Méthodes d'inférence démographique récente utilisant les polymorphismes et leur liaison génétique

Elise Kerdoncuff

École Doctorale 227
Sciences de la nature et de l'Homme : évolution et écologie

SMILE, CIRB, Collège de France.
ABI, ISYEB, Muséum National d'Histoire Naturelle.

Thèse de doctorat

Discipline : Génétique et Génomique des Populations

Méthodes d'inférence démographique récente utilisant les polymorphismes et leur liaison génétique.

Présentée pour obtenir le grade de
Docteur de Sorbonne Université par
Élise Kerdoncuff

dirigée par
Guillaume Achaz & Amaury Lambert

Présentée et soutenue publiquement le 10 novembre 2021 devant un jury composé de :

Lounès Chikhi	Directeur de Recherche, CNRS	Rapporteur
Aurélien Tellier	Professeur, Technische Universität München	Rapporteur
Flora Jay	Chargée de Recherche, CNRS	Examinatrice
Bertrand Servin	Directeur de Recherhce, INRAE	Examinateur
Emmanuelle Porcher	Professeure, MNHN	Examinatrice
Guillaume Achaz	Professeur, Université de Paris	Directeur
Amaury Lambert	Professeur, École Normale Supérieure	Directeur

Table des matières

Chapitre 1

Introduction

Contents

1.1 Contexte

1.1.1 Sixième crise de la biodiversité

L'accroissement de la population humaine et de la consommation induit, entre autres, le changement climatique, la surexploitation des écosystèmes, la dégradation et la perte d'habitat pour certaines espèces ainsi que les invasions biologiques (Primack 2012). Ces différentes pressions entraînent des disparitions d'espèces, de la perte de biodiversité. Depuis les 200 dernières années, le nombre d'extinction a augmenté considérablement. Le taux moyen de vertébrés disparus dans le siècle dernier, est supérieur à cent fois le taux en l'absence de pression anthropique (Ceballos et al. 2015). Actuellement, nous vivons probablement la 6ème crise d'extinction (Barnosky et al. 2011), la 5ème étant la crise Crétacé-Tertiaire ayant vu les dinosaures disparaître. Ces extinctions de masse touchent des taxons et des habitats variés : 26% des mammifères sont en voie d'extinction, 41% des amphibiens et 41% des gymnospermes (Chiffres UICN 2021), ainsi que les insectes, qui représentent plus de la moitié de la biodiversité, dont l'abondance diminuerait de 1 à 2% par an (Wagner et al. 2021). Une perte importante de biodiversité est donc en cours sur l'ensemble de la planète.

Les espèces et leurs interactions jouent un rôle fonctionnel important au sein des écosystèmes, une perte importante de la biodiversité affecte les équilibres des écosystèmes (Cardinale et al. 2012). Ces changements d'équilibre vont altérer les différents services, appelés services écosystémiques, procurés par les écosystèmes et les espèces qui les composent. La quantification des risques d'extinction et l'attribution de statuts de conservation aux différentes espèces permettent de se rendre compte de l'ampleur de cette crise d'extinction. L'étude des dynamiques des différentes espèces au cours du temps, qui permet l'attribution de statuts de conservation, est donc indispensable pour appréhender l'évolution ainsi que les futurs équilibres des écosystèmes.

La ressource majeure pour étudier l'extinction des espèces est la liste rouge de l'Union Internationale pour la Conservation de la Nature (UICN ou *IUCN* en anglais). Cette liste regroupe les statuts de conservation disponibles pour les différentes espèces (Fig 1.1). Chaque espèce ou sous-espèce peut être classée dans l'une des neuf catégories suivantes : Éteinte (*Extinct* EX), Éteinte à l'état sauvage (*Extinct in the wild* EW), En danger critique (*Critically endangered* CR),

En danger (*Endangered* EN), Vulnérable (*Vulnerable* VU), Quasi menacée (*Near threatened* NT), Préoccupation mineure (*Least concern* LC), Données insuffisantes (*Data deficient* DD), Non évaluée (*Not evaluated* NE).

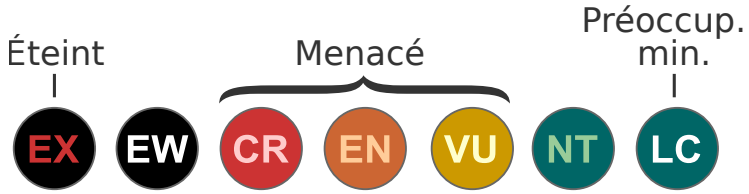


Figure 1.1: Catégories UICN informant sur le statut de conservation de l'espèce, ou sous-espèce. (Source UICN - Version 2006)

Pour attribuer des statuts de conservation, l'UICN utilise des critères basés sur différents facteurs biologiques comme la taille de la population, l'aire de répartition géographique ou la fragmentation de la répartition. L'UICN se base sur les changements de ces différents facteurs au cours du temps (Rapport IUCN 2012). La liste rouge de l'UICN est actuellement la meilleure compilation disponible des statuts de conservation (Lamoreux et al. 2003). Elle est utilisée dans de nombreux processus de décision, notamment pour définir les espèces à conserver en priorité ou pour étudier les changements au niveau de la biodiversité. Son efficacité est cependant limitée en vue de ces objectifs (Possingham et al. 2002; Rodrigues et al. 2006). En effet, actuellement, seulement 6% des espèces décrites ont un statut de conservation : 135,000 espèces ont été évaluées par l'UICN alors que plus de 2.1 millions d'espèces ont été décrites (et il en existerait au total, probablement, plus de 7 millions (Mora et al. 2011)). Parmi ces espèces, 37,500 sont menacées d'extinction, ce qui représente 28% des espèces ayant un statut de conservation et 1.7% des espèces décrites. Les statuts de conservation ainsi que les risques d'extinction restent néanmoins inconnus pour approximativement 94% des espèces connues. Ce manque d'information est problématique pour pouvoir définir les espèces à conserver en priorité et avoir une vision globale de la dynamique de la biodiversité.

Les statuts de conservation ne sont pas collectés de manière homogène sur l'ensemble de l'arbre du vivant. La quasi-totalité des mammifères, des oiseaux, des amphibiens et des gymnospermes sont évalués. Les autres groupes n'ont pas une couverture suffisante afin d'estimer le pourcentage d'espèces menacées au sein de ces groupes. Par exemple, seulement 2% des espèces d'insectes connues ont un statut de conservation : 25,000 espèces d'insectes ont été évaluées alors que quasiment 1.5 million d'espèces ont été décrites ; les insectes représentent plus de la moitié de la biodiversité sur Terre. Ce biais taxonomique est sans doute lié aux mesures nécessaires pour étudier les critères de l'UICN. Les comptages d'individus, les mesures de leur aire de répartition, sont applicables facilement à certains groupes d'individus (mammifères, oiseaux ...) mais difficilement à d'autres (insectes, algues ...) pour lesquels ces données peuvent être lourdes à récolter et demander une expertise taxonomiste. Ainsi, les critères et mesures nécessaires à l'utilisation des méthodes de l'UICN sont difficilement applicables à l'ensemble des espèces. Il existe donc un

réel besoin de mettre en place de nouvelles méthodes universelles, utilisables pour l'ensemble des taxons du vivant.

En 2015, un groupe de chercheurs incluant Amaury Lambert et Guillaume Achaz, a mis en place un modèle stochastique traitant des données d'occurrence (Régnier et al. 2015). Ils se sont intéressés à 200 espèces de mollusques terrestres échantillonnées au hasard parmi la totalité des espèces de mollusques connues. Il s'agit d'un groupe pour lequel peu de statuts de conservation sont disponibles, certains ont, tout de même, été attribué par des experts taxonomiques. Régnier et al. se sont intéressés, pendant la période 1850-2000, aux années et localisations où les espèces avaient été vues, celles où elles n'ont pas été vues alors qu'elles avaient été cherchées et celles où il n'y a pas eu de relevés. A l'aide de ces données, ils ont développé un modèle inférant un changement dans le taux d'extinction de l'espèce ainsi que l'année d'extinction de cette espèce, permettant de déterminer si l'espèce n'a pas été observée parce qu'elle a disparu ou parce que l'effort d'échantillonage n'était pas suffisant. A partir de ce modèle, Régnier et al. ont pu attribuer des statuts d'extinction aux différentes espèces. Les statuts qu'ils ont attribués aux espèces étaient globalement conformes à ceux attribués par les experts taxonomiques et par l'UICN (lorsque ces derniers étaient disponibles). Ils ont également pu attribuer des statuts de conservation ainsi que des probabilités d'extinction (probabilités que les espèces soient déjà éteintes) à des espèces pour lesquelles les experts et l'UICN n'avaient pas suffisamment de données pour déterminer un statut. L'utilisation de modèles mathématiques permet d'attribuer des statuts de conservation à des espèces pour lesquelles les critères habituels ne sont pas applicables. Cependant, les données d'occurrence sont lourdes à récolter, demandent de nombreux relevés et l'avis d'experts. Elles ne permettent pas de *passer à grande échelle*. Afin de mettre en place une méthode qui permettrait d'attribuer des statuts de conservation ainsi que des risques d'extinction à une majorité des espèces, il est nécessaire d'utiliser un autre type de données, plus facilement récoltables pour l'ensemble des espèces. Dans le cadre de cette thèse, j'ai exploré l'exploitation de données génomiques à des fins de conservation.

Certaines méthodes permettant d'estimer la taille d'une population utilisent les séquences d'ADN des individus de la population. Le coût du séquençage de l'ADN diminue drastiquement et devient donc de plus en plus accessible. Les génomes de référence, facilitant les séquençages, se généralisent dans l'arbre du vivant. Les méthodes utilisant l'ADN des individus afin d'estimer la taille de la population pourraient, dans l'avenir, servir à attribuer des statuts de conservation. Ces méthodes sont basées sur des modèles de génétique des populations.

1.1.2 L'origine de la génétique des populations

Cette partie historique, non exhaustive, est en partie inspirée des chapitres 10 et 11 de l'ouvrage *Penser l'évolution* d'Hervé Le Guyader, paru en 2012.

1.1.2.1 Le concept de gène

Le développement de la biologie moléculaire a permis de séquencer des fragments d'ADN porteurs d'information génétique, unité de base de l'hérédité, qu'on appelle aujourd'hui gène. Cependant le concept de gène est pourtant bien plus ancien. Il provient de trois caractéristiques du gène, il s'agit d'une unité de transmission, une unité de mutation et une unité de recombinaison.

Unité de transmission

Le premier à publier un article de *génétique* est Gregor Mendel (1822-1884) en 1865 *Recherches sur des hybrides végétaux* (Mendel 1865). Lors de ces expériences de croisement de plantes du genre *Pisum*, il observa la régularité à laquelle revenait les mêmes formes d'hybrides. Il étudia la loi des variations des caractères entre des hybrides et leurs descendants, et ce au fil des générations. En s'intéressant aux descendants hybrides en fonction de la provenance de l'ovule ou du pollen, Mendel fût le premier à conceptualiser la transmission de facteurs responsables des caractères observables par les cellules sexuelles. Il trouva que ces facteurs étaient apportés, à parts égales, par le gamète mâle et le gamète femelle. Ils font partie intégrante de la cellule et chaque cellule possède, en proportions égales, une combinaison de caractères résultant de la fécondation croisée précédente. Mendel va même conclure que les caractères différenciant deux plantes proviennent de différences dans la composition et le regroupement des éléments à l'intérieur des cellules à l'origine des deux plantes.

Ces facteurs ne seront jamais observés par Mendel, il s'agit, à cette époque, seulement d'un objet conceptuel.

Cette première particularité des cellules sexuelles apparut également dans les travaux d'August Weismann (1834-1914), dans sa théorie *continuité du plasma germinatif* (Weismann 1892). Il fit la distinction entre *soma* et *germen*, le soma étant l'ensemble des cellules du corps à l'exception des cellules germinales. Ces dernières engagées dans la reproduction sexuée forment le germen.

Le concept-clé de Darwin est celui de la descendance avec modification, qui postule que les caractères héréditaires sont variables (publié dans son ouvrage majeur *L'origine des espèces* en 1859).

Le brassage génétique définit par Mendel, concernait des caractères fixes : il ne vont pas changer au cours des générations, mais pouvant générer, par des combinaisons différentes, une variabilité de descendants. Ces deux concepts, à l'époque où les protagonistes vivaient, ne semblaient pas pouvoir s'accorder.

Unité de mutation

Hugo De Vries (1848-1935) permit la réconciliation de ces deux concepts, on ajoutant l'idée de la mutation aux lois de Mendel (de Vries 1901). Il appelle *mutation* une variation brusque et discontinue, qui diffère de celles continues et limitées

déjà décrites. Elles pourraient être à l'origine de la variabilité de l'espèce et donner prise à la sélection sexuelle.

Le terme *gène* fut définit par Wilhelm Johannsen (1857-1927) (Johannsen 1909), il fit également la distinction entre génotype, l'ensemble des gènes, et phénotype, l'ensemble des caractères. Les principes mendéliens sont également étendus aux espèces animales, notamment par Lucien Cuénot (1866-1951) (Cuénot 1902) et William Bateson (1961-1926) (Bateson 1903), qui est l'auteur du mot *génétique* (Bateson et al. 1906).

Unité de recombinaison

Walter Sutton (1877-1916), après avoir observé le comportement des chromosomes au moment de la méiose, propose la théorie chromosomique de l'hérédité : les chromosomes seraient le support des gènes (Sutton 1902, 1903). Cette théorie fut, notamment démontré par les travaux sur les drosophiles de Thomas H Morgan (1866-1945) (Morgan et al. 1915).

Les anciens facteurs mendéliens, devenus gènes, ont donc une réalité physique et même un emplacement physique grâce aux cartes chromosomiques de Morgan. Ces cartes ont permis de mettre en évidence la liaison entre certains gènes et l'indépendance d'autres ; mais également la recombinaison.

Le concept de gène : son caractère héréditaire, sa capacité à varier avec la mutation et sa transmission plus ou moins liées avec la recombinaison, sont centrales dans la modélisation en génétique des populations.

1.1.2.2 Naissance de la génétique des populations

La génétique des populations est le domaine de la biologie qui étudie la composition génétique des populations d'organismes vivants et les changements qui surviennent par l'action de divers facteurs. Les modèles de génétique des populations, pouvant être très théoriques, permettent de proposer des scénarios en biologie évolutive, de mettre en avant les différents processus ayant façonnés la diversité observée aujourd'hui.

Premier modèle neutre : Hardy & Weinberg

Le premier modèle fut développé indépendamment par Godfrey Hardy (1877-1947) (Hardy 1908) et Wilhelm Weinberg (1862-1937) (Weinberg 1908) en 1908. Ce modèle, le plus simple possible, permet de montrer que, sous les lois de Mendel, il n'est pas possible de prédire qu'un allèle dominant se fixera dans la population et qu'un allèle récessif disparaîtra, les deux allèles resteront donc présents dans la population, à fréquence constante. Hardy considère les hypothèses les plus simples possibles et rappelle que des croisements non aléatoires, ainsi qu'un caractère lié au sexe des individus ou ayant une influence sur la fertilité des individus, changeront les résultats du modèle.

Aujourd'hui, il est possible de résumer les hypothèses d'Hardy-Weinberg de la manière suivante : les croisements entre les individus se font de manière aléatoire, la population est de grande taille, les processus évolutifs (la sélection, la mutation, la migration, la dérive) ne jouent pas de rôle. Si ces hypothèses sont respectées alors la population est à l'équilibre : ses proportions génotypiques, ses fréquences alléliques sont constantes d'une génération à l'autre : c'est l'équilibre d'Hardy-Weinberg.

Par la suite, les modèles seront de plus en plus complexes, afin de prendre en compte les processus évolutifs qui entraînent des écarts à l'équilibre d'Hardy-Weinberg.

La sélection : Haldane & Fisher

La sélection peut être intégrée dans le modèle grâce à la valeur sélective ou *fitness* en anglais, introduite pas John Haldane (1892-1964). Dans le premier de ses dix articles contenus dans *A mathematical theory of natural and artificial selection* paru en 1924 (Haldane 1924), il présente des équations permettant de faire correspondre une intensité de sélection au taux auquel la proportion d'individus portant un certain caractère augmente ou diminue (connaissant le système de reproduction, le mode de transmission ainsi que la présence ou non du caractère chez les différents sexes). Il considère deux phénotypes A et B et il introduit k un coefficient de sélection. Chaque individu B produit, en moyenne, autant de descendants que $(1-k)$A. Le coefficient de sélection k peut être compris entre 1, aucun descendant de B n'est produit et $-\infty$, aucun descendant de A n'est produit. Haldane décrit ces cas comme de la 'sélection complète'.

En 1937 dans *The Effect of Variation of Fitness* (Haldane 1937), Haldane démontre également que dans le cas où le génotype hétérozygote présente la valeur sélective la plus élevée, il existe un équilibre polyallélique. Cela explique comment deux allèles peuvent être maintenus dans une population à l'équilibre. Il introduit également la notion d'équilibre mutation-sélection : la mutation introduit l'allèle dans la population, la sélection, si elle est négative, l'élimine de la population.

Ronald Fisher (1890-1962), dans son livre de 1930, développe, entre autres, des travaux sur la sélection sexuelle et le sexe-ratio. Il introduisit notamment, les notions de préférence dépendant du sexe de l'individu (souvent la femelle) et de sélection sexuelle. C'est à lui qu'on doit l'idée d'*emballement* expliquant l'exagération de caractères sexuels secondaires sous la pression sélective du choix de partenaire basé sur ce caractère. En supposant que le coût de production d'un mâle et d'une femelle est identique, il expliqua pourquoi, dans la plus part des espèces le sexe ratio équilibré est la stratégie évolutivement stable.

Il contribua également énormément à la modélisation de la génétique des populations et intégra la stochasticité au sein des modèles en génétique des populations. Il considéra que l'effet de la fluctuation aléatoire peut être négligé dans le cas des populations de grande taille. Fisher est également connu pour ses travaux sur l'analyse de la variance dont proviennent les tests de type ANOVA/ANCOVA ainsi

que pour son travail sur la mesure de l'information utilisée, notamment, dans la recherche de maximum de vraisemblance.

La dérive : Wright

Sewall Wright (1889-1988) va, quant à lui, s'intéresser aux populations de petite taille et introduire un processus majeur : la dérive génétique, *random drift* en anglais. Il s'agit du processus produisant des modifications aléatoires de fréquence allélique dans une petite population, dû à un biais d'échantillonnage.

Cependant, il a également travaillé sur les effets de la sélection naturelle. En effet, il va être le premier à introduire l'idée de paysage adaptatif (Wright 1931), il va représenter la valeur sélective d'un individu en fonction de la combinaison des gènes qu'il possède et les enchainements possibles pour passer d'une combinaison de gène à une autre. Ces paysages pouvant être composés de pics et de vallées, il va exprimer le fait que pour évoluer les espèces ne doivent pas être sous une sélection stricte, les descendants d'individus doivent pouvoir survivre avec une valeur sélective plus basse que leur parents.

Il va également représenter les différences de paysages adaptatifs en fonction de la rapidité du taux de mutation ou de la force de la sélection, tous les deux dépendant de la taille de la population. Dans les populations de petite taille, les fluctuations aléatoires des fréquences alléliques sont beaucoup plus importantes et peuvent entraîner la fixation d'allèle sans avantage sélectif (ou avec un désavantage sélectif). Plus une population est petite, plus la probabilité de fixation est grande (elle dépend de la taille de population). Sans la connaissance de la dérive, ces fixations pouvaient être considérées comme de la sélection.

Théorie Synthétique de l'Evolution

L'alliance entre la théorie darwinienne et la génétique donne naissance à la théorie synthétique de l'évolution, principalement due à Theodosius Dobzhansky (1900-1975), Ernst Mayr (1905-2005) et Julian Huxley (1887-1975) qui lui donne ce nom (Huxley 1942).

La Théorie Synthétique de l'évolution repose grandement sur ce qui est décrit précédemment et ces principales hypothèses ont été résumées de la manière suivante par Hervé Le Guyader dans son ouvrage *Penser l'évolution* (Le Guyader 2012) :

1. L'hérédité est particulaire, et d'origine exclusivement génétique.
2. Il y a une *énorme* variabilité dans les populations naturelles.
3. L'évolution se déroule dans des populations distribuées géographiquement.
4. L'évolution procède par modification graduelle des populations.
5. Les changements dans les populations sont le résultat de la sélection naturelle.

6. Les différences observées entre des organismes sont, pour une grande part, des adaptations.
7. La macro-évolution n'est que la prolongation, avec le temps, de ces mêmes processus qui contrôlent l'évolution des populations.

La plupart des différences observées étant due à des adaptations, cette théorie considère la sélection naturelle comme étant le principal processus évolutif. Ce courant fait donc parti du sélectionnisme.

1.1.2.3 Le neutralisme

Remise en cause du sélectionnisme

La biologie moléculaire, avec notamment la découverte de la molécule d'ADN en 1953, a eu un impact important sur la biologie évolutive. Il est devenu possible de comparer des séquences d'acides aminés, et donc d'avoir accès à la variabilité nucléotidique, et non plus seulement la variabilité génétique. En comparant des séquences nucléotidiques d'espèces proches dont la date de spéciation est connue, il est possible d'estimer un taux de substitution de nucléotides. Zuckerkandl et Pauling (1965) montrent que les taux de substitution sont constants entre lignées et formulent l'hypothèse de l'horloge moléculaire : les mutations s'accumulant à vitesse constante dans les génomes, elles peuvent être utilisées pour dater des évènements. Cette hypothèse, qui implique que l'accumulation des mutations de dépend pas de l'environnement dans lequel se trouve l'espèce, ne semble pas en accord avec la sélection comme processus à l'origine du polymorphisme observé.

De plus, la notion de *coût de la sélection* formulée par Haldane, implique que si un allèle est fixé par la sélection d'autres ont nécessairement été définitivement perdus. Cela signifie que si toutes les substitutions observées sont issues de la sélection, le nombre de substitutions s'étant réellement produit est bien supérieur à celui observé. Les taux de substitution estimés deviennent bien supérieur à ceux estimés précédemment. Ces observations remettent en cause le sélectionnisme et sont à l'origine du neutralisme.

La découverte de la redondance du code génétique permit d'identifier des mutations de nucléotides sans effet sur l'acide aminé traduit. Les pseudogènes, introns, régions intergéniques qui ne participent pas à l'élaboration des protéines, sont aussi porteurs de mutations. L'étude des régions variantes, sans effet sur le génotype a soutenu la théorie neutraliste (Takahata 2007).

Théorie neutraliste de l'évolution

La théorie neutraliste de l'évolution a été formulée par Motoo Kimura (1924-1994) en 1968, puis explicitée dans son livre en 1983. C'est une théorie de l'évolution moléculaire selon laquelle la majorité des changements évolutifs à l'échelle moléculaire proviennent de mutations neutres. Ce ne serait donc pas la sélection darwinienne mais la dérive génétique qui façonne la diversité observée. L'existence,

ainsi que l'effet, de la sélection naturelle ne sont pas remis en cause, mais seule une infime fraction des changements de l'ADN au cours du temps serait due à ce processus. La grande majorité des mutations n'auraient aucun effet sur le phénotype de l'individu, ni sur sa valeur reproductive et dériveraient aléatoirement dans la population. La théorie neutraliste nécessite une différenciation entre l'évolution au niveau moléculaire et celle au niveau phénotypique. La sélection naturelle agit sur certains phénotypes, par leur valeur sélective et peut avoir un effet indirect sur les gènes. La dérive génétique, quant à elle, a un effet direct, sur tous les nucléotides. La majorité des polymorphismes observés au sein d'une espèce serait donc neutre et dans une phase transitoire entre la fixation ou la disparition et non pas maintenue par une sélection balancée.

La théorie neutraliste attribue donc la diversité observée au sein des populations à la dérive génétique, lui donnant un poids plus important que celui de la sélection.

La théorie *presque* neutre

La théorie presque neutre de l'évolution est une extension de la théorie neutraliste de l'évolution, développée par Tomoko Ohta (1933-) en 1973. Elle considère l'existence de mutations intermédiaires, dont l'effet est compris entre celui des mutations neutres et celui des mutations sélectionnées et qui ont leur importance au niveau moléculaire. Des mutations légèrement délétères ne disparaissent de la population que si leur coefficient de sélection est supérieur à l'inverse de la taille de la population, si l'effet de la sélection est plus fort que celui de la dérive. Tomoko Ohta propose que la plupart des substitutions d'acides aminés soit légèrement délétère, ce qui a pour effet d'augmenter le taux de fixation dans les petites populations par rapport aux grandes populations : moins de mutations disparaissent. Ce qui expliquerait que l'on retrouve des taux de fixation similaires entre des organismes de

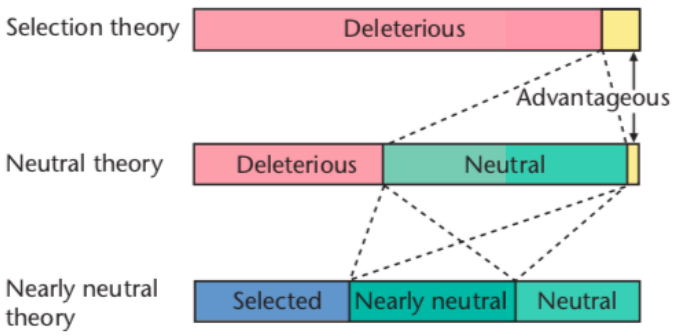


Figure 1.2: Classification des nouveaux mutants suivant la théorie sélectionniste (en haut), la théorie neutraliste (au milieu) et la théorie presque neutre (en bas). Les mutants peuvent être délétères (rouge), avantageux (jaune) ou neutre (vert). L'intensité de leur sélection peut être très faible, *Nearly neutral* (vert légèrement plus foncé) ou plus importante, *Selected* (bleu). Figure issue de Ohta 2013.

grande taille, avec des temps de générations longs et des petites populations et des organismes de petites tailles, aux temps de générations courts et aux grandes populations. D'après le concept d'horloge moléculaire, un nombre plus important de mutations devraient être accumulées dans les populations au temps de génération court, ce qui n'est pas le cas dans les observations. Tomoko Ohta propose donc une nouvelle classification des nouvelles mutations arrivant dans une population (Fig 1.2 de Ohta 2013), elle y résume la proportion de chaque type de mutation suivant le courant considéré.

1.2 Modélisation en génétique des populations

1.2.1 Modèle Standard Neutre

De nombreux modèles mathématiques ont découlé du neutralisme, j'en décris seulement deux ici, le modèle de Wright-Fisher en temps prospectif et le coalescent de Kingman en temps rétrospectif.

1.2.1.1 Modèle de Wright-Fisher

Un des modèles les plus classiques en génétique des populations est celui développé par Fisher (1930) et Wright (1931). Dans ce modèle, la taille de la population N est constante. Les générations ne sont pas chevauchantes. A chaque génération, tous les individus meurent et sont remplacés par de nouveaux individus. Les génotypes de ces nouveaux individus sont tirés aléatoirement parmi ceux de la génération précédente (tirage avec remise). Chaque individu a la même probabilité de voir son génotype tiré, la reproduction est aléatoire. Les bifurcations dans la généalogie d'un individu sont créées quand deux enfants sont issus du même parent. Cela arrive avec une probabilité de $1/N$, elle dépend donc de la taille de la population.

Une variante de ce modèle est le modèle de Moran (1958), pour lequel un seul individu meurt à la fois et le génotype de son descendant est tiré aléatoirement de la même manière. Dans ce second modèle, le temps est continu.

1.2.1.2 Coalescent de Kingman

On appelle évènement de coalescence la jonction de deux lignées en une, quand deux individus tirent le même parent dans la génération précédente. L'arbre de coalescence est l'arbre des lignées ancestrales d'un ensemble d'individus, jusqu'à leur ancêtre commun (ou *Most Recent Common Ancestor* (MRCA) en anglais). Le processus à l'origine de l'arbre a été décrit par John Kingman (1982) et correspond à des résultats obtenus par d'autres approches comme celles de Wrigth-Fisher et Moran. Le n-coalescent de Kingman est exactement l'arbre d'un echantillon uniforme de n individus dans le modèle de Wright-Fisher, dans la limite $N \to \infty$. Dans l'approche coalescente, on remonte le temps du présent vers le passé, en temps rétrospectif, à l'inverse de Wright-Fisher et Moran.

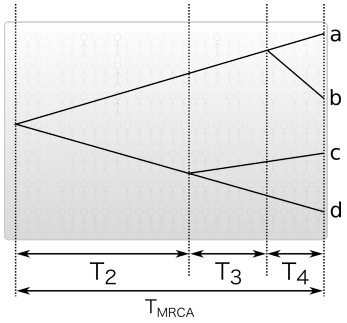


Figure 1.3: Arbre de coalescence de 4 individus (a,b,c et d). a et b coalescent au temps T_4, c et d au temps $T_3 + T_4$, les deux dernières lignées coalescent (et forment le MRCA) au temps $T_{MRCA} = T_2 + T_3 + T_4$.

Le n-coalescent est composé de $n-1$ évènements de coalescence, à chaque évènement deux lignées fusionnent, le nombre d'individus présents diminue donc de 1. On nomme T_i le temps qui s'écoule entre i et $i-1$ lignées (Fig 1.3). Le dernier temps de coalescence est donc T_2, entre les deux dernières lignées. Tous les T_i sont indépendants. Le temps de l'ancêtre commun le plus récent, T_{MRCA} est la somme des temps de coalescence précédent (Fig 1.3) :

$$T_{\text{MRCA}} = \sum_{i=2}^{n} T_i.$$

La somme de toutes les longueurs de branches s'exprime de la manière suivante :

$$T_{total} = \sum_{i=2}^{n} iT_i.$$

Les propriétés du coalescent dépendant de sa taille N, on exprime le temps en unités de N générations, unité de temps coalescent.

Kingman a montré que, quand N tend vers l'infini et que les hypothèses de neutralité, de taille constante et de panximie sont respectées, ainsi que la probabilité de trifurcation (coalescence de trois individus) est négligeable par rapport à la probabilité de bifurcation, les temps T_i suivent une loi exponentielle de paramètre $\binom{i}{2}$. Ce qui entraîne les propriétés suivantes :

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

ainsi que :

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

1.2.2 Mutation

1.2.2.1 Modélisation

Dans le cadre du modèle neutre, les mutations n'ont aucun effet sur le potentiel reproducteur des individus et donc aucun effet sur la généalogie des lignées. Le

processus de mutation est indépendant du processus généalogique. La probabilité d'observer un évènement de mutation au cours d'une génération est faible. Le taux de mutation μ par génération et par site est, la plupart du temps, considéré constant au cours du temps et le long du génome. La variable aléatoire M le nombre de mutation suit une loi de poisson :

$$M \sim Poisson(\mu T_{total})$$

avec T_{total} mesuré en générations (et non N générations comme précédemment).

1.2.2.2 Diversité génétique

Une paramètre décrivant la diversité génétique au niveau de la population est $\theta = 2N\mu$, le double du nombre moyen de mutations introduites dans la population de taille N à chaque génération. Il représente le nombre moyen de différences attendues entre deux loci échantillonnés dans la population. Plus une population est petite (plus N est petit), plus le temps de coalescence entre deux individus est faible, moins de mutations ont lieu et donc moins il y a de différences entre deux loci de la population.

Estimateur de Watterson (1975) A partir d'un alignement de séquences, il est possible de mesurer le nombre de sites polymorphes M (parfois appelé S), dont l'espérance est égale à :

$$E(M) = \mu E(T_{total}) = 2N\mu a_n$$

avec

$$a_n = \sum_{i=1}^{n-1} \frac{1}{i}.$$

Ce qui permet d'estimer θ, à l'aide de l'estimateur de Watterson noté $\hat{\theta}_w$ (ou $\hat{\theta}_S$), qui se calcule de la manière suivante :

$$\hat{\theta}_w = \frac{M}{a_n}.$$

Il est possible de généraliser cet estimateur en dehors du modèle standard neutre en remplacant a_n par un équivalent dépendant de la longueur totale des branches de l'arbre. En effet, sous le modèle standard neutre :

$$a_n = \frac{E[T_{total}]}{2},$$

il est donc possible de remplacer a_n par $\frac{E[T_{total}]}{2}$, estimé suivant le scénario considéré (T_{total} exprimé en N générations).

1.2.3 Recombinaison

À chaque site le long d'un chromosome correspond un arbre de coalescence. Deux sites peuvent avoir la même généalogie ou des généalogies différentes s'il y a eu un évènement de recombinaison entre ces deux sites depuis le dernier ancêtre commun (MRCA) de leur généalogie. Les évènements de recombinaison vont entremêler des segments d'ADN avec des histoires différentes. Deux sites voisins ont plus de chances de partager la même généalogie que deux sites éloignés. Les généalogies des sites sont interdépendantes, elles vont dépendre des évènements de recombinaison qui ont eu lieu le long des séquences et de l'ensemble des généalogies des lignées. En effet, chaque segment de séquence non recombiné a sa propre généalogie, mais ces généalogies sont liées puisqu'elles se retrouvent sur le même génome. La modélisation d'un génome complet se fait grâce au coalescent avec recombinaison.

1.2.3.1 Coalescent avec recombinaison

Le coalescent avec recombinaison décrit la généalogie d'individus, en commençant au présent et en remontant dans le temps jusqu'à rencontrer un évènement de coalescence (deux lignées coalescent en une lignée ancêtre) ou un évènement de recombinaison (une lignée ancêtre est séparée en deux).

Les évènements de recombinaison sont rares. Soit ρ/N le taux de recombinaison par génération, qui est généralement considéré constant au cours du temps et le long du génome, la distance sur le chromosome à laquelle se produit un évènement de recombinaison suit une loi exponentielle de paramètres $\rho/2 \times T_{total}$ (Wiuf and Hein 1999). Ainsi pour les arbres longs, la distance entre deux évènements de recombinaison sera plus faible. Inversement, pour les arbres courts, la distance entre deux évènements de recombinaison sera plus importante.

Ancestral Recombination Graph Le stockage de l'histoire des lignées ne s'effectue plus dans un arbre mais dans un graphe nommé l'*Ancestral Recombination Graph* (ARG) (Griffiths and Marjoram 1997). Toute l'information des ancêtres est contenue dans ce graphe (Fig 1.4). En bref, un brin d'ADN est représenté par un intervalle [0,1] sur lequel se produisent des évènements de recombinaison. Chaque évènement de recombinaison se produisant sur le brin d'ADN entraînera la création d'un nouvel sous-intervalle possédant sa propre généalogie. Chaque arbre de coalescence dépend de l'ensemble des évènements de recombinaison apparus sur le chromosome, il est donc nécessaire de mémoriser tous les évènements de recombinaison pour former un ARG.

L'inférence de l'ARG est complexe puisqu'à chaque temps, les taux globaux de coalescence et de recombinaison vont dépendre des lignées ancêtres présentes. Il y a cependant des tentatives (Rasmussen et al. 2014).

Tree Sequence Une autre approche permettant de décrire le coalescent avec recombinaison et de considérer chaque généalogie par segment non recombiné. Le

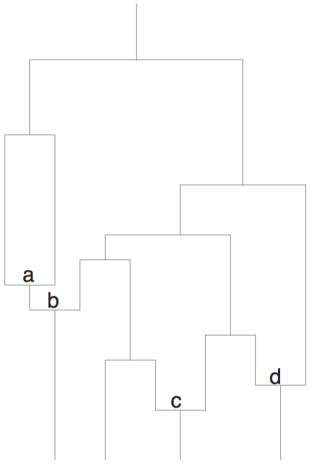


Figure 1.4: Illustration d'un *Ancestral Recombination Graph*. Il y a eu 4 évènements de recombinaison, aux points a, b, c et d. Figure issue de Griffiths and Marjoram 1997.

génome est donc représenté par différents segments, chacun ayant sa propre généalogie, cette représentation est appelée *Tree Sequence* (ou *succinct tree sequence*) (Kelleher et al. 2019).

Il est possible de simuler les généalogies en avançant le long de la séquence d'ADN. On appelle G_k la généalogie du k-ème segment. On commence par une généalogie initiale G_0 puis on construit la généalogie d'après G_1 à partir de la précédente G_0 et ainsi de suite (Wiuf and Hein 1999). G_k dépend non seulement de G_{k-1} mais également de toutes les généalogies précédentes. Ceci implique qu'il faut garder en mémoire tous les évènements de recombinaison (comme pour l'ARG). Il s'agit d'une structure non-markovienne complexe.

Cette représentation est notamment utilisée dans le simulateur *msprime* (Kelleher et al. 2016).

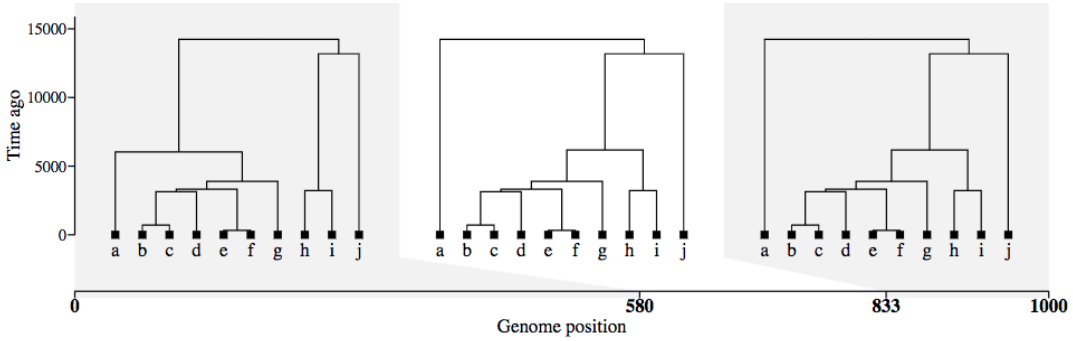


Figure 1.5: Illustration d'une *Tree Sequence*. Le génome peut être divisé en 3 segments, un de 0 à 580 possédant la généalogie G_0, un de 580 à 833 (G_1) et un de 833 à 1000 (G_2). La séquence est composée de 3 segments possédant chacun leur généalogie. Figure issue du tutoriel de tskit (Kelleher et al. 2019).

1.2.3.2 Approximation : *Sequentially Markovian Coalescent*

Il existe une approximation du coalescent avec recombinaison qui facilite les simulations en supposant que la tree sequence (G_k) est markovienne : le *Sequentially Markovian Coalescent* (SMC) (McVean and Cardin 2005). L'arbre de chaque segment dépend uniquement de l'arbre du segment d'avant sur lequel s'est produit un évènement de recombinaison. Cette approximation permet de réduire les temps de calcul et la mémoire nécessaire à la simulation du coalescent avec recombinaison. Elle permet également d'avoir une intuition sur l'effet des évènements de recombinaison sur le génome.

L'idée générale est de simuler un arbre de coalescence G_k auquel correspond une longueur de segment non recombiné. Sur cet arbre un évènement de recombinaison se produit aléatoirement à un temps précis, sur une branche précise, la branche sur laquelle est arrivée l'évènement de recombinaison est disjointe de l'arbre et se rattachera suivant le coalescent de Kingman. L'arbre nouvellement créé a une généalogie différente (G_{k+1}) de G_k et correspond au segment non recombiné suivant.

Il existe une variante de SMC appelée SMC' (Marjoram and Wall 2006). Le SMC' permet à la branche de l'arbre sur laquelle l'évènement de recombinaison s'est produit de coalescer avec son ancienne branche ancêtre (avant l'évènement de recombinaison), ce que ne permettait pas le SMC originel.

1.2.3.3 Conséquences

Mosaïque En étudiant un génome (ou un chromosome) entier, nous avons donc accès à une mosaïque d'histoires évolutives. Les génomes sont une compilation de segments non recombinés qui ont chacun des histoires différentes, une combinaison de loci avec des histoires différents. Des loci sélectionnés au hasard sur le génome peuvent donc être considérés comme indépendants. Un alignement de séquences reflète donc non pas un arbre de coalescence mais un ensemble d'arbres. Les mesures effectuées sur le génome renvoie donc une valeur moyenne de ces différents arbres.

Liaison La liaison physique des loci proches physiquement sur le génome, peut également entraîner une liaison génétique. Ces loci auront tendance a être transmis ensemble d'un individu à sa descendance. La liaison génétique est une notion statistique, elle correspond à la probabilité d'association entre allèles à des loci différents. Un évènement de recombinaison peut briser la liaison entre deux loci, qui seront, par la suite, transmis indépendamment. Le nombre d'évènements de recombinaison a donc un impact sur la liaison entre des loci.

1.2.3.4 Test des 4 gamètes

Il est possible de détecter certains évènements de recombinaison en utilisant le test des 4 gamètes (Hudson and Kaplan 1985). On considère un modèle à sites infinis : le nombre de sites pouvant muter est infini, les mutations ne peuvent se produire qu'une fois au même site et il n'y a pas de recombinaison. Sous ses hypothèses, si on considère deux sites différents dialléliques A/a et B/b, un seul arbre

généalogique, il est possible d'observer seulement 3 associations d'allèles. Dans la Figure 1.6, on observe les combinaisons A-B, a-B et a-b, il manque la combinaison A-b. Si les quatre combinaisons sont présentes, les deux sites ne sont pas compatibles avec la même généalogie, ce qui indique qu'au moins une des hypothèses du modèle n'est pas respectée. La généalogie d'une population correspond souvent à des temps courts, une mutation apparaissant deux fois sur le même site est peu probable. Cela indique donc la présence d'un évènement de recombinaison entre les deux sites.

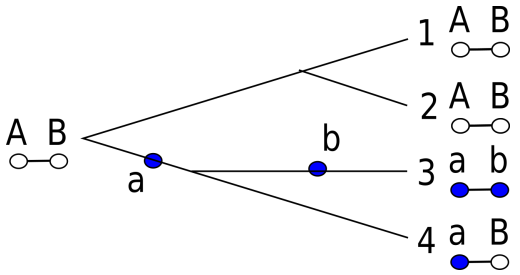


Figure 1.6: Illustration de 3 génotypes possibles pour deux sites avec la même généalogie et l'absence de mutations redondantes. Allèles ancestraux A et B (blanc), allèles dérivés a et b (bleu).

Tous les évènements de recombinaison ne sont pas détectables à l'aide du test des 4 gamètes, certains évènements de recombinaison ne vont pas modifier la topologie non-racinée de l'arbre, les combinaisons d'allèles restent identiques et aucune incompatibilité n'est générée. Dans les cas où des incompatibilités sont générées, il est nécessaire d'observer des mutations sur les branches permettant de les dévoiler.

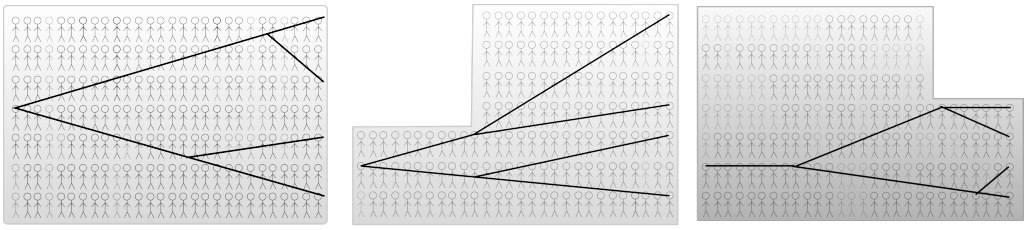
1.2.4 Variations par rapport au modèle neutre

1.2.4.1 Démographie

Changement progressif

Croissance Pour une population en croissance, rétrospectivement, la taille de la population N diminue, la probabilité de coalescence augmente avec le temps, les coalescences sont donc plus rapprochées les unes des autres que dans le cas où la population est constante. L'arbre d'une population en croissance est donc plus petit que l'arbre d'une population constante. Afin de comparer l'arbre d'une population en croissance à celui d'une population constante, il est possible de normaliser les arbres pour qu'ils aient le même T_{MRCA}. Dans ce cas là, l'arbre en croissance a des branches récentes plus longues et des branches ancestrales plus courtes que l'arbre d'une population constante. En effet, proche du temps présent, la taille de la population est plus proche de celle de la population constante tout au long du temps, mais l'arbre final étant plus petit, la mise à la même échelle « tire » sur tout l'arbre et donc également sur les branches terminales.

Décroissance Pour une population décroissante, rétrospectivement, la taille de la population N augmente, la probabilité de coalescence diminue, les coalescences sont de plus en plus éloignées. L'arbre final est donc plus grand que celui d'une population constante. Si on compare un arbre constant et un arbre en décroissance à la même échelle, ce dernier aura des branches ancestrales plus grandes et des branches terminales plus petites.



(a) Constant **(b)** Croissance **(c)** Décroissance

Figure 1.7: Arbres de coalescence de 4 individus suivant 3 scénarios : population constante (a), population en croissance (b) et population en décroissance (c).

Changement soudain Un scénario démographique facile à considérer est le scénario de changement soudain au temps τ, d'une intensité κ de la taille de la population N (Fig 1.7). La généalogie va dépendre de deux tailles de population : la taille de la population au temps présent N_0 et la taille de la population dans le passé (avant le changement de taille) N_∞. On note $\kappa = N_\infty/N_0$.

— Si $\kappa > 1$, la population était plus grande dans le passé, il s'agit d'un scénario de décroissance.

— Si $\kappa < 1$, la population était plus petite dans le passé, il s'agit d'un scénario de croissance.

— Si $\kappa = 1$, la population est constante au cours du temps.

Dans ce modèle, une partie des évènements de coalescence se déroule quand $N = N_0$ et l'autre quand $N = \kappa N_0$.

La déformation de l'arbre dépend de la valeur de κ. Plus κ est différent de 1, plus la différence entre les deux tailles de populations est importante, plus l'arbre de coalescence est déformé. La différence entre un arbre ayant subi un changement soudain et un arbre d'une population constante dépend également du temps auquel a lieu le changement. Pour des temps très récents ou très lointains, aucune différence ne sera observable, tous les temps de coalescence ayant eu lieu avant ou après le temps de changement de taille.

1.2.4.2 Structure

Une population peut être structurée en plusieurs sous-populations, on parle de méta-population. Ces sous-populations (ou dème) peuvent être de tailles différentes

et inter-connectées ou non. La diversité observée dans la population globale dépend de ces paramètres. Dans un modèle typique de population structurée, on considère comme panmictiques les individus d'une même population et on appelle migrants les individus se déplaçant d'une population à une autre.

Deux sous-populations Le cas le plus simple à considérer est celui de deux sous-populations de même taille $N/2$ avec m le taux de migration symétrique entre les deux sous-populations. Le taux de coalescence à l'intérieur de chaque population dépend de $N/2$, il est donc deux fois plus rapide que celui d'une population unique de taille N.

En l'absence de migration $m = 0$, chaque sous-population a un arbre de coalescence indépendant deux fois plus petit que celui d'une population de taille N. Il n'y a pas de coalescence entre deux individus de sous-populations différentes.

Si m est faible, il faut un temps important à deux lignées de deux populations différentes pour coalescer. Il y a donc une grande différence de temps de coalescence si on considère deux lignées du même dème ou de deux dèmes différents. Deux lignées de même dème coalescent rapidement suivant la taille de la population de leur dème. Deux lignées de dèmes différents coalescent plus lentement, leur lignées devant d'abord rejoindre le même dème pour coalescer au sein de ce dème. Leur temps de coalescence, qui sera toujours plus grand que celui de lignées se trouvant dans le même dème, dépend du taux de migration. Échantillonner des individus au sein d'un même dème ou dans des dèmes différents nous apporte des informations différentes. L'allure globale de l'arbre de coalescence dépendra de l'échantillonnage des individus dans les deux populations.

Si m est très élevé, il n'y a quasiment pas de différence entre la population structurée en deux sous-populations et une grande populations contenant tous les individus.

Dans ce modèle particulier, pour $m \neq 0$, l'espérance du temps de coalescence par dème ne dépend pas de la valeur de m. La variance, quant à elle, est très affectée.

Modèle continent-île Un modèle communément étudié est le modèle continent-île. Ce modèle considère que la migration s'effectue entre une grande population « continent » (ou source) vers une population plus petite « île » (ou puits). A chaque génération, la migration s'effectue de la population continent vers la population île et, dans certains cas, de la population île vers la population continent, avec un taux de migration inférieur.

Au sein de l'île, la population étant plus petite, le taux de coalescence est élevé et donc les coalescences sont rapprochées. Dans le continent, la population est beaucoup plus importante, les taux de coalescence beaucoup plus faibles, les coalescences sont donc plus lentes. En échantillonnant les individus dans l'île, l'arbre de coalescence a donc des coalescences récentes : celles des îliens, et des coalescences anciennes : celles des migrants provenant du continent. L'arbre de coalescence d'un modèle continent-île a donc les mêmes caractéristiques (mais pas la même loi) que

celui d'une population en décroissance, malgré l'absence de changement de taille de population. Il existe de nombreux cas où la démographie et la structure ont le même effet sur certaines caractéristiques du coalescent (Mazet et al. 2016).

Population fantôme La présence d'une population non-connue (ou population fantôme) et non échantillonnée, dans la structure d'une population a un effet sur l'arbre de coalescence de la population. En effet, les temps de coalescence des lignées vont dépendre des caractéristiques de la population (taille, taux de migration, durée de la migration etc) comme expliqué précédemment, mais l'échantillonnage est indifférencié et ne permet pas d'identifier les différents facteurs façonnant les arbres de coalescence. Ces effets peuvent ressembler à de la démographie mais également à de la sélection (Marchi and Excoffier 2020).

1.2.4.3 Sélection

Les mutations considérées jusqu'ici sont neutres, elles n'ont pas d'effet sur la généalogie des séquences. Nous allons considérer maintenant que les mutations peuvent être avantageuses ou non, et étudier l'effet sur l'arbre de coalescence.

Balayage sélectif Un allèle fortement avantageux va voir sa fréquence augmenter rapidement dans la population, on parle de balayage sélectif. À ce locus, l'ancêtre commun de la population est très récent, les branches sont donc courtes. Tous les nucléotides liés génétiquement à cet allèle vont voir leur fréquence augmenter également, c'est l'effet d'auto-stop génétique (Smith and Haigh 1974). L'ancêtre commun de ces sites est également récent. Si on échantillonne peu de temps après le balayage, on observera un arbre ressemblant à un arbre correspondant à une croissance de population.

Une forte sélection positive va entraîner une augmentation du nombre de descendants de l'individu sélectionné. Cela peut avoir comme effet de générer des multifurcations dans la généalogie : plusieurs lignées qui coalescent simultanément ou presque (Fig 1.8). Le coalescent de Kingman ne considère que des bifurcations, il faut donc considérer une autre classe de coalescent pour prendre en compte ce phénomène (Schweinsberg 2003; Eldon and Wakeley 2006).

Sélection négative La sélection négative tend à faire disparaître l'allèle de la population. L'individu portant l'allèle a moins de descendants que les autres, ce qui réduit le nombre d'individus à se reproduire, qui a pour effet d'augmenter le taux de coalescence.

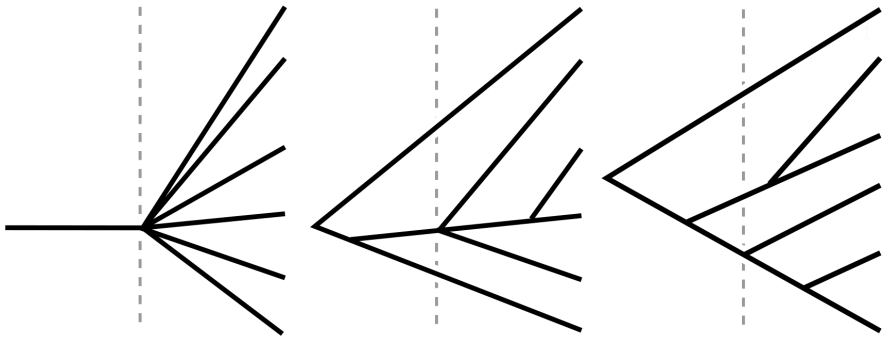


Figure 1.8: Illustration de balayage sélectif. Trois arbres sont représentés, celui d'un locus sélectionné (à gauche), celui d'un locus lié au locus sélectionné (au centre) et celui d'un locus non lié (à droite).

1.3 Inférence d'histoire évolutive

1.3.1 Inférence statistique et optimisation

L'inférence statistique en génétique des populations, est le procédé permettant de déduire les caractéristiques de la population à partir de celles des génomes (complets ou partiels) d'un échantillon d'individus de cette population, en testant des hypothèses et en estimant des paramètres. Elle permet de sélectionner le modèle statistique du procédé qui génère les données et d'en déduire ses paramètres.

La conclusion d'une inférence est une proposition statistique, dont la forme peut être, entre autres : une estimation ponctuelle de paramètre, une estimation d'intervalle de confiance ou le rejet ou non d'une hypothèse.

Généralement, les méthodes d'inférences en génétique des populations sont basées sur des méthodes probabilistes comme la vraisemblance ou l'inférence bayésienne, ou sur des méthodes utilisant des mesures de distance. En génétique des populations, les jeux de données sont de plus en plus larges et complexes, il est difficile de réaliser des inférences rapides et exactes. L'utilisation de méthode d'optimisation pour estimer les paramètres recherchés est également nécessaire.

1.3.1.1 Méthodes probabilistes

Vraisemblance La vraisemblance est une mesure d'adéquation entre un modèle statistique paramétré et des données observées. La vraisemblance est la probabilité (ou la densité de probabilité) des observations sous le modèle, pour un paramètre θ donné. La vraisemblance du modèle θ sur la réalisation x s'écrit $L(x \mid \theta)$. Comme on cherche le θ qui explique le mieux les données, L est une fonction de θ.

Il est courant de calculer la vraisemblance d'un modèle par rapport à un échantillon d'observations indépendantes, $\{x_1, x_2, \ldots, x_n\}$ avec n la taille de l'échantillon, et non pas une seule observation x. La vraisemblance de l'ensemble d'observations

s'écrit comme le produit de la vraisemblance de chaque observation :

$$L(x_1, x_2, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} L(x_i \mid \theta).$$

Numériquement, il est souvent plus pratique d'utiliser le logarithme de la vraisemblance : la log-vraisemblance. Dans le cadre de l'inférence, on cherchera à atteindre le maximum de la vraisemblance, la fonction logarithme étant strictement croissante, la vraisemblance et la log-vraisemblance atteignent leur maximum au même point. Le logarithme du produit des vraisemblances individuelles s'écrit comme la somme des log-vraisemblances :

$$\log L(x_1, x_2, \ldots, x_n \mid \theta) = \sum_{i=1}^{n} \log L(x_i \mid \theta).$$

Il est courant de faire l'hypothèse que les généalogies à différent loci le long du génome sont indépendantes, même si ce n'est pas le cas, comme vu à la section précédente. Un génome serait donc composé d'observations indépendantes de la même histoire évolutive, ces calculs de vraisemblances sont donc applicables. On parle de pseudo-vraisemblance ou de vraisemblance composite.

Maximum de Vraisemblance Le maximum de la fonction de vraisemblance correspond au modèle θ le plus probable d'après les données. Trouver le maximum de la fonction de vraisemblance permet donc d'inférer le modèle et ses paramètres expliquant le mieux la réalisation des données. Si la fonction est connue, dérivable et concave, le maximum de vraisemblance est déterminé facilement. Lorsque le calcul est impossible, il est possible d'utiliser des méthodes numériques, itératives comme la méthode de ascension de gradient qui consiste à estimer le gradient local, suivre sa direction jusqu'à ce que cette dernière change, puis recommencer jusqu'à ce qu'on ne se déplace plus dans l'espace des paramètres. Pour des cas plus complexes, une myriade d'autres méthodes ont été développées.

Test du rapport de vraisemblance Le *Likelihood Ratio Test* (LRT) en anglais, est un test statistique qui permet de comparer l'adéquation de deux modèles statistiques emboîtés par rapport aux données, un modèle doit être compris dans l'autre. Ce test compare les vraisemblances des deux modèles. On compare deux modèles M_0 avec n_0 paramètres et une vraisemblance L_0 et M_1 avec n_1 paramètres et une vraisemblance L_1 avec $n_0 < n_1$, M_0 doit être compris dans M_1. On teste l'hypothèse selon laquelle l'ajout de paramètres n'augmente pas significativement la vraisemblance. On définit la statistique du test comme :

$$\lambda = -2 \log \frac{L_0}{L_1} \quad \text{ou} \quad \lambda = -2(\log L_0 - \log L_1).$$

La statistique du test suit approximativement une loi du χ^2 à $n_1 - n_0$ degrés de libertés. On rejette donc le test avec le risque d'erreur α lorsque la statistique du test est supérieure au quantile d'ordre $1 - \alpha$ de la loi du χ^2 à $n_1 - n_0$ degrés de libertés.

Inférence bayésienne L'inférence bayésienne repose sur l'hypothèse que les estimateurs des paramètres du modèle sont distribués. Elle permet d'intégrer des connaissances à priori sur les valeurs que peuvent prendre les paramètres. On commence donc avec une distribution a priori des paramètres (intégrant par exemple les connaissances antérieures) et l'on cherche la distribution, dite a posteriori, des paramètres conditionnellement aux nouvelles observations, en utilisant le théorème de Bayes. La connaissance de la distribution a posteriori est affinée à chaque itération, à chaque nouvelle observation.

***Markov chain Monte Carlo* (MCMC)** Il s'agit d'algorithmes servant à échantillonner une distribution d'intérêt. Ces chaînes de Markov sont conçues de telle sorte que leur loi stationnaire est la distribution d'intérêt que l'on cherche à échantillonner. Le nombre d'étapes dans la chaîne peut jouer sur la précision de la loi stationnaire échantillonnée.

De nombreux algorithmes existent pour construire la chaîne, un des plus connus est celui de Metropolis–Hastings (Hastings 1970). Cette méthode se base sur deux caractéristiques à définir : une densité de probabilité g pour choisir ces nouveaux pas et une loi afin de rejeter un déplacement (voir Décision si dessous).

Exemple d'algorithme de Métropolis-Hastings Soit $f(x)$, une fonction proportionelle à la distribution d'intérêt :

1. Initialisation. On choisit un point x_0 arbitrairement et une probabilité de transition g afin de choisir le candidat suivant. Par exemple, un tirage uniforme avec une distance maximale par rapport au point précédent.
2. Itérations (t) :
   - Tirage aléatoire du nouveau point x' suivant g.
   - Calcul du taux d'acceptation $\alpha = \dfrac{f(x')}{f(x_t)}$.
   - Décision : Tirage aléatoire d'un nombre uniformément $u \in [0, 1]$, si $u \leq \alpha$ accepter le nouveau point $x_{t+1} = x'$, sinon rejeter le nouveau point et $x_{t+1} = x_t$. Dans le cas où $\alpha > 1$, le point est rejeté sans tirage de u.

1.3.1.2 Méthodes basées sur les distances

Root Mean Square Error La racine de l'écart quadratique moyen (ou *Root Mean Square Error* (RSME) en anglais) est fréquemment utilisée pour mesurer des différences entre des valeurs prédites par un modèle, donc d'estimateurs et les valeurs observées. Ces écarts peuvent être appelés résidus ou erreurs suivant sur quelles données ils sont calculés. Résidus : s'ils sont calculés sur les données qui ont servi à l'estimation, erreurs : s'ils sont mesurés sur d'autres données que celles utilisées pour l'estimation. La RMSE est une mesure de précision qui sert à comparer les erreurs de différents modèles. Elle est toujours positive et une valeur de 0 indique un ajustement parfait entre les données et le modèle. La RMSE d'un

estimateur $\hat{\theta}$ par rapport à la vraie valeur du paramètre θ est définie comme la racine carrée de l'erreur quadratique moyenne :

$$\text{RMSE}(\hat{\theta}) = \sqrt{E((\hat{\theta} - \theta)^2)}.$$

Afin d'éprouver la capacité de différentes méthodes d'inférence à estimer un paramètre, on minimise le RMSE sur un ensemble de méthodes de façon à identifier la « meilleure » méthode d'inférence. Il est également courant d'utiliser la *Mean Square Error* (MSE) :

$$\text{MSE}(\hat{\theta}) = E((\hat{\theta} - \theta)^2).$$

Régression La régression recouvre plusieurs méthodes statistiques permettant d'estimer les relations/dépendances entre une variable à expliquer y et des variables explicatives indépendantes x.

La régression linéaire est une des plus utilisées. Elle consiste à établir des relations *linéaires* entre la variable à expliquer y et les variables explicatives x. Pour chaque individus i, y_i s'écrit comme une combinaison linéaire du vecteur x_i des variables explicatives, avec β le vecteur des paramètres du modèle :

$$y_i = x_i\beta + \varepsilon_i,$$

ε_i représente l'erreur.

Une des méthodes utilisée pour estimer les paramètres du modèle β est la méthode des moindres carrées. Elle consiste à minimiser la somme des carrés des écarts entre chaque point y_i et son projeté, parallèlement à l'axe des ordonnées, sur la droite de régression $\hat{y_i}$.

***Approximate Bayesian Computation* (ABC)** La méthode ABC consiste à rejeter les valeurs des paramètres sous lesquelles les simulations du modèle produisent des échantillons trop éloignés des données observées. Cette notion de distance est quantifiée à partir d'une distance classique entre vecteurs de statistiques résumées obtenues à partir des simulations d'une part et à partir des données d'autre part. Elle a été introduite en génétique des populations par Tavaré et al. 1997 puis généralisée par Pritchard and Rosenberg 1999.

Exemple d'algorithme de rejet issu de Beaumont et al. 2002, considérant un seul paramètre d'intérêt ϕ, ici le taux de mutation, qui sera estimé grâce à S, ici le nombre de polymorphismes ou SNP (*Single Nucleotide Polymorphism*) :

1. choisir une statistique résumée S et mesurer s sur le jeu de données,
2. choisir une tolérance ε,
3. échantillonner ϕ' sous sa distribution a priori,
4. simuler un arbre généalogique sous le modèle choisi : la valeur de ϕ' échantillonnée,
5. simuler les allèles ancestraux présents à la racine de l'arbre et les évènements de mutation le long de l'arbre afin de générer le jeu de données

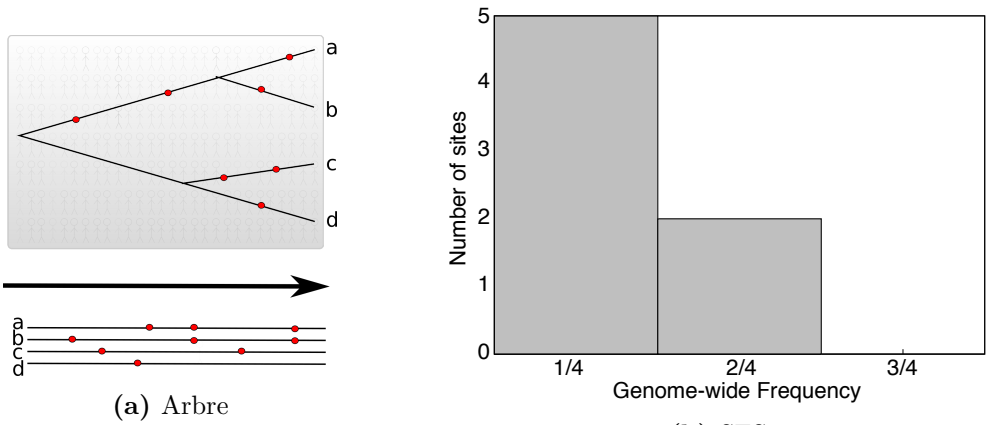
6. mesurer s', la valeur de S pour le jeu de données simulé
7. si $|s' - s| < \varepsilon$ alors accepter ϕ', sinon le rejeter
8. répéter les étapes 3 à 7 jusqu'à avoir accepté k fois.

Cet algorithme peut être utilisé sur un vecteur d'intérêt (dans notre exemple, de statistiques résumée $\vec{S}$) à la place d'une unique variable d'intérêt.

Il existe de nombreuses variantes à cet algorithme, notamment sur la décision de rejet ou sur l'obtention de la distribution a priori. Par exemple, l'ABC peut être combiné à une méthode de Monte-Carlo par chaînes de Markov (MCMC décrit ci-après) : le MCMC-ABC (Marjoram et al. 2003).

1.3.2 Mesures statistiques sur le génome

1.3.2.1 Site Frequency Spectrum



(a) Arbre

(b) SFS

Figure 1.9: Exemple d'un arbre sur lequel se produisent des mutations (a) et du SFS correspondant (b). Cet arbre supporte 4 génomes haploïdes, 7 mutations se produisent, 5 sont portées par un seul individu, 2 sont portées par deux individus, aucune n'est portée par 3 individus.

Définition Le Spectre de Fréquence Allélique, ou *Site Frequency Spectrum* (SFS) représente la distribution des fréquences alléliques des mutations tout le long du génome. Il reporte le nombre de mutations présent à chaque fréquence. Le SFS d'un échantillon de n individus diploïdes est décrit comme un vecteur ξ tel que $\xi = (\xi_1, \xi_2, \ldots, \xi_{2n-1})$, où ξ_i est le nombre de mutations à la fréquence $i/2n$ avec $i \in [1 : 2n-1]$ (on ne s'intéresse pas aux mutations à une fréquence 0 ou 1). La fréquence de mutation se calcule uniquement pour des sites bi-alléliques (avec seulement deux allèles différents), la fréquence est celle de l'allèle dérivé, qui apparaît dans la population après un évènement de mutation. Cependant, pour éviter des erreurs d'orientation entre l'allèle ancestral et l'allèle dérivé, il est possible de travailler avec un SFS dit *replié* (ou *folded* en anglais). Le SFS plié considère la fréquence

de l'allèle le moins fréquent. Il est représenté par le vecteur $\eta = (\eta_1, \eta_2, \ldots, \eta_n)$, où $\eta_i = \xi_i + \xi_{2n-i}$ sauf $\eta_n = \xi_n$, lorsque l'absence d'orientation ne nous permet pas de savoir si un allèle est à la fréquence $i/(2n)$ ou $(2n-i)/(2n)$.

Attendu théorique Sous le modèle standard neutre, les probabilités associées au nombre d'allèles distincts observés un certain nombre de fois dans un échantillon ont été caractérisées par Ewens (Ewens 1972). Le SFS attendu a pu en être déduit (Fu 1995). L'attendu théorique peut être calculé à partir d'une approche de coalescence (Fu 1995) ou de diffusion (Crow et al. 1970). Dans le cadre du modèle standard neutre, le SFS attendu est le suivant :

$$E(\xi_i) = \frac{\theta}{i}$$

avec $\theta = 2N\mu$ le taux de mutation populationnel.

Représentation graphique Pour faciliter la visualisation du SFS et plus précisément de ses écarts au modèle standard neutre, il a été proposé de normaliser les SFS par leur attendu théorique (Nawa and Tajima 2008; Achaz 2009; Lapierre et al. 2017) :

— SFS déplié : $i\xi_i$ pour $i \in [1 : 2n-1]$

— SFS plié : $\frac{i(2n-i)}{2n}\eta_i$ pour $i \in [1 : n-1]$ et $i\eta_i$ pour $i = n$.

Le SFS ainsi transformé sera uniforme en espérance s'il est issu du modèle standard neutre (Achaz 2009).

Tests de Neutralité Il existe de nombreux tests pour vérifier la concordance des données avec le modèle standard neutre s'appuyant sur les fréquences alléliques, et donc sur le SFS. Les tests peuvent avoir des sensibilités différentes en fonction des types de déviations au modèle, influant sur le déséquilibre de l'arbre (Ferretti et al. 2017). La majorité des tests de neutralité se basant sur l'information du SFS compare deux estimations du taux de mutation populationnel θ. Quand le modèle standard neutre est avéré, les valeurs de ces différents estimateurs non biaisés sont égales (Tajima's D (Tajima 1989), Fu and Li's F (Fu and Li 1993), ou les tests de Fay and Wu 2000; Achaz 2008). Les statistiques de ces tests de neutralité peuvent être exprimés comme des combinaisons linéaires du SFS, il est donc possible de générer de nombreux tests de neutralité utilisant l'information du SFS (Achaz 2009).

Impact des différents processus

Démographie Nous avons vu précédemment que la démographie d'une population avait un impact sur la topologie de l'arbre de coalescence. Plus une population est grande, plus un arbre a de longues branches et inversement. Une population en croissance a des branches terminales plus courtes et une population en décroissance des branches terminales plus longues (en comparaison avec une population

constante). Nous avons également vu que le nombre de mutations dépend de la taille de la branche, comparativement, une population en croissance a donc moins de mutations sur les branches terminales et une population en décroissance en a plus. Les branches terminales supportent moins d'individus, les mutations localisées sur ces branches sont donc en plus faible fréquence dans la population.

Le SFS d'une population en croissance a donc un déficit de mutations à faibles fréquences et un excès de mutations à fortes fréquences. Inversement, une population en décroissance a un excès de mutations à faibles fréquences et un déficit de mutations à fortes fréquences.

Structure La structure de la population ayant aussi un effet sur la topologie de l'arbre de coalescence, elle a un effet sur le SFS. Les structures de population peuvent prendre des formes plus ou moins complexes, influençant la généalogie de la population de manière différente, il existe donc de nombreuses déformations du SFS dû à la structure de la population. Comme dit précédemment, la structure et la démographie peuvent avoir le même effet sur certaines caractéristiques du coalescent, notamment le SFS. Une même déformation de la topologie et des longueurs de l'arbre, entraine le même patron de mutations et donc le même SFS.

Recombinaison Le SFS d'un alignement multiple ne correspond pas à un arbre de coalescence mais à un ensemble d'arbres de coalescence entrecoupés par des évènements de recombinaison, formant une tree sequence. Ainsi, le SFS de l'échantillon est, en réalité, le spectre moyen de toutes les généalogies.

S'il y a de nombreuses recombinaisons entre deux sites, ces derniers peuvent être considérés comme statistiquement indépendants. Si c'est le cas pour la majorité des sites, le SFS mesuré est celui de la moyenne des arbres échantillonnés, qui correspond, sous le modèle neutre à l'attendu théorique (Adams and Hudson 2004). La liaison entre certains sites entraînent une absence d'indépendance statistique, des sites liés contiendront moins d'information que des sites non liés, rendant l'inférence plus difficile. Cependant, la recombinaison sous le modèle neutre a très peu d'effet sur le SFS attendu, seule sa variance diminue (Wall 1999).

En l'absence de recombinaison, d'autres méthodes sont utilisables pour inférer la démographie d'une population (Beerli and Felsenstein 2001; Kuhner et al. 1998; Nielsen 2000).

Sélection Quand un locus est sous sélection positive, sa fréquence augmente. Les locus qui lui sont physiquement liés vont voir leur fréquence augmenter également. En l'absence de recombinaison, les allèles liés au locus sélectionné vont se fixer dans la population en même temps que l'allèle sélectionné. La recombinaison casse la liaison entre les sites, la fréquence de certains sites augmentera jusqu'à ce qu'un évènement de recombinaison brise la liaison physique du site avec le locus sélectionné. Cela va donc avoir un effet sur le SFS avec un excès de mutations à fortes fréquences.

Conversion génique biaisée Lors d'une erreur d'appariement de nucléotides (par exemple entre G et T), le système de réparation privilégie la réparation par un couple d'allèle avec une liaison forte à trois hydrogènes (C—G) plutôt qu'un couple avec une liaison faible à deux hydrogènes (A—T), ce phénomène s'appelle le biais de conversion génique (Duret and Galtier 2009). Cela entraîne une conversion d'allèles à liaison faible en allèles à liaison forte et provoque un excès de mutations à fortes fréquences. Dans le cadre du SFS, il est possible de corriger ce biais en considérant seulement les sites bi-alléliques composés des couples d'allèles A$\longleftrightarrow$T et C$\longleftrightarrow$G, ces derniers ne peuvent pas être produit par une réparation biaisée (Pouyet et al. 2018).

Méthodes d'inférence Il existe plusieurs types de méthodes utilisant les SFS pour faire de l'inférence d'histoire de population.

Méthodes *paramétriques* Dans les méthodes paramétriques, il est nécessaire de renseigner un modèle de population dont les paramètres seront estimés pour générer un SFS le plus proche de l'observé. Le SFS observé peut être comparé au SFS estimé à partir de simulations Monte Carlo d'arbres de coalescence pour le SFS d'une (Nielsen 2000; Coventry et al. 2010) ou plusieurs populations (Excoffier et al. 2013). D'autres approches existent utilisant la densité de SNP en fonction du temps (Gutenkunst et al. 2009).

Méthodes *non-paramétriques* Les méthodes non-paramétriques ne demandent pas de modèle sous-jacent à la population. L'expression exact de l'espérance du SFS d'une population constante, ou exponentielle, par morceaux est connue (Bhaskar et al. 2015), des méthodes infèrent une démographie constante (ou exponentielle) par morceaux à partir du spectre (Liu and Fu 2015, 2020), sans aucune connaissance préalable.

Ces méthodes ne demandent pas de modèle sous-jacent, mais ce n'est pas pour cela qu'elles sont sans contraintes. En effet, considérer que la démographie d'une population peut être résumée par un certain nombre de morceaux de taille constante est une hypothèse forte et peut être très éloignées de la véritable histoire. Par exemple, si la population a une croissance exponentielle, il sera très difficile de représenter son histoire à partir de morceaux de taille constante.

1.3.2.2 Linkage Disequilibrium

Définition Le Déséquilibre de Liaison ou *Linkage Disequilibrium* (LD) en anglais, représente la dépendance entre des allèles à des sites différents. En effet, les allèles ayant des liens physiques sur le génome peuvent être transmis plus souvent ensemble que s'ils étaient indépendants. Le LD s'intéresse donc à l'association statistique entre des allèles à des sites différents, et mesure l'écart de cette association à une association aléatoire, c'est une covariance.

Pour calculer la covariance D entre deux loci bi-alléliques A/a et B/b, en considérant A et B comme étant les allèles dérivés et f_A et f_B étant respectivement les

fréquences de A et de B et f_{AB}, la fréquence de la co-occurrence AB, D est défini par : $D_{AB} = f_{AB} - f_A f_B$.

De nombreuses autres statistiques mesurent le LD, la grande majorité s'intéressent au LD entre deux loci. Les plus communément utilisées sont r^2 et $|D'|$, des mesures de D normalisé.

La mesure r^2 (Hill and Robertson 1968) est le coefficient de corrélation au carré (corrélation de Pearson) entre les deux fréquences :

$$r^2 = \frac{D^2}{f_A f_a f_B f_b}.$$

La valeur de r^2 est comprise entre 0 et 1. La valeur 0 indique une absence de corrélation, ce qui correspond à la présence d'évènements de recombinaison cassant la liaison ; et 1 indique une complète corrélation, une forte liaison entre les deux sites. La mesure r^2 a la même valeur peu importe la qualification (ancestrale ou dérivée) de chaque allèle et varie avec l'histoire évolutive (McVean 2002). Le r^2 est notamment utilisé lors des études d'association pangénomique (ou *Genome-Wide Association Study* GWAS en anglais).

La mesure |D'| (Lewontin 1964) est définie comme la valeur absolue du ratio entre le D observé et la plus extrême valeur qu'il pouvait prendre sachant les fréquences alléliques individuelles :

$$|D'| = \begin{cases} \dfrac{-D_{AB}}{\min(f_A f_B, f_a f_b)} & D_{AB} < 0 \\ \dfrac{D_{AB}}{\min(f_A f_b, f_a f_B)} & D_{AB} > 0 \end{cases}$$

L'utilisation principale de $|D'|$ est la mise en évidence d'un évènement de recombinaison. $|D'|$ est compris entre 0 et 1. $|D'|$ ne peut être inférieur à 1 que si les 4 combinaisons d'allèles sont présentes (comme dans le cadre du Test des 4 gamètes de Hudson and Kaplan 1985). La mesure $|D'|$ a donc sa valeur maximale (égale à 1) quand il n'y a que 3 combinaisons ou moins, c'est-à-dire que les deux sites sont compatibles avec le même arbre. Pour des valeurs de $|D'|$ proches de 0 plusieurs évènements de recombinaison se sont sans doute produits.

Il est important de noter qu'une valeur de $|D'|$ égale à 1 ne veut pas dire qu'il y a pas eu d'évènements de recombinaison. En présence d'haplotypes à faible fréquence, les 4 combinaisons peuvent ne pas être observées malgré un évènement de recombinaison. Deux sites peuvent ne pas être corrélés, avoir un r^2 proche de 0, et ne pas comptabiliser les 4 combinaisons, et donc avoir un $|D'|$ égal à 1. Ces deux mesures du LD peuvent donc être contradictoires (McVean 2008).

Impact des différents processus Le LD étant mesuré à partir des fréquences des mutations et de leur association, il dépend de la topologie de l'arbre et des positions des évènements de mutations le long de cet arbre. Le LD dépend donc des différents processus ayant un effet sur la topologie de cet arbre.

Démographie Même dans des régions sans recombinaison, le LD va dépendre de la démographie (Slatkin 1994). En effet, si deux mutations se produisent sur la même branche, elles vont être présentes chez les mêmes individus et donc être en LD « parfait » ($r^2 = 1$). Si deux mutations se produisent sur des branches différentes, elles ne seront que peu corrélées, surtout si ces branches sont dans des parties différentes de l'arbre et encore plus si une mutations, ou les deux sont sur des branches terminales ($r^2 << 1$). Le LD dépend donc fortement de la forme de l'arbre de coalescence. Un scénario de décroissance forte, produit un arbre dominé par des longues branches ancestrales, sur lesquelles se produisent des mutations en LD parfait. Au contraire, un scénario de croissance produit un arbre avec de longues branches terminales, sur lesquelles se produisent des mutations présentes chez seulement un individu, aux faibles corrélations entre elles (McVean 2008).

Structure Le LD est affecté par la structure de l'arbre de coalescence, s'il est composé de différents sous-groupes bien distincts, le LD sera hétérogène avec des groupes possédant des allèles très liés et des groupes ne comprenant aucune association d'allèles identiques. Le LD est affecté par la structure de population, par les évènements d'admixture. La structure a tendance à faire augmenter les valeurs de LD, car il peut y avoir une association forte entre des sites non liés génétiquement, mais étant issus d'individus liés car provenant des mêmes sous-populations (Pritchard and Przeworski 2001).

Recombinaison La recombinaison a comme effet de casser la liaison entre les sites et donc faire diminuer le LD. La variation du taux de recombinaison le long du génome a donc un fort effet sur le LD, créant des zones où le LD est plus ou moins fort.

Plus deux sites sont éloignés sur le génome, plus il y a eu d'évènements de recombinaison entre eux et plus leur LD sera faible. La relation entre r^2 et la distance génétique au niveau de la population (en fonction de la taille de la population et du taux de recombinaison), a été approximée analytiquement (Sved 1971; Ohta and Kimura 1971). Le LD peut donc être utilisé pour inférer le taux de recombinaison d'une population ou la carte de recombinaison d'un chromosome (Consortium 2003).

Sélection Un autre processus affectant le LD est la sélection. En effet, la présence d'allèles avantageux peut entraîner des effets de *hitchhiking* génétique. Tout un groupe de sites liés voit sa fréquence augmenter grâce à l'allèle sélectionné. Le LD à l'intérieur de ce groupe sera très élevé. Cependant, tous les loci ne sont pas affecté par la sélection (Slatkin 2008).

Méthode d'inférence

Loci sous sélection Le LD est donc très utilisé pour identifier des loci sous sélection. En comparant les variations de LD à celles attendues sous un modèle standard neutre, il est possible d'identifier des loci possiblement sous balayage sélectif, ce travail a notamment été fait chez l'homme (Sabeti et al. 2006). Une classe de méthode permet d'étudier l'adéquation d'un modèle précis de balayage sélectif avec la variation génétique observée (Coop and Griffiths 2004; Kim and Stephan 2002; Nielsen 2005; Przeworski 2003).

Genome-Wide LD L'étude des variations de LD le long du génome permet de mettre en avant d'autres phénomènes comme la présence de goulot d'étranglement (Schmegner et al. 2005; Zhang et al. 2004) et même des trajectoires démographiques plus complexes (Santiago et al. 2020).

Le LD entre des sites situés à une grande distance a permis, quant à lui, de mettre en évidence des évènements d'admixture passés. Un évènement unique d'admixture entre deux populations possédant des fréquences alléliques différentes génère du déséquilibre de liaison. Au moment de l'admixture, il peut exister du déséquilibre de liaison entre deux loci non liés (Pritchard and Rosenberg 1999). Le déséquilibre s'atténuera au fil des générations.

1.3.2.3 Segments *Identical By Descent* et *Runs Of Homozygosity*

Définitions

Identity By Descent

> Les segments *Identical By Descent* ont été défini par les mathématiciens comme pouvant être partagés par n individus. Dans les faits, seuls les segments partagés par deux individus sont étudiés. Nous nous restreignons donc au cas de segments partagés par deux individus.

Deux segments homologues d'ADN sont dits Identiques par Descendance (*Identical By Descent* (IBD) en anglais), lorsque qu'ils sont hérités d'un même ancêtre sans évènement de recombinaison les séparant. Deux segments homologues sont dits Identiques par État (*Identical By State* (IBS) en anglais), lorsqu'ils sont parfaitement identiques à tous les sites (homozygotes). Un segment IBD ne sera pas IBS si une mutation est apparue plusieurs fois chez des individus différents ou qu'il y a eu un évènement de recombinaison sans effet sur l'haplotype. La distribution des longueurs de segments IBD passés au cours des générations a été étudiée (Stam 1980; Chapman and Thompson 2003; Stefanov 2000). La fréquence des recombinaisons dépend du nombre de générations écoulées, donc de la taille de l'arbre de coalescence et de la date de l'ancêtre commun. Plus l'ancêtre commun est récent, moins il y a d'évènements de recombinaison est plus le segment IBD sera long. Inversement, plus l'ancêtre commun était lointain, plus le segment IBD est court.

Runs Of Homozygosity Les *Runs Of Homozygosity* (ROH), sont des séquences d'homozygotie, ils peuvent être définis comme deux séquences contiguës IBS sans mutations sur les chromosomes homologues d'un individu. Les deux copies proviennent d'un ancêtre commun aux deux parents, ils sont donc également IBD (Gibson et al. 2006). Comme pour les segments IBD, les longs ROH sont hérités d'ancêtres plus récents et les courts ROH sont hérités d'ancêtres plus lointains.

Les ROH peuvent s'apparenter à une mesure de la consanguinité. Le degré de consanguinité F est la probabilité qu'un individu reçoive deux allèles IBD à un locus donné (Wright 1922). Cela correspond à la proportion du génome autozygote, la somme des segments ROH (McQuillan et al. 2008).

Impact des différents processus La distribution populationnelle des segments IBD sans mutations et ROH va dépendre de l'histoire évolutive de la population, les deux types de segments dépendent des évènements de recombinaison se produisant sur l'arbre de coalescence. Les segments IBS sans mutations et ROH covarient positivement.

Démographie La taille des segments IBS sans mutations et ROH dépend de la fréquence des évènements de recombinaison au niveau de la population. Plus une population est grande, plus le temps écoulé entre l'ancêtre commun et les individus au temps présent est long, plus il y a d'évènements de recombinaison se déroulant pendant cette période, plus les segments sont petits. Dans une population petite, la consanguinité est plus forte, les ROH seront donc plus longs.

Structure Les populations issues de mixage, auront des ROH plus petits que les populations ancestrales. Dans le cas d'un modèle continent-île, les individus de l'île auront des plus longs ROH que ceux du continent. L'effet de la structure sur les segments dépend donc du scénario considéré.

Recombinaison La recombinaison a un effet très direct sur la taille des segments, plus le taux de recombinaison est élevé plus les segments sont petits et inversement. Un taux de recombinaison variable changera donc la distribution de ces segments.

Les hotspots de recombinaison sont des portions du génome où le taux de recombinaison est beaucoup plus élevé que la moyenne. Dans ces endroits, les segments sont donc très fins et rarement détectables.

Sélection Par l'effet du hitchhiking, la sélection positive tend à allonger les segments non recombinés. Cet effet reste cependant localisé dans le temps et sur le génome aux balayages sélectifs récents.

Méthodes

Détection Malgré le fait que les limites des blocs IBD ne sont généralement pas observables lors de la comparaison de deux régions homologues, des blocs IBD *assez longs* ($\geq$ 1cM) peuvent être extraits en appliquant l'une des méthodes suivantes à une paire de séquences (Purcell et al. 2007; Gusev et al. 2009; Browning and Browning 2010). Ces méthodes reposent par exemple la détection de longs segments partagés identiques (Gusev et al. 2009). Certaines utilisent les régions partagées qui abritent de multiples variants rares (Purcell et al. 2007; Browning and Browning 2010). En effet, si deux individus partagent le même variant rare, ils peuvent également partager la région chromosomique environnante, en particulier parce que les variants plus rares sont les plus susceptibles d'être relativement récents. Comme la plupart de ces méthodes prennent en compte les erreurs de séquençage, les segments IBD inférés peuvent avoir de séquences pas complètement identiques. La précision de la détection des blocs IBD dépend de l'algorithme utilisé (Browning and Browning 2013). Ces méthodes conçues pour détecter des blocs IBD, sont applicables à la recherche des segments ROH au sein d'un même individu (Ceballos et al. 2018).

Inférence Certaines méthodes d'inférence démographique sont basées sur la distribution des longueurs de blocs IBD par paire de génomes échantillonnés dans la population. Palamara et al. (2012) calculent la distribution des longueurs attendues des blocs IBD pour un modèle démographique paramétré donné. Browning et Browning (2015) calculent le temps attendu jusqu'à l'ancêtre commun le plus récent (TMRCA) d'un bloc IBD en fonction de sa longueur. Ensuite, ils utilisent la densité empirique des longueurs des blocs IBD pour estimer la distribution de TMRCA et donc les variations de la taille effective de la population dans le temps. D'autres méthodes utilisent la longueur des ROH (Hayes et al. 2003; MacLeod et al. 2009). Des outils ont été développés pour appliquer ces méthodes afin d'en déduire une inférence démographique à partir de données génomiques (MacLeod et al. 2013; Harris and Nielsen 2013).

1.3.2.4 Inverse Instantaneous Coalescence Rate

Définition Le *Inverse Instantaneous Coalescence Rate* (IICR) est la fonction inverse du taux de coalescence de deux lignées en fonction du temps. Il a été introduit par Mazet et al. 2016 dans le cadre d'une étude de $f_{T_2}(t)$, la densité de probabilité des temps de coalescence pour un échantillon de $n = 2$ (T_2) :

$$IICR(t) = \frac{P(T_2 > t)}{f_{T_2}(t)}$$

où t est exprimé en N générations. L'IICR est l'équivalent de la taille de population dans un régime panmictique, mais va être influencé par les différents processus affectant le taux de coalescence.

Impact des différents processus L'effet des différents processus sur les temps de coalescence est déjà décrit dans une partie précédente.

Dans un modèle de type Wright-Fisher en temps continu et à taille de population constante N le taux de coalescence est inversement proportionnel à la démographie de la population, le IICR est donc la démographie de la population.

Dans le cas d'une population structurée, l'IICR présente une tendance de population en déclin, entièrement due à de la structure de population (Mazet et al. 2016).

Suivant les cas, l'IICR va être plus ou moins influencé par la structure de la population ou par sa démographie. En comparant les distributions d'IICR de différents individus d'une même population, il peut être possible de différencier un scénario de démographie d'un scénario de structure (Chikhi et al. 2018). Pour des échantillons plus grands, la distribution conjointe des événements de coalescence $[T_2, T_3, \dots]$ peut être utilisée, en théorie, pour démêler la structure de la démographie (Grusea et al. 2019).

Méthodes d'inférence L'estimation de l'IICR est utilisée dans de nombreuses méthodes d'inférence. Dans PSMC, Li et Durbin (2011), utilisant le même principe que le SMC de McVean et Cardin (2005), ont conçu un modèle de Markov caché (HMM) qui déduit l'IICR à partir des positions de sites hétérozygotes le long d'une paire de séquences, puis estime un modèle démographique constant par morceaux. MSMC, l'extension de PSMC (Schiffels and Durbin 2014), utilise le temps de la première coalescence entre n individus, il a besoin de données phasées.Ces méthodes sont gourmandes en calcul (à ce jour, MSMC ne peut pas déduire l'histoire démographique de plus de 10 individus) et regroupent la diversité sur des fenêtres de 100 pb, supposées former un seul locus à deux états, hétérozygote ou homozygote. Le MSMC2 (Malaspinas et al. 2016) permet d'utiliser l'information portée par plus individus en mesurant l'IICR entre chaque paires de génome haploïde.

Chapitre 2

Utilisation des polymorphismes et des segments non recombinés pour inférer la démographie passée

Contents

2.1 Inférences de changement de taille de population à partir de SFS

Ce travail a été réalisé dans le cadre du stage de Master 2 de Pierre Imbert intitulé *Estimation de l'histoire démographique de populations à partir de données génomiques.*

2.1.1 Motivation

Le SFS est communément utilisé pour étudier l'histoire évolutive d'une population (Section Inférence d'histoire évolutive), notamment pour son histoire démographique : les différentes tailles par lesquelles la population est passée. Des logiciels utilisant des méthodes d'inférences différentes (Gutenkunst et al. 2009; Excoffier et al. 2013; Liu and Fu 2015, 2020) ont été développés pour inférer de nombreuses histoires démographiques.

Cependant, la gamme de scénario et ses paramètres pour lesquels la déformation du SFS peut mettre en évidence des changements d'histoires de population est peu connue. Nous avons ici, étudié la déformation du SFS sous le scénario d'un unique changement soudain de taille de population ainsi que la possibilité de détection du changement et d'estimation des paramètres de ce changement de deux logiciels utilisant l'information du SFS : $\partial a \partial i$ (Gutenkunst et al. 2009) et Stairway Plot 2.0 (Liu and Fu 2020).

2.1.2 Matériel et méthode

2.1.2.1 Modèle de changement de taille de population

L'ensemble des SFS utilisés lors de cette étude a été généré avec le simulateur msprime (Kelleher et al. 2016). Chaque simulation comporte 20 génomes haploïdes, contenant au moins 100,000 SNPs, avec un taux de recombinaison égal au taux de mutation ($\rho = \mu = 0.08$). Le scénario considéré a été le suivant : la population est de taille actuelle N_0 identique pour toutes les simulations, un changement soudain de taille de population d'intensité κ arrive un temps τ dans le passé, exprimé en N générations. La taille de la population ancestrale N_∞ est κN. Pour un temps t :

— si $t < \tau$, $N = N_0$,

— si $t > \tau$, $N = N_\infty = \kappa N_0$.

L'intensité κ détermine le sens du changement de taille de la population.

— Si $\kappa = 1$, il n'y a pas de changement de taille de population.

— Si $\kappa < 1$, la population était plus petite dans le passé, il s'agit donc d'un scénario de croissance de population.

— Si $\kappa > 1$, la population était plus grande dans le passé, il s'agit donc d'un scénario de décroissance de population.

2.1.2.2 Méthode

Distance Pour comparer les SFS simulés (SFS^{sim}) sous le scénario de changement de taille de population avec l'attendu théorique (SFS^{th}) (Fu 1995), un calcul de distance pondérée entre les deux SFS est effectué, d'après la formule suivante :

$$\chi^2(\mathrm{SFS}^{th}, \mathrm{SFS}^{sim}) = \sum_{i=1}^{n-1} \frac{(\mathrm{SFS}_i^{th} - \mathrm{SFS}_i^{sim})^2}{\mathrm{SFS}_i^{th}},$$

avec n le nombre de génomes simulés, et i la i^{eme} case du SFS.

La comparaison des paramètres estimés par rapport aux paramètres simulés dans le cadre des méthodes d'inférence se fait aussi par distance pondérée au carré :

$$d^2(\text{simulé}, \text{estimé}) = \frac{(\text{simulé} - \text{estimé})^2}{\text{simulé}}.$$

Test du rapport de vraisemblance Dans le cadre de notre étude, les logiciels considérés utilisent la vraisemblance afin d'inférer le meilleur modèle en fonction des données. Pour tester ces logiciels nous comparons la vraisemblance du modèle de changement de taille unique de la population (M_1 avec la vraisemblance L_1) à celui d'une population constante (M_0 avec la vraisemblance L_0). Le *Likelihood Ratio Test* (LRT), se calcule de la manière suivante : $\text{LRT} = 2\,[log(L_1) - log(L_0)]$. Pour *∂a∂i* comme pour Stairway Plot, il y a deux degrés de liberté. Le risque d'erreur choisi est de 5%.

2.1.2.3 Logiciels

Stairway Plot 2 (Liu and Fu 2020) est une méthode dite *non-paramétrique*, il n'est pas nécessaire de renseigner le modèle sous-jacent à l'histoire de la population. La méthode utilisée par Stairway Plot pour l'inférence est un modèle multi-époques fondé sur le skyline plot (Drummond et al. 2005). Elle consiste à inférer une démographie constante par morceaux à partir du SFS observé, par maximum de vraisemblance composite. A chaque tour d'inférence, le nombre de blocs constants (ou dimensions) est augmenté de 1, la taille et la longueur du bloc sont inférées puis la vraisemblance totale du scénario est mesurée puis comparée à celle du scénario précédent (avec une dimension de moins). L'inférence s'arrête quand le gain en vraisemblance n'est pas statistiquement significatif.

Le modèle de population constante (M_0) est donc le modèle à une dimension et le modèle à un changement brutal de population (M_1) est celui à deux morceaux. Nous avons tout de même gardé le nombre de dimensions final pour chaque inférence. Le logiciel Stairway Plot ne demandant pas de paramètres à estimer pour inférer la démographie, nous avons estimé ces paramètres de la manière suivante. L'intensité du changement κ est le rapport entre la taille de population la plus grande et la plus petite (ou l'inverse suivant laquelle est la plus proche du temps présent, afin de déterminer s'il s'agit d'une croissance ou d'une décroissance). La date de changement de taille τ est le temps moyen entre ces deux tailles de population. Les paramètres par défaut, conseillés dans la notice du logiciel, ont été utilisés pour l'optimisation.

∂a∂i (Gutenkunst et al. 2009) utilise une méthode dite *paramétrique* pour inférer une démographie. Il est nécessaire de renseigner un modèle de population simplifié dont *∂a∂i* va inférer les paramètres. *∂a∂i* va donc optimiser les paramètres du modèle simplifié en simulant des SFS et en mesurant une log-vraisemblance (les sorties du logiciel sont le SFS simulé, l'estimation des paramètres et la log-vraisemblance).

On peut noter que le modèle simplifié n'est pas forcément un modèle démographique, il peut s'agir de populations structurées. Il est cependant indispensable de définir le modèle sous-jacent.

Nous testons donc deux modèles : celui d'un changement brutal de taille de population et un modèle de population constante.

2.1.3 Résultats

2.1.3.1 Comparaison SFS et attendue théorique

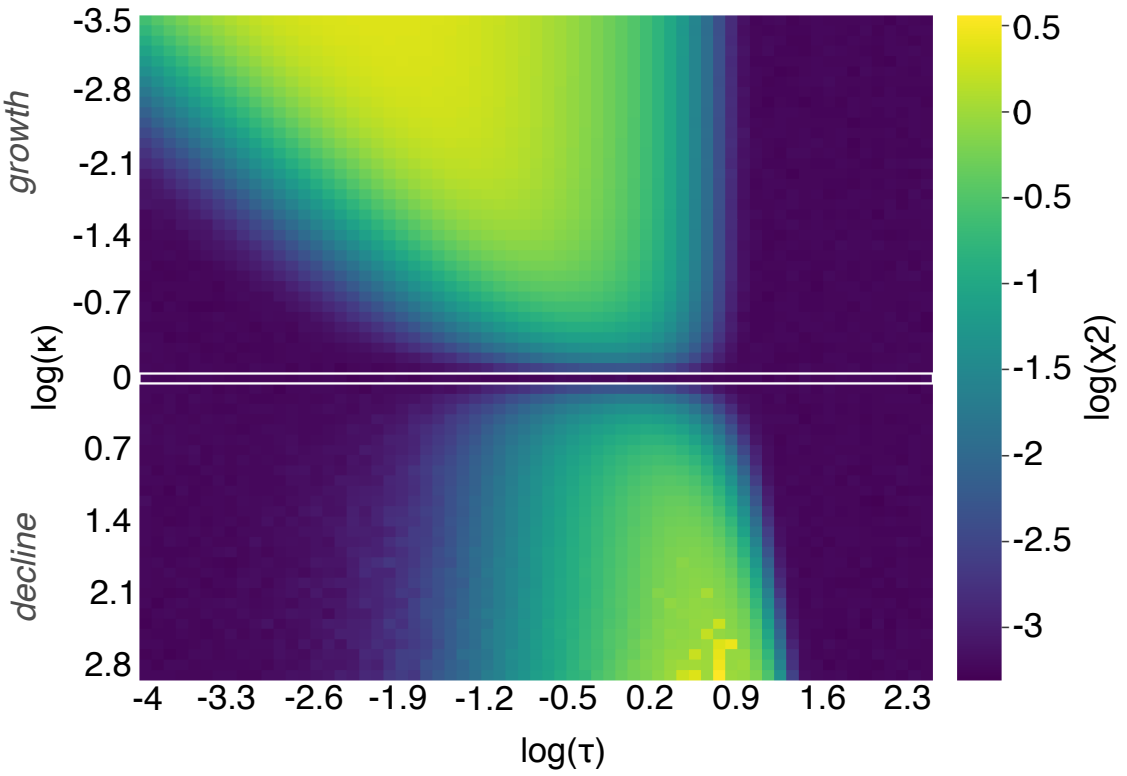


Figure 2.1: Distance entre le SFS d'un changement de taille de population soudain et le SFS théorique sous modèle standard neutre, en fonction de la date du changement τ (abscisses) et de l'intensité du changement κ (ordonnées). Le bleu foncé représente des SFS très similaires ($\chi^2 = 0.001$) et le jaune des SFS très différents ($\chi^2 = 3$). La ligne encadrée en blanc représente le population constante, au dessus se situe les populations croissantes et en dessous les populations décroissantes.

En observant la distance entre le SFS ayant subi un changement de taille de population et le SFS théorique sous le modèle standard neutre (Figure 2.1), plusieurs observations sont à faire.

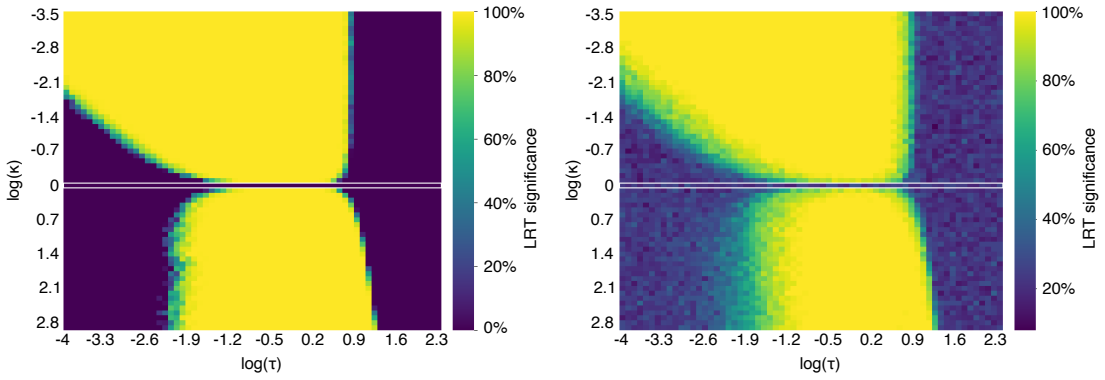
Premièrement, à partir d'un certain temps ($\tau = 10$), il n'est plus possible de voir une différence entre les SFS ayant subi un changement de taille de population et le SFS théorique constant. En effet, ces changements de taille sont tellement tardifs que le temps de l'ancêtre commun est plus récent que le changement de taille (et ce peu importe le nombre d'arbre échantillonnés). Le changement ne pourra donc jamais être détecté.

Deuxièmement, il y a un forte asymétrie entre une croissance de population et un déclin de population. En effet, la différence entre les SFS en déclin et le SFS constant est plus faible et disparaît plus rapidement pour les temps récents que pour les SFS en croissance. Cette différence peut se comprendre de la manière

suivante : la taille initiale est identique pour tous les scénarios, le taux de mutation également, pour une même date de changement, peu importe la taille précédente de la population, il y aura en moyenne le même nombre de mutations présentes sur la partie à taille N_0 de l'arbre. Dans le cas d'une décroissance, la population était plus grande avant le changement, cela augmente la taille totale de l'arbre et donc le nombre de mutation. Les mutations présentes dans la partie récente de taille N_0 sont donc « noyées » dans celles de la partie de grande taille. L'information de cette taille N_0 est donc plus difficilement accessible, et il n'y a pas de différence observable entre le SFS ayant subi un déclin et le SFS constant. (Le nombre total de mutation n'est, quant à lui, pas concordant avec une population constante de taille N_0.) Dans le cas d'une population en croissance, l'inverse se produit. L'arbre étant réduit, moins de mutations se produisent sur cette partie de l'arbre, rendant la partie plus proche du présent plus importante quantitativement. Plus la croissance est importante, plus la partie ancienne est plus petite et plus la partie récente prend de l'importance. C'est ce qui explique que pour une croissance très importante, une différence est toujours marquée entre le SFS en croissance et le SFS théorique alors que pour la même date de changement il n'y a pas de différence pour une croissance faible.

2.1.3.2 Détection et estimation par Stairway Plot 2 et $\partial a \partial i$

L'existence de différence entre un SFS ayant subi un changement de taille de population brutal et un SFS constant ayant été démontrée, il est intéressant de déterminer si les logiciels sélectionnés sont capables de détecter ce changement et d'estimer ses paramètres.



(a) Stairway Plot 2 **(b)** $\partial a \partial i$

Figure 2.2: Significativité du test de rapport de vraisemblance entre le modèle constant et le modèle avec un changement de taille de population (donc deux tailles différentes pour l'histoire de la population) : en jaune significativité de 100% en bleu foncé de 0%, en fonction de la date de changement (τ) (abscisse) et de l'intensité du changement κ (ordonnée), pour deux logiciels différents : Stairway Plot 2 (a) et $\partial a \partial i$ (b).

Détection Les deux logiciels d'inférence utilisant le SFS sont capables de détecter un changement de taille de population pour toutes les valeurs auxquelles une différence est remarquable entre les SFS (Fig 2.2). La seule différence entre les deux méthodes se situe à leur frontière de significativité. Stairway Plot 2 passe d'une significativité de 100% à une significativité de 0% très rapidement (Fig 2.2a), plus rapidement que $\partial a \partial i$ et perd le 100% de significativité plus tard que $\partial a \partial i$. $\partial a \partial i$ a une frontière plus épaisse et passe par des valeurs intermédiaires (Fig 2.2b). De ce point de vue, Stairway Plot semble plus efficace que $\partial a \partial i$.

Estimation du nombre de dimension pour Stairway Plot Stairway Plot étant une méthode libre, le logiciel infère le nombre de tailles différentes par lesquelles la population est passée, le nombre de morceaux. Stairway Plot était le meilleur logiciel pour détecter une population non constante. Cependant, comme on peut l'observer dans la Figure 2.3, il a tendance à surestimer le nombre de changement, allant jusqu'à inférer 12 tailles de populations à la place des deux simulées. Plus la différence est marquée entre le SFS de population changeante par rapport au constant théorique, plus le signal peut être considéré comme fort mais plus le logiciel infère de dimensions. Stairway Plot est donc le logiciel qui détecte le mieux un changement, mais il est également celui qui surestime beaucoup les changements, soucis déjà montré par Lapierre et al. 2017 pour la version précédent du logiciel.

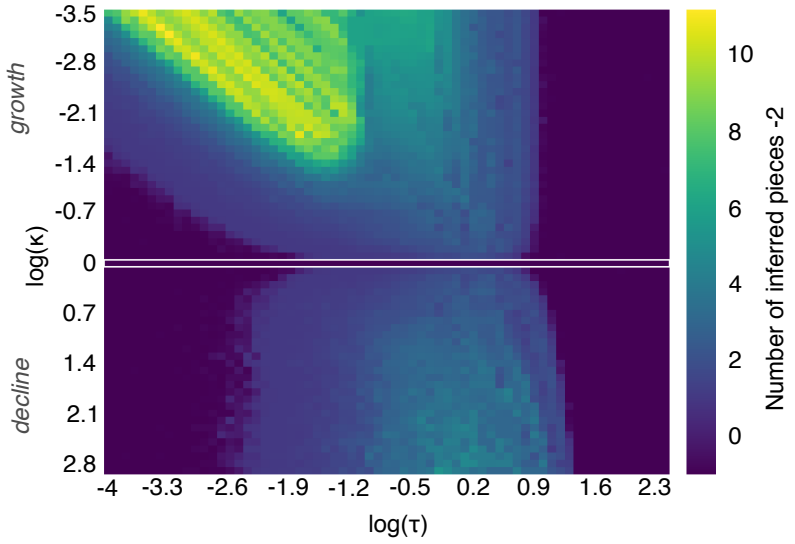
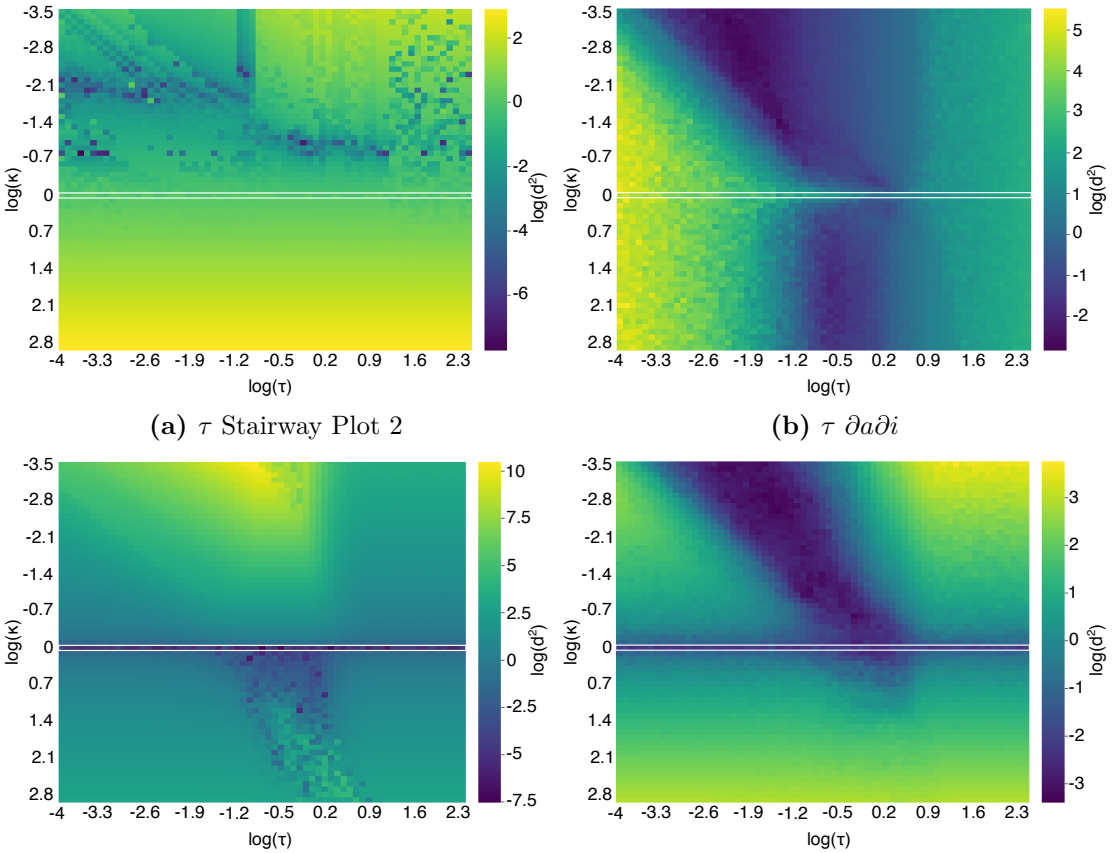


Figure 2.3: Écart entre le nombre de morceaux inféré par Stairway Plot et le nombre de morceaux réel (2), allant du même nombre de dimension (bleu) à 10 dimensions d'écart (jaune), en fonction du couple τ (abscisse) κ (ordonnée) simulé.

Estimation des paramètres de changement de taille Il est naturel que pour tous les SFS n'ayant aucun signe de changement et n'étant pas détecter comme tel, l'écart entre les paramètres inférés et ceux simulés soit important.

Dans le cas de Stairway Plot, il existe une grande disparité entre les paramètres estimés et les simulés, même pour les gammes de paramètres où le changement de taille à été détecté, et même pour les couples pour lesquels le bon nombre de dimensions est estimé aussi bien pour τ (Fig 2.4a) que pour κ (Fig 2.4c).

Le logiciel $\partial a \partial i$ est, quant à lui, capable d'estimer les bonnes valeurs de paramètres. La distance relative entre l'estimation de τ et sa vraie valeur est faible pour quasiment toute la gamme de détectabilité du changement (Fig 2.4b). Les valeurs de κ estimées avec $\partial a \partial i$ sont très proches des vraies valeurs pour les cas de la croissance, mais beaucoup plus dispersées dans les cas de déclin (Fig 2.4d).



(a) τ Stairway Plot 2 **(b)** τ $\partial a \partial i$

(c) κ Stairway Plot 2 **(d)** κ $\partial a \partial i$

Figure 2.4: Distance relative entre le paramètre estimé et l'observé (le simulé) (en bleu des estimations très proches, en jaune des très éloignées) pour l'estimation de τ (a et b) et κ (c et d), en fonction du couple τ (abscisse) et κ (ordonné) utilisé pour la simulation et du logiciel utilisé : Stairway Plot 2 (a et c) et $\partial a \partial i$ (b et d).

2.1.4 Conclusion

Nous avons étudié l'impact de la démographie sur les SFS. L'utilisation d'un simulateur tel que msprime a permis de générer des scénarios variés de changements de taille de population à un temps τ dans le passé et d'intensité κ. Cela a permit

d'apprécier la gamme de paramètres τ et κ pour laquelle les SFS permettent de repérer une variation de la taille de la population, ainsi que leurs limitations. En effet, les SFS ne semblent pas offrir la possibilité de détecter un changement s'il est (i) trop récent : croissance de faible intensité ou décroissance, ou (ii) trop ancien, qu'il s'agisse de croissance ou de décroissance. L'utilisation des SFS reste néanmoins possible afin de déterminer l'histoire démographique d'une population donnée, ces derniers constituant de bons marqueurs pour l'inférence. Les SFS semblent cependant mieux adaptés pour inférer une croissance qu'un déclin, notamment dans le cas de variations récentes dans la taille des populations étudiées.

A partir des SFS, deux logiciels d'inférence de population ont été utilisés : Stairway Plot et $\partial a \partial i$. Le premier n'utilise que l'information des SFS pour réaliser son inférence. Le deuxième, au contraire, repose sur une méthode paramétrique qui nécessite la connaissance d'un modèle simplifié de population pour réaliser l'inférence. Bien que la méthodologie utilisée soit différente, ces deux logiciels ont permis de détecter un changement de taille de population lorsque les SFS simulés diffèrent du SFS théorique constant. Quant à l'évaluation des paramètres τ et κ, c'est le logiciel $\partial a \partial i$ qui semble plus précis. En effet, dans le cas de Stairway Plot, on observe une disparité relativement importante entre les paramètres estimés et observés. $\partial a \partial i$ produit quant à lui des résultats encourageants, notamment pour (i) estimer τ dans le cas d'une décroissance à un temps moyen ou d'une croissance, et (ii) estimer κ, uniquement en cas de croissance. Cependant, l'utilisation de ce logiciel requiert la mise en place d'un modèle et donc des connaissances préalables sur la population étudiée, connaissances qui ne sont pas toujours disponibles. A cet égard, Stairway Plot ne nécessite pas de telles connaissances : même s'il surestime souvent le nombre de dimensions rendant le scénario plus complexe que la réalité, il permet de bien détecter un changement de taille de population. Stairway Plot a tout de même des hypothèses fortes puisqu'il considère que l'histoire démographique de la population peut se résumer en une succession de morceaux de taille constante.

Maintenant que les SFS et les logiciels $\partial a \partial i$ et Stairway Plot ont été éprouvés dans le cas de scénarios simples à deux paramètres (τ et κ), il est possible de mettre en place des scénarios plus complexes, pouvant même allier démographie et structure. Cependant, Stairway Plot n'est pas adapté à des scénarios de structure, contrairement à $\partial a \partial i$ et ne pourra donc pas être utilisé dans ce cas.

En outre, il n'a été présenté ici que des méthodes d'inférence qui s'appuient sur les SFS. L'évaluation de méthodes reposant sur d'autres statistiques (comme le LD), ou d'autres bases d'information, comme les évènements de recombinaison, avec notamment les segments IBD permettraient sans doute, d'avoir une autre gamme de paramètres pour lesquels un changement de taille de population est détectable.

2.2 Testing for population decline using maximal linkage disequilibrium blocks

Ce chapitre a fait l'objet d'une publication du même nom publiée dans *Theoretical Population Biology* en 2020. L'article et les analyses complémentaires sont rédigés en anglais.

2.2.1 Résumé de l'article

2.2.1.1 Motivation

Seulement 6% des espèces décrites ont un statut de conservation. Les méthodes utilisées pour attribuer les statuts de conservation sont souvent basées sur un suivi du nombre d'individus au cours du temps, ou un suivi de la taille de l'aire géographique de la population. Ces méthodes sont difficiles à mettre en place pour l'ensemble des espèces. Il existe des méthodes d'inférence démographique s'appuyant sur les modèles de génétiques de populations. Cependant, comme nous l'avons montré avec le SFS dans le chapitre précédent, ces méthodes ne sont pas forcément utilisables pour la détection de déclin récent. Nous avons donc décidé de mettre en place une nouvelle méthode d'inférence démographique, ne s'appuyant pas sur l'information des évènements de mutation (comme c'est le cas du SFS), mais sur l'information contenue par les évènements de recombinaison et en considérant tout particulièrement le cas d'un déclin récent.

2.2.1.2 Principaux résultats

Nous avons étudiés les blocs non recombinés au sein d'un aligement multiple ou blocs *Maximal Recombination Free* (MRF). Ces segments d'alignement multiple sont hérités d'un même ancêtre commun par tous les individus échantillonnés. Ils sont plus grands dans des petites populations que dans des larges populations. Nous avons utilisés la distribution de ces blocs normalisés par leur longueur moyenne pour créer un test détectant des déclins récents de population.

Cependant, les blocs MRF sont difficiles à détecter, nous avons donc considéré les blocs *Maximal Linkage Disequilibrium* (MLD), découpés grâce au test des quatre gamètes (Hudson and Kaplan 1985) qui représentent un ensemble de nucléotides adjacents, compatibles avec une seule topologie d'arbre. Utilisant ces blocs, qui se comportent de la même manière que les blocs MRF, nous avons créé un nouveau test du déclin récent de population à l'aide de la distribution des longueurs de blocs normalisés par la longueur moyenne. Il possède un pouvoir de détection de 50% pour des populations ayant perdu la moitié de leur taille de population N il y a $0.05N$ générations.

2.2.1.3 Conclusion

Nous avons pu démontrer que l'utilisation de l'information des évènements de recombinaison au sein d'une population permet de détecter des déclins récents avec une puissance de détection importante et une gamme de date de déclins allant jusqu'à 3 ordres de grandeur.

Il est cependant important de noter que le test mis en place est plus précisément un test d'écart à un scénario de population constante non structurée. Il n'est pas possible de différencier un scénario de déclin d'un scénario de bottleneck ou de migration entre différentes populations.

La normalisation par la longueur moyenne utilisée dans ce test nécessite un bon échantillonage de la distribution de longueurs de blocs. Ce n'est malheureusement pas le cas avec les données actuellement disponibles. Il est donc nécessaire de prendre en compte cette contrainte afin de proposer une alternative à la normalisation.

2.2.2 Article

L'article est présenté dans les pages suivantes, il est suivi d'analyses complémentaires réalisées pendant l'étude mais non incluses dans la publication.

Theoretical Population Biology 134 (2020) 171–181

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

Testing for population decline using maximal linkage disequilibrium blocks

Elise Kerdoncuff [a,b,*], Amaury Lambert [b,c], Guillaume Achaz [a,b]

[a] Atelier de Bioinformatique, UMR 7205 ISYEB, Sorbonne Université, CNRS, EPHE, Muséum National d'Histoire Naturelle, Paris, France
[b] SMILE (Stochastic Models for the Inference of Life Evolution), UMR 7241 CIRB, Collège de France, CNRS, INSERM, PSL Research University, Paris, France
[c] Laboratoire de Probabilités, Statistique et Modélisation (LPSM), UMR 8001, CNRS, Sorbonne Université, Paris, France

ARTICLE INFO

Article history:
Received 9 July 2019
Available online 9 April 2020

Keywords:
Coalescent theory
Recombination
Demography
Conservation biology

ABSTRACT

Only 6% of known species have a conservation status. Methods that assess conservation statuses are often based on individual counts and are thus too laborious to be generalized to all species. Population genomics methods that infer past variations in population size are easy to use but limited to the relatively distant past. Here we propose a population genomics approach that tests for recent population decline and may be used to assess species conservation statuses. More specifically, we study Maximal Recombination Free (MRF) blocks, that are segments of a sequence alignment inherited from a common ancestor without recombination. MRF blocks are relatively longer in small than in large populations. We use the distribution of MRF block lengths rescaled by their mean to test for recent population decline. However, because MRF blocks are difficult to detect, we also consider Maximal Linkage Disequilibrium (MLD) blocks, which are runs of single nucleotide polymorphisms compatible with a single tree. We develop a new method capable of inferring a very recent decline (e.g. with a detection power of 50% for populations whose size was halved to N, 0.05 $\times N$ generations ago) from rescaled MLD block lengths. Our framework could serve as a basis for quantitative tools to assess conservation status in a wide range of species.

1. Introduction

The severe and rapid changes imposed by human activities upon living organisms are suspected to be a major factor leading to short-term mass extinctions (Barnosky et al., 2011). The most comprehensive list of endangered species is the Red List of the International Union for Conservation of Nature (IUCN) (Rodrigues et al., 2006). Criteria used in the list to assess the species conservation status are based on geographical range, population trends, threats to habitat and ecology. Despite being very robust and reliable, these criteria are hard to establish for many species. To quantify the ongoing crisis for a wider range of organisms, there is a crucial need to develop quantitative measures of extinction risk to efficiently monitor species in real time and at a global scale. Previous attempts were developed to estimate quantitatively extinction rates, including two of the present authors, based on occurrence data (Régnier et al., 2015; Ceballos et al., 2017; Sánchez-Bayo and Wyckhuys, 2019) or genetic data (from museum specimen (Díez-del Molino et al., 2018; van der Valk et al., 2019)). The genetic methods measure the genetic diversity at different time to estimate the population size at these times and conclude on a general trend. The limitation of these methods is the difficulty to obtain time series data.

A handful of genomes sampled in a population at a single time point can help infer the past demography of this population (Gutenkunst et al., 2009; Li and Durbin, 2011; Excoffier et al., 2013; Harris and Nielsen, 2013; Sheehan et al., 2013; Schiffels and Durbin, 2014; Lapierre et al., 2017; Ringbauer et al., 2017; Terhorst et al., 2017; Beichman et al., 2018). In standard population genetic inferences, the periods when variations of population size can be estimated are of the order of N_e generations back in time. N_e denotes the so-called effective population size (Wright, 1931). Recent methods such as MSMC (Schiffels and Durbin, 2014) can provide inferences on more recent past but hardly scale up to large data sets of complete genomes because of their computational load. With the development of next generation sequencing, complete genomes from multiple individuals of the same species are now released routinely (Gibbs et al., 2015; Alonso-Blanco et al., 2016). Actual methods cannot be applied to test for recent decline of populations, the models and methods we present in this manuscript specifically target very recent past

* Corresponding author at: SMILE (Stochastic Models for the Inference of Life Evolution), UMR 7241 CIRB, Collège de France, CNRS, INSERM, PSL Research University, Paris, France.
E-mail address: elise.kerdoncuff@college-de-france.fr (E. Kerdoncuff).

https://doi.org/10.1016/j.tpb.2020.03.004
0040-5809/

when considering small populations and are meant to be applied to data sets of arbitrary size.

Methods using whole genome sequences to infer demography use different measures of genomic polymorphism. One of these measures is the so-called Site Frequency Spectrum, or SFS (Fu, 1995). The SFS, that is the genome wide distribution of the frequencies of polymorphic alleles in a sample of the population, is strongly distorted by the demographic history of the species (Adams and Hudson, 2004; Marth et al., 2004). SFS-based methods (e.g. Gutenkunst et al., 2009) can handle arbitrarily large numbers of loci and genomes but disregard correlations between sites caused by genetic linkage. Using genetic linkage information may help overcoming the SFS-based methods limitations (e.g. difficulty to discriminate between different scenarios Lapierre et al. (2017) and to infer recent demography).

Recombination is the process by which two DNA sequences are intermixed to create a new sequence that combines segments of different ancestries. When two homologous regions of the genome are inherited from the same ancestor without having undergone recombination, they are said IBD: *Identical By Descent*. The probability distribution and the length of IBD regions passed through generations have been studied (Stam, 1980; Chapman and Thompson, 2003; Stefanov, 2000).

Recombination patterns are also characterized by Linkage Disequilibrium (LD). LD arises when individuals of a finite population share chunks of DNA inherited from a common ancestor (IBD blocks). Specifically, two variants located at two distinct sites are in linkage disequilibrium (LD) when their joint frequency differs from what is expected under independence. More specifically, LD is defined as the covariance $f_{A_1B_1} - f_{A_1}f_{B_1}$, where f_{A_1} is the frequency of allele 1 at locus A (Lewontin and ichi Kojima, 1960). When $f_{A_1} \times f_{B_1} = f_{A_1B_1}$, the two variants are said in complete linkage equilibrium. On average, LD decreases exponentially with genetic distance due to recombination. The pattern of LD is distorted by demography (Hill and Robertson, 1968) and thus can be used to infer the past demography of a population (Hollenbeck et al., 2016; Patin et al., 2014).

Importantly, despite the fact that breakpoints between IBD blocks are usually not observable when comparing two homologous regions, "long enough" IBD blocks can be retrieved by applying one of several recent methods to a pair of sequences (Purcell et al., 2007; Gusev et al., 2009; Browning and Browning, 2010). These methods are based on detecting long identical shared segments (Gusev et al., 2009) or shared regions that harbor multiple rare variants (Purcell et al., 2007; Browning and Browning, 2010). If two individuals share the same rare variant, they may also share the surrounding chromosomal region, particularly because rarer variants are more likely to be relatively recent. Most methods take sequencing errors into account, allowing IBD blocks to not be totally identical. The accuracy of IBD block detection depends on the algorithm used (Browning and Browning, 2013).

Some demographic inference methods are based on the distribution of lengths of inferred pairwise IBD blocks in a population. Palamara et al. (2012) have calculated the distribution of expected lengths of pairwise IBD blocks for a given parameterized demographic model. Browning and Browning (2015) have calculated the expected time to the most recent common ancestor (TMRCA) of an IBD block as a function of its length. Then they use the empirical density of IBD block lengths to estimate the distribution of TMRCA and thus the variations of effective population size through time.

Other methods use the length of identical shared segments of chromosome within a diploid individual (Hayes et al., 2003). Two identical shared segments may be inherited from a common ancestor without recombination event (and then be IBD) or may not be IBD as there are invisible recombination events that may have occurred within it. The probability that the two haplotypes of an individual share identical alleles for a given number of adjacent positions can be predicted (Hayes et al., 2003; MacLeod et al., 2009). Tools have been developed to apply these methods to infer demographic inference from genomic data (MacLeod et al., 2013; Harris and Nielsen, 2013).

Yet another approach to infer demographic history from IBD blocks is to reconstruct the genome-wide distribution of the TMRCA between two haploid genomes. In PSMC, Li and Durbin (2011) devised a Hidden Markov Model that infers the TMRCA from the positions of heterozygous sites along a pair of sequences and then estimate a step-wise demographic pattern. MSMC, the extension of PSMC (Schiffels and Durbin, 2014), analyzes the heterozygosity pattern from multiple individuals and uses first coalescence events between any two haploid genomes of the sample. These methods are computationally intensive (as of today, MSMC cannot infer the demographic history of more than 8 individuals) and pool the diversity on windows of 100 bp, that are assumed to form a single locus with two states, heterozygous or homozygous.

Importantly, the previous methods infer stepwise changes of the "effective population size" ($N_e(t)$) that are estimated from the density of coalescence events. This motivated Mazet et al. (2015, 2016), Chikhi et al. (2018), Rodríguez et al. (2018) to propose to replace $N_e(t)$ by the more explicit *Inverse Instantaneous Coalescence Rate*. IICR only matches the instantaneous population size when the population is panmictic. It is nonetheless always possible to find a population model with constant size but spatial structure that corresponds to any IICR of a size-changing population for the TMRCA of 2 sequences (Chikhi et al., 2018). For larger samples, the joint distribution of coalescence events $[T_2, T_3, \ldots]$ can be used, in theory, to disentangle structure from demography (Grusea et al., 2019).

Existing methods for demographic inference using recombination information often use the whole genome of few individuals (less than 10) or use a smaller part of the genome. These methods only consider the joint history of two individuals (e.g. the pairwise IBD length distribution or the time of the first coalescence event between any two haploid genomes) whose algorithmic complexity increases drastically with the number of individuals (e.g. detection of pairwise IBD blocks is quadratic) and generates a computational load limiting in most cases the application of the methods to a larger number of individuals. On the other hand, with few individuals, demographic inferences are unable to detect recent changes of population size.

Following the idea of Tiret and Hospital (2017), we decided to study the IBD concept extended to a multilocus segment and a larger number of individuals ($n > 2$). Some studies have been conducted on the amount of genetic material shared IBD with $n > 2$, considering closely related individuals (Donnelly, 1983; Ball and Stefanov, 2005). We extend the concept at a population level while relaxing the need for *identical* sequence (without mutation), which is why we decided to define a new term. We call 'MRF blocks' homologous segments that are entirely inherited from the same ancestor without recombination; these segments may or may not harbor different alleles, because of mutations. An MRF block is a segment of an alignment of haploid genomes that share the same coalescent tree. A recombination event along the sequence cuts the genome alignment into two MRF blocks, one on each side of the recombination point. By definition there is no recombination within MRF blocks so that all variants located within an MRF block are necessarily in complete linkage disequilibrium. The reciprocal is not true, as variants in complete LD do not necessarily belong to the same MRF block. MRF blocks carry the information of any recombination event that happened among the sampled individuals. As for IBD blocks, MRF blocks are usually not observable.

Outline. We have developed a new test to detect very recent population declines of endangered species. We first consider the full length distribution of MRF blocks in a sample of haploid genomes ($n \geq 2$). Second, as MRF blocks are not directly observable from sequence alignments, we devised a simple and efficient algorithm to chop an alignment of $n \geq 4$ haploid genomes in Maximal Linkage Disequilibrium (MLD) blocks, that are segments whose variants are in complete LD. From the length distributions of MRF blocks or MLD blocks, we devised a summary statistic to test whether a population has been declining in the very recent past. Our method is not limited by the number of genomes in the sample.

2. Model and methods

In the absence of recombination, ancestral relationships between genomes can be represented in the form of a genealogical tree. Individual haploid genomes at present time are the leaves of the tree, the MRCA of these individuals is the root. The fusion of two lineages into one (a common ancestor) is named coalescence event (Kingman, 1982), hence the name of "coalescent tree". The sum of all branch lengths that separates two genomes up to their common ancestor is the time of divergence between them, usually expressed in generations. In the Wright–Fisher model with constant population size N, branch lengths measured in number of generations scale like N. In particular, if we define T as the total length of the coalescent tree, the expectation of T is proportional to N. Large populations generate coalescent trees with deep nodes, whereas small populations have shallow coalescent trees.

In the presence of recombination, two loci of an alignment have the same coalescent tree only if no recombination event happened since their MRCA. We name MRF block, a maximal interval along the alignment of sites sharing the same coalescent tree. MRF blocks are consequently separated by recombination points, corresponding to recombination events. It is standard to assume that conditional on the total length T of the coalescent tree of a site, the length L of its MRF block is exponentially distributed with rate ρT, where ρ is the recombination rate (expressed in an arbitrary unit proportional to Morgan). Then for a fixed ρ, T and L are negatively correlated: recombination is more likely to occur in deep trees, which thus are carried by shorter blocks. As mentioned above, as T is proportional to N, MRF blocks are also shorter in larger populations. More accurately, because the law of T/N does not depend on N, neither does the law of NL (for $n = 2$ it alludes to results of Carmi et al., 2014). In other words, if population 1 has size N_1 and population 2 has size N_2, the distribution of MRF block lengths in population 2 can be deduced from that in population 1 by a scaling factor N_1/N_2, both populations having identical demography otherwise. For example if $N_2 = 2N_1$, the MRF blocks in population 2 are twice smaller than those of population 1.

Note that for a given N the lengths $(L_1, L_2, \ldots)$ of successive adjacent blocks have the same distribution, but they are not independent, because the coalescent trees of adjacent MRF blocks are not. The dependencies between these trees are encoded in the so-called Ancestral Recombination Graph (ARG) (Griffiths and Marjoram, 1997). Because these dependencies have a complex structure (Wiuf and Hein, 1999), a popular way of approximating them is the Sequentially Markovian Coalescent (SMC) (McVean and Cardin, 2005; Marjoram and Wall, 2006). This approximation neglects coalescences between lineages with no overlapping ancestral material and assumes Markovian dependencies of coalescent trees along the sequence: the genealogy of an MRF block only depends on the genealogy of the adjacent ones.

Although genealogies of different MRF blocks are not independent, they are asymptotically independent as the distance between them increases.

Throughout this article, we use msprime to generate MRF blocks directly from the ARG (Kelleher et al., 2016) but very similar results were obtained with a local SMC implementation. We assume constant recombination and mutation rate along the genome. We simulated the alignment of $n = 10$ haploid genomes at present time.

Demographic scenario. We consider a single change of population size (Fig. 1(a)). Here N_t represents the population size at time t, $t = 0$ is the present time and positive values represent the past. We denote by κ the ratio of the two sizes: $\kappa = N_\infty/N_0$, and by τ the time at which the population size changes in coalescent units of N_0 generations. If $\kappa = 1$, $N_\infty = N_0$: there is no change. If $\kappa = 10$, $N_\infty = 10N_0$: the population size has been divided by 10, τN_0 generations in the past.

3. MRF blocks

3.1. Distribution of block lengths

Impact of population decline on tree length. For declining populations ($\kappa > 1$), the coalescent trees have two distinct time scales: a first one for the shallow part of the tree ($t < \tau$), expressed in N_0 generations, and a second one for the deep part of the tree ($t > \tau$) that is expressed in κN_0 generations. When the declining population tree is compared to a standard coalescent tree (constant population size), it has shorter external branches if the reference time scale is expressed in κN_0 generations *or* longer internal ones if the reference time scale is expressed in N_0 generations. When it is compared to a reference tree with population size chosen so as to have the same T_{MRCA}, its external branches are too short *and* its internal branches are too long. Similarly, the distribution of the total length T of the tree is overdispersed when compared to the length of the standard coalescent tree with the same mean.

Impact of population decline on lengths of MRF blocks. For a declining population, the distribution of the length L of MRF blocks will depend not only on ρ and N_0 but also on κ and τ. As the tree relative branch lengths are distorted and the distribution of T is overdispersed, so is the distribution of L. In a declining population, the distribution of L can be seen as a mixture of the two distributions that correspond to the two population sizes, N_0 and κN_0. The strength of the decline (κ) tunes the difference between the distributions; the date of decline (τ) tunes in what proportion the two distributions are mixed. When $\tau \to 0$ (practically, $\tau < 10^{-4}$ times N_0 generations for a sample size $n \in [10, 100]$), the distribution of L is indistinguishable from that of block lengths in a population with constant size equal to κN_0. At the opposite, for $\tau \to \infty$ (practically, $\tau > 10$ times N_0 generations for a sample size $n \in [10, 100]$), the distribution of L is indistinguishable from that of block lengths in a population with constant size equal to N_0. As a result for $\tau \in [10^{-4}, 10]$, the distribution of L has an excess of MRF blocks smaller than the N_0 reference and an excess of MRF blocks longer than the N_∞ reference (Fig. 1(b)). The small blocks correspond to the trees whose total length T is mostly driven by the distant N_∞ time scale and the long ones to the trees whose total length T is mostly driven by the recent N_0 time scale.

As mentioned in the previous section, in a population with constant size N, the distribution of L, briefly denoted L_N, scales like $1/N$, in the sense that the distribution of $\tilde{L} := NL_N$ does not depend on N. In particular, the distribution of $L' := L/E[L]$ does not depend on N in a population with constant size and follows the law of $\tilde{L}/E[\tilde{L}]$. However, the distribution of L' is distorted when

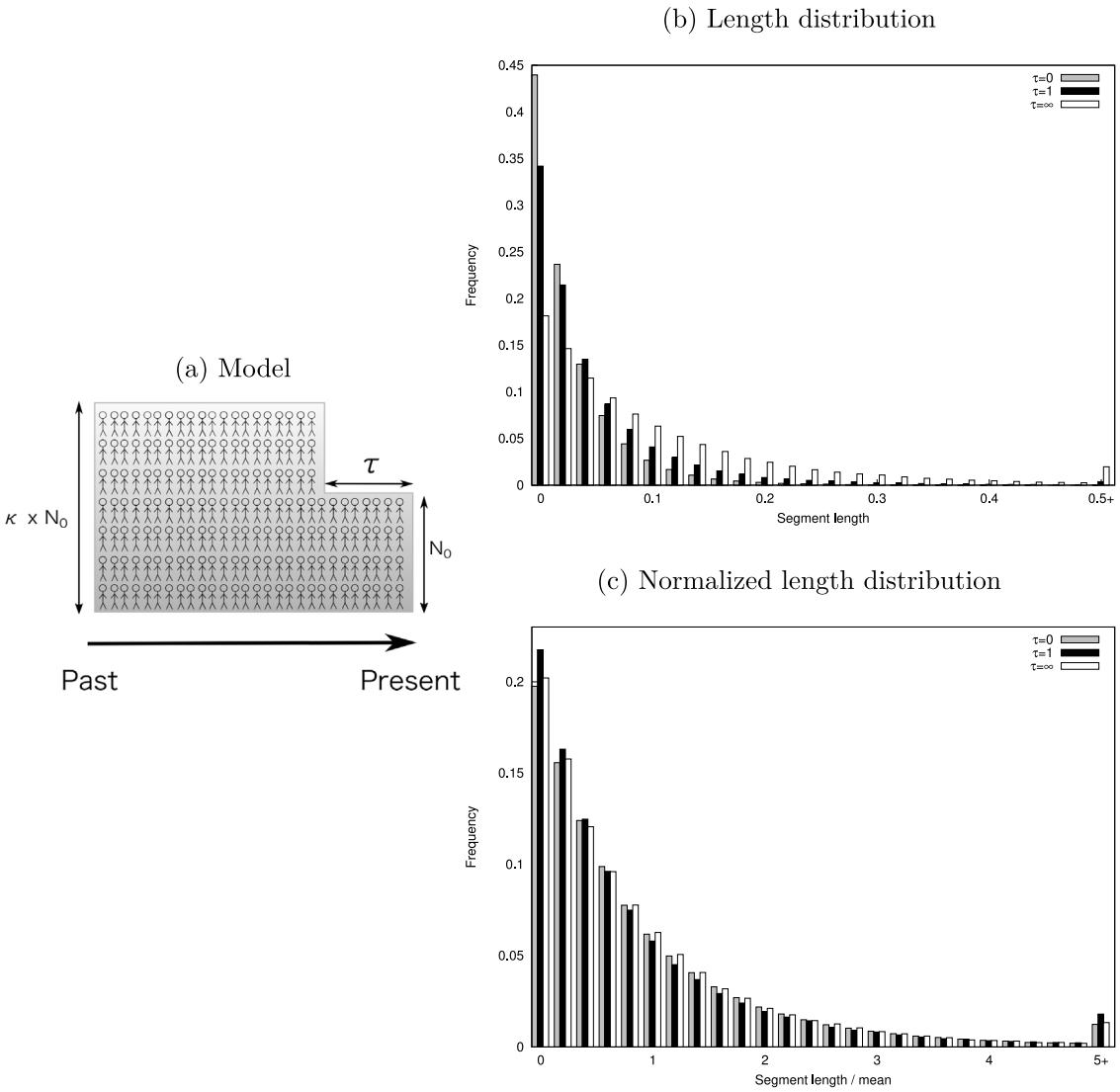


Fig. 1. Impact of the demography on the distribution of MRF block lengths. (a) The demography considered here is a sudden size change tuned by 3 parameters : N_0, the actual population size, τ the date of decline (in backward time) and κ the strength of decline. Time is expressed in N_0 generations. (b) Distribution of L for $\rho = 1$, $\kappa = 3$ with $\tau = \{0, 1, \infty\}$. When $\tau = 0$ (gray) or $\tau = \infty$ (white), population size is constant. (c) Distribution of $L' = L/E[L]$ under the same values of $\rho = 1$, κ and τ. In case of a decline, the distribution is overdispersed, with an excess of both short and long normalized MRF blocks.

there is a size change. For a declining population, the distribution of L' is overdispersed, it has an excess of small blocks (*i.e.* less than 0.2) and an excess of long blocks (*i.e.* more than 5), as can be seen in Fig. 1(c).

Note that we always have $E[L'] = 1$, but here $E[L]$ has

$$\frac{1}{N_\infty}E[\tilde{L}] = E[L_{N_\infty}] < E[L] < E[L_{N_0}] = \frac{1}{N_0}E[\tilde{L}].$$

As the block distribution of L_N is a mixture of the one of L_{N_0} and L_{N_∞}, $E[L]$ is bounded by $E[L_{N_\infty}]$ and $E[L_{N_0}]$ that depend on the population size.

4. MLD blocks

4.1. Definition

All recombination events are not directly visible in a genome alignment. First, adjacent MRF blocks may have coalescent trees sharing the same topology and the same branch lengths, so that mutations occurring on either tree show exactly the same pattern on either block. Second, adjacent MRF blocks may have coalescent trees sharing the same topology but not the same branch lengths, so that mutations occurring on either tree display the same bipartitions (compare the second and third tree in Fig. 2). Third, even if two adjacent MRF blocks have trees with different topologies, it is possible that branches distinguishing these topologies do not carry mutations (see the second block in Fig. 2).

Importantly, recombination events that happen between the two oldest lineages do not impact the topology of the tree, so are never detectable because they do not impact the possible bipartitions.

A possibility used in the literature to detect breakpoints between MRF blocks is to detect the changes in the density of polymorphic sites along the sequence due to the change of coalescent tree (like in PSMC, Li and Durbin, 2011).

Here we used instead the incompatibilities between bipartitions displayed by polymorphic sites to place the minimal number of recombination events on the alignment. Two bipartitions are said incompatible when they are not compatible with a common tree.

In what follows, we will assume that the mutation rate μ is constant through time and along the genome.

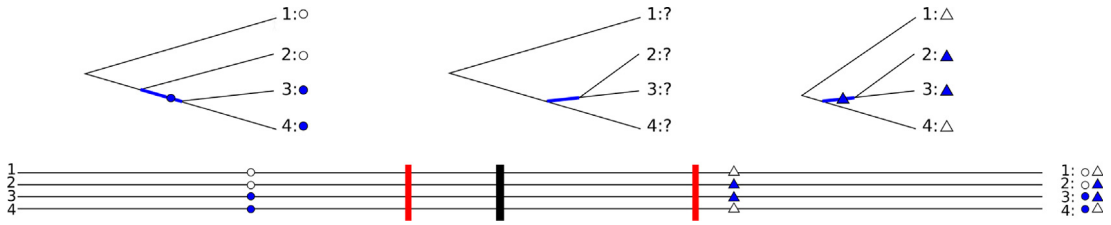


Fig. 2. Detection of recombination in three MRF blocks. The four lines represent four haploid genomes, circles and triangles are mutation events, red lines are the true recombination events delimiting MRF blocks and above each MRF block is represented its true tree. Mutation events occur on certain lineages as represented on the trees. The first recombination event generates an incompatibility between the blue branches of the first two MRF blocks, but as no mutation occurs on the second MRF block, this recombination event cannot be detected. The second recombination event does not change the topology of the tree and thus this second event cannot be detected either. However, the first and the third MRF blocks carry mutations that are not compatible; thus a minimum of one recombination event can be inferred between the two mutations, as indicated by a vertical thick black line arbitrarily placed in the middle.

4.1.1. The four-gamete test

From now on, we assume that each site can be hit at most once by a mutation, so that a polymorphic site is always bi-allelic, an assumption known as the "infinitely-many sites model". The four-gamete test (Hudson and Kaplan, 1985) serves to detect incompatibilities between bipartitions displayed by two polymorphic sites. For any two biallelic sites (A/a and B/b) there are at most four gamete haplotypes in the population (A-B A-b a-B and a-b). Under the infinitely-many sites model, the four possible haplotypes cannot be observed in a sample if the two sites share the same genealogy. Then if the four possible haplotypes are observed in the sample, a recombination event must have occurred between them — but not necessarily the other way round. This property can be used to compute a lower bound for the number of recombination events in a genome alignment (Hudson and Kaplan, 1985) or even to estimate the recombination rate (Hey and Wakeley, 1997). We used it to compute and place the minimal number of breakpoints in a genome alignment.

Two polymorphic sites are said *incompatible* if the four possible haplotypes are present in the sample. When a sequence of adjacent polymorphic sites contains no pairwise incompatibility, we speak of a sequence of compatible sites. Note that a sequence of compatible sites is in complete linkage disequilibrium. We thus define an MLD block, for *Maximal Linkage Disequilibrium* block, as any maximal sequence of compatible sites.

We now explain how to extend this notion originally designed for haploid genomes (or phased diploid genomes) to an unphased diploid genome, that is, a diploid sequence lacking the linkage information. For an unphased diploid genome, the two original haplotypes can be determined if the diploid genome is homozygous at least one the two sites:

- When the genome is homozygous at both loci (A/A-B/B), both haplotypes must be A-B.
- When the genome is homozygous at one locus and heterozygous at the second one (A/A-B/b), the haplotypes must be A-B and A-b.

The four-gamete test can then be extended to a sample of unphased diploid genomes by saying that two sites are incompatible in this sample if they are incompatible in the subsample of haplotypes that have been inferred thanks to the previous remark. When the haplotype is ambiguous, the sites are considered compatible and do not bring more information about a recombination event.

4.1.2. The chopping algorithm

We used the four-gamete test to detect incompatibilities in the genome alignment and to chop it into MLD blocks (Fig. 3). To avoid computing the full matrix of pairwise incompatibilities between all polymorphic sites of the genome, we only compute the incompatibilities for sequences of P adjacent polymorphic sites (by default $P = 150$). Each pair of incompatible sites (i, j) defines an interval that contains at least one MLD breakpoint. To place the MLD breakpoint, we seek the shortest interval that is sufficient to explain the incompatibilities.

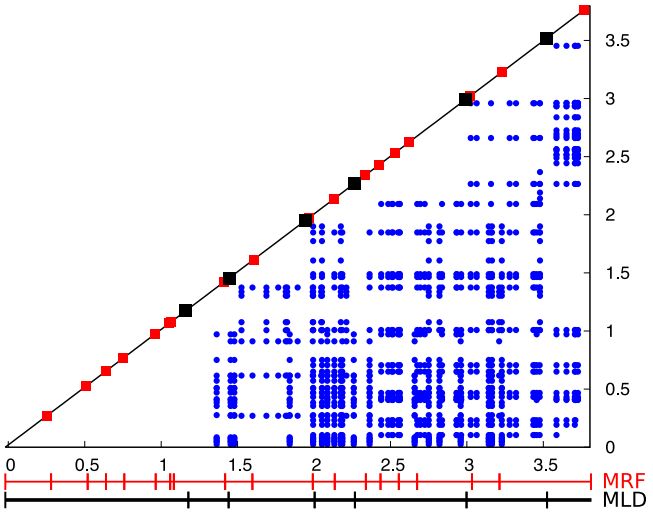


Fig. 3. The incompatibility matrix and the chopping algorithm. X and Y-axis are positions on the genome alignment. Blue dots represent a pair (x, y) of incompatible sites. The red squares are the true positions of recombination events (MRF breakpoints) and the black squares are MLD breakpoints inferred by the chopping algorithm. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Algorithm. We retrieve all intervals and sort them in increasing order of site positions along the genome (first by i the first site position and when equal, by j the second site position). As we scan two times the list of intervals, the algorithm complexity is linear with the number of polymorphic sites:

1. **Discarding and shortening.** For this step, we scan the list in reverse order, from the last (N) to the first interval. (The algorithm can be done in the forward order, the distribution will be slightly different but it will not affect the study.) Each interval containing another entire interval is discarded: for two intervals (i_N, j_N) and (i_{N-1}, j_{N-1}), if $i_N \leq i_{N-1} \leq j_{N-1} \leq j_N$, then (i_N, j_N) is discarded. When two intervals overlap, they are replaced by their intersection (the two original ones are discarded): for the two intervals (i_N, j_N) and (i_{N-1}, j_{N-1}), if $i_{N-1} \leq i_N \leq j_{N-1} \leq j_N$, both are replaced by a new interval (i_N, j_{N-1}), that is then compared to (i_{N-2}, j_{N-2})...
2. **Positioning.** From the final list of disjoint intervals, we place an MLD breakpoint at the middle of each interval.

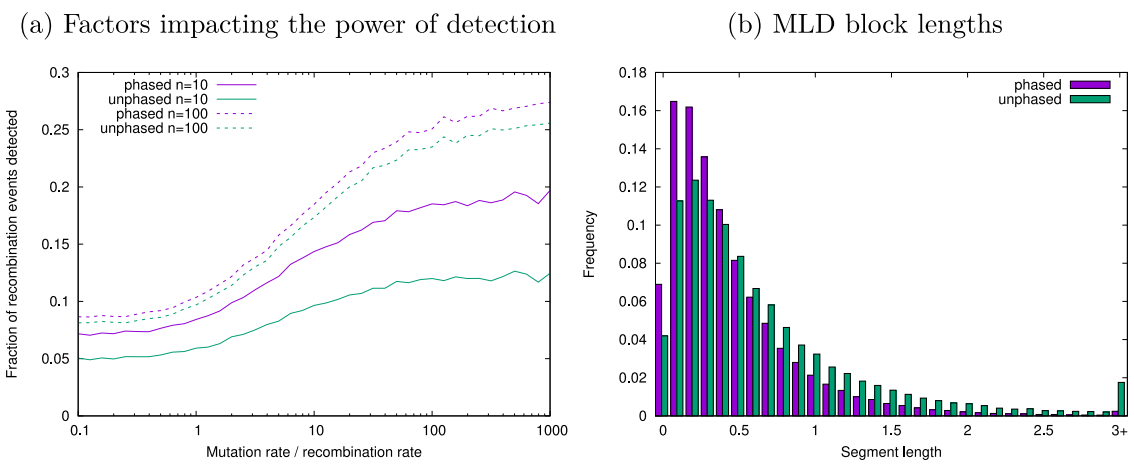


Fig. 4. Detection of recombination events and its impact on MLD block length (L_c) distribution under a constant population. (a) Fraction of recombination events that are detected as a function of μ/ρ for different sample sizes ($n = 100$, dashed lines and $n = 10$ plain lines) and for phased (purple) or unphased (green) diploid genomes. (b) Distribution of MLD block lengths for phased (purple) and unphased (green) diploid genomes in a population of constant size ($\mu = 10$, $\rho = 1$, $n = 10$). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

MLD breakpoints partition the genome alignment into MLD blocks.

4.2. Length distribution

The distribution of the length L_c of a typical MLD block does not only depend on the distribution of L (MRF block length) but also on the fraction p of recombination events that are detected. This fraction increases with the ratio μ/ρ, as illustrated in Fig. 4(a). When many mutations occur in two different MRF blocks ($\mu \gg \rho$), the probability that they occur on incompatible branches of their respective coalescent trees increases and so does the detection efficiency, up to a point of saturation due to cases when these MRF blocks share the same tree topology. The number of sampled individuals also impacts the efficiency of detection (Fig. 4(a)): the larger the sample size, the higher the probability to observe incompatible mutations. The four-gamete test for unphased diploid genomes has obviously less power to detect recombination than for phased genomes (Fig. 4(a)).

The lower the power to detect recombination, the longer the MLD blocks. In particular, phased genomes have smaller MLD blocks than unphased ones (Fig. 4(b)). Furthermore increasing the sample size results in more detectable recombination points and thus smaller MLD blocks. In Fig. 4(b), the average block length, in our arbitrary unit for $n = 10$ phased haploid genomes is $\bar{L_c} = 0.497$ ($\mu = 10, \rho = 1$). Considering smaller sample size will result in larger MLD blocks (e.g. $\bar{L_c} = 1.32$ for $n = 5$). This implies that the total number of blocks can be limiting for small sample size, and that these long blocks will be harder to detect in scaffolds of partial genomes. In (very) large samples, MLD blocks are shorter: $\bar{L_c} = 0.132$ for $n = 600$ (ten times smaller than for $n = 5$) and $\bar{L_c} = 0.103$ for $n = 6000$. On a side note, the theoretical pitfall of having too small "undetectable" blocks can always be overcome by subsampling.

Here, we consider the block lengths normalized by the average length $L'_c = L_c/\bar{L_c}$. Similarly to the MRF blocks, the distribution of L'_c does not depend on the value of N but does depend on the demographic scenario (Fig. 5). However, it still depends on our ability to detect recombination and so on the ratio μ/ρ and n the number of sampled individuals. To compare distributions, it is then important that they have the same ratio μ/ρ and the same n.

Similar to what we have observed for MRF blocks, a declining population exhibits both an excess of small blocks ($L'_c < 0.2$) and large blocks ($L'_c > 5$) (Fig. 5). The shape of the distribution of

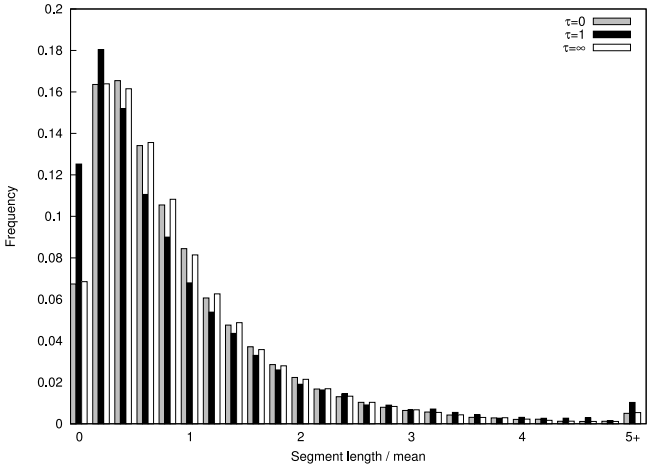


Fig. 5. Distribution of L'_c for a population of constant size (white, $N = N_0 = 1$ and gray, $N = \kappa N_0 = 3$) and for a declining population (black for $\tau = 1$) with $\rho = 1$, $\mu = 10$ and $n = 10$.

L'_c (Fig. 5) differs from the one for L' (Fig. 1(c)): MLD blocks are longer than MRF blocks. Indeed, they contain a variable number of MRF blocks and below a certain size, MRF blocks are not detectable as recombination points at the edges of an MRF block can be detected only when mutations have occurred inside the block. MLD blocks are always longer than MRF blocks.

5. Statistical tests for population decline

5.1. Test

To test for population decline, we use the excess of small and large blocks that we observe when comparing samples from a declining *vs* a constant population size. More specifically, we compute the fraction of blocks whose normalized length is either smaller than 0.2 or larger than 5, both in the case of MRF blocks ($f = f_{L'<0.2} + f_{L'>5}$) and of MLD blocks ($f_c = f_{L'_c<0.2} + f_{L'_c>5}$). To set an empirical threshold value under H_0, we simulate 10,000 genomes of 10^5 MRF blocks under a constant population size for 10 haploid genomes and compute both $f^{5\%}$ and $f_c^{5\%}$ as upper limits for one-tailed tests: $f^{5\%} = 0.214236$ and $f_c^{5\%} = 0.075824$. As the threshold is empirical, simulations need to be redone for a change

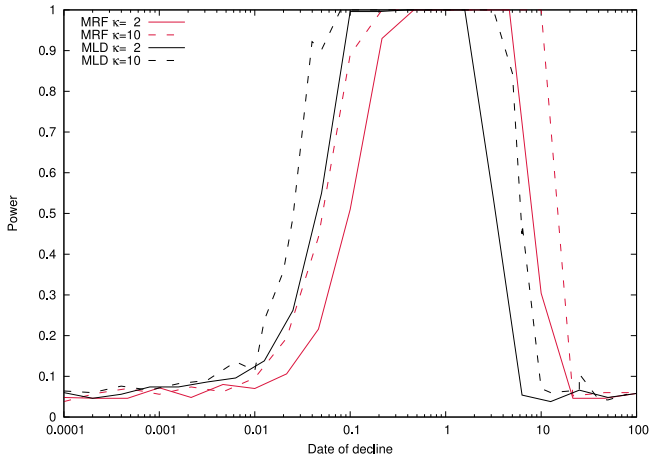


Fig. 6. Power to detect population decline. The test based on MRF blocks (f) is pictured in red, whereas the one based on MLD blocks (f_c) is represented in black. We assess the power of the two tests for $\kappa = 2$ (plain line) and $\kappa = 10$ (dashed line) with $\tau \in [0.0001, 100]$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

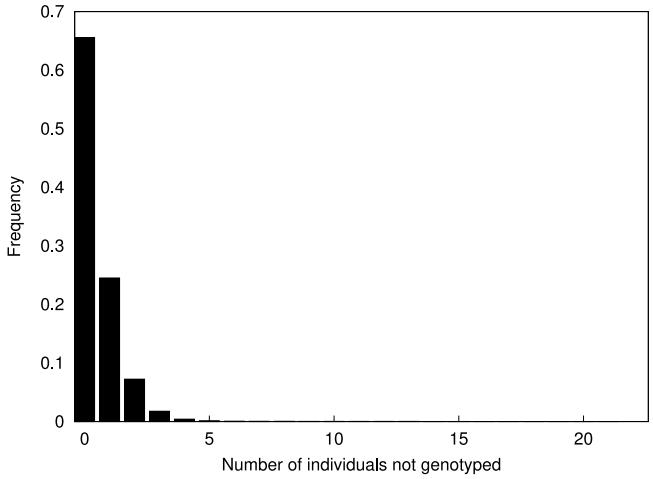


Fig. 7. Distribution of the number of individuals not genotyped per SNP on chromosome 1 of the Gorillas data set (Prado-Martinez et al., 2013).

in sampled size or in null model. Time needed for simulations depends on the algorithm/software used and the specific features of the model.

5.2. Power

To assess the power of this test, we simulated 1000 replicates under population decline (H_1 with various τ and κ) and report the fraction of runs where $f^{H_1} > f^{5\%}$ for MRF blocks or $f_c^{H_1} > f_c^{5\%}$ for MLD blocks. When the power is 1, the decline was significant in all runs. When the power is 5%, the decline is not detectable, the test cannot differentiate H_0 and H_1.

Without surprise, results show that the power of the test to detect population decline depends on both the decline strength (κ) and the date of decline (τ) (Fig. 6). For both tests based either on MRF or on MLD blocks, the power outreaches the 5% risk only for a range of τ. The type I error of the test is 5% as expected. For both tests, the range of detection is wider when the decline is stronger (compare dashed to solid lines in Fig. 6). The surprise is that the test based on MLD blocks (f_c) detects more recent declines than the test based on MRF blocks (f). Therefore, we recommend using the f_c test when searching for very recent decline even if MRF blocks are known (which is generally not the case).

6. Application to data: the case of the western lowland Gorillas

6.1. Handling the low quality of real genomes

Genomic data sets often include sequencing errors and regions that are not genotyped. Consequently, the f_c test cannot be run as is on these data sets. We present some modifications to our test to handle the poor quality of data. We show in this section that adjustments can be made to get the information from the L_c' distribution.

Simulations with lower quality

Difficulties in applying the f_c test to real data sets can stem from the low quality of DNA sequences. We replicated in the simulated genomes the two main issues, namely the interruptions of DNA tracts and the absence of genotyping for some SNPs in some individuals.

DNA tract interruptions truncate MLD blocks and make their detection difficult. The number, size and location of these interruptions will have an effect on the detection of MLD blocks and thus will alter the L_c' distribution. To handle the effect of interruptions, we placed the interruptions at the same positions in our simulated chromosome as in the real chromosome.

As for the partial genotyping issue, we artificially lowered the genotyping quality in the simulated chromosomes. We used the empirical distribution of missing individuals (e.g. chr1 of *Gorilla gorilla*, Fig. 7) to pick random positions in the simulated chromosome and erase the genotypes of some individuals.

Mutation rate and recombination rate

To cope properly with the issue of genotyping, we simulated chromosomes with the same number of mutations and the same MLD length mean as in the studied data set. We use the Watterson estimator (Watterson, 1975) for the mutation rate and fixed the recombination rate so that simulated and real chromosomes had the same average length of MLD block.

6.2. Application to Chr1 of Gorilla gorilla gorilla

We applied this methodology on chromosome 1 of twenty-three unrelated western lowland Gorillas (*Gorilla gorilla gorilla*) from the Great Ape Genome Project (Prado-Martinez et al., 2013). The chromosomes have 247,249,719 base pairs. The 23.1% of sites that are considered "low coverage"(Prado-Martinez et al., 2013) divide the chromosome alignment into 6,277,293 uninterrupted stretches. The 5,388,083 interruptions due to a single site were not considered as *interruptions*. To speed up simulations, we considered stretches longer than 499 sites, as smaller stretches often carry no entire MLD block. We chopped chromosome 1 using a window of 150 polymorphic sites, into 7082 MLD blocks with an average length of 307.897 bp.

Distribution of L_c'

The distribution of L_c' for our sample of gorilla sequences has an excess of small and long MLD blocks compared to the L_c' distribution of a constant population with the same characteristics (same number of mutations and same average length of MLD blocks) (Fig. 8). The excess of small blocks is even larger than what we see in simulated declines (see above). The truncation of long MLD blocks due to the inclusion of low quality of genotyping can potentially inflate this excess.

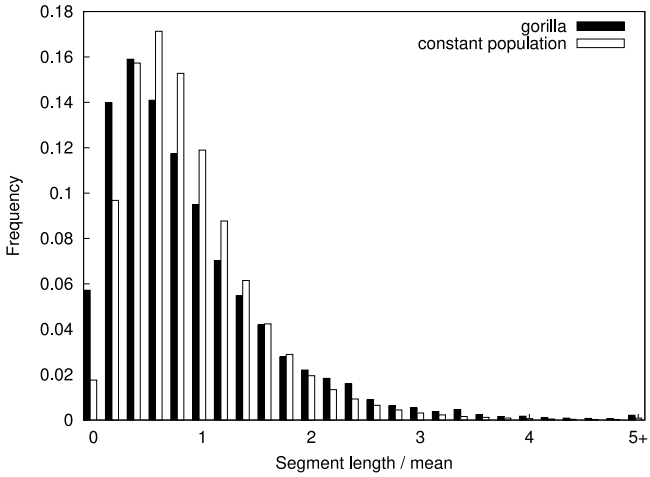


Fig. 8. Distribution of L_c' for a population of constant size (white, mutation rate = 0.000375, recombination rate = 0.012) and for the chromosome 1 of the gorillas (black).

With the low quality of genotyping and the chosen mutation and recombination rates, the threshold value $f_c^{5\%}$ is 0.041627. As we measure $f_c^{gor} = 0.0631178$ for the gorillas, the test significantly rejects H_0. However, it is possible that other designs of similar tests (tweaking the lower or upper bounds) may be more relevant to analyze demography from low quality chromosome alignments.

However, and this may be even more important, misspecifications of the model can also make the test significant. Among all, we have chosen to explore the impact of recovery after the decline and of spatial structure.

7. Misspecification of H_1

To appreciate how the f_c test, that was specifically designed to detect population decline, is sensitive to other violations of H_0, we explored their sensitivity to a scenario of bottleneck (decline followed by recovery) and to a scenario with structure but no demography.

7.1. Bottleneck

In the bottleneck scenario, we model a population that experienced a sudden strong decline ($\kappa = 10$) at time $\tau = 1$ in the past and recovered to its original size after a duration of $x \in [0, 1]$. If $x = 0$, there is no population decline. If $x = 1$, the population has not recovered and the bottleneck scenario is identical to our original H_1. When the bottleneck lasts long enough ($x > 0.02$), it is detected by the f_c test (Fig. 9(b)). On the contrary, when the bottleneck is too short ($x < 0.02$), the distribution of L_c' is similar to the one under H_0 (Fig. 9(a)). This shows that even if the population has recovered, the signal of decline will be observable in the excess of short and long MLD blocks.

7.2. Island-mainland structure

Structured populations generate signals of population size change, even when the population is stationary (Mazet et al., 2015). For example if the size of the sample is $n = 2$, for any population model with spatial structure, there exists a model without structure but specifically designed variations of population size which has the same distributions of coalescence times (Mazet et al., 2016). We consider here a larger sample size ($n = 10$, as in the other scenarios). We assume that genomes are sampled from an island with population size N and the island receives migrants with individual rate m from the mainland, which has population size $10N$. The shape of the distribution of L_c' is impacted by the migration rate and the ratio of population sizes between the island and the mainland (data not shown). When the migration rate gets too small ($m < 0.001$) or too large ($m > 10$), the distribution of L_c' is the same as under H_0 (Fig. 9(c)). For intermediate values (*i.e.* $0.001 < m < 10$), an excess of short and long blocks will be observed (Fig. 9(d)). However, the shape of the distribution of L_c' is visually different from the one of a declining population, that is, the excess of small blocks is higher than under a declining scenario. For a value of $m = 0.1$, the proportion of blocks between 0 and 0.1 times the mean is significantly higher than the proportion of blocks between 0.1 and 0.2 times the mean, which is not observed for the distribution of block lengths under decline. This suggests that the distribution could be used to differentiate between the effects of demography and structure.

8. Discussion

We have explored the impact of demography, more specifically recent population decline, on the pattern of recombination in a sample of n genomes, where $n \gg 2$. We have shown that the distribution of the distances between recombination breakpoints (MRF block lengths) is strongly affected by the demography. More specifically, a decline will result in an overdispersion of the distribution, that is, a relative excess of short and long blocks. As most recombination breakpoints are difficult, and sometimes impossible, to detect in a sequence alignment, we have proposed to restrict ourselves to the ones that can be detected using the four-gamete test. These detectable breakpoints delineate blocks in full linkage disequilibrium that we named MLD blocks.

Although different from the distribution of MRF blocks, the distribution of MLD block lengths is also overdispersed when the population has been declining recently. Using simple tests based on an excess of small and long blocks (f and f_c), one can detect declines for a wide range of different dates and strengths.

Surprisingly, the f_c test based on MLD blocks has more power for very recent declines ($\tau \approx 0.01$) than the f test based on MRF blocks. The past demography of the population impacts the distribution of the length L of MRF blocks but also the fraction p of MRF breakpoints that also correspond to MLD breakpoints. When recombination occurs at distant times when only $k \ll n$ ancestor lineages are present (*i.e.* the most ancient times of the tree), it rarely produces incompatibilities detectable with the four-gamete test (never when $k = 2$). For a declining population, these ancient lineages have longer branches than the ones of a constant population scenario, so that recombination events occur more frequently in these lineages. This results in a smaller p for declining populations and thus in more numerous (ancient, small) MRF blocks per MLD block. The relative abundance of long recent MLD blocks becomes thus more important in the distribution. This effect fades away for distant declines. In summary, the effect of recent declines on the L_c distribution is the result of both a change in the L distribution and a change in the fraction p of breakpoints detected, which can explain the difference in power between the f and the f_c tests.

We also show that using the f_c statistic, the decline can still be detected even if the population has recently recovered its original size (bottleneck scenario). Finally, we showed that local sampling of a small deme with constant size also leads to rejection of H_0 for f_c but that the distribution of block lengths seems distorted in a way that can help distinguish the two scenarios. We leave this for future work.

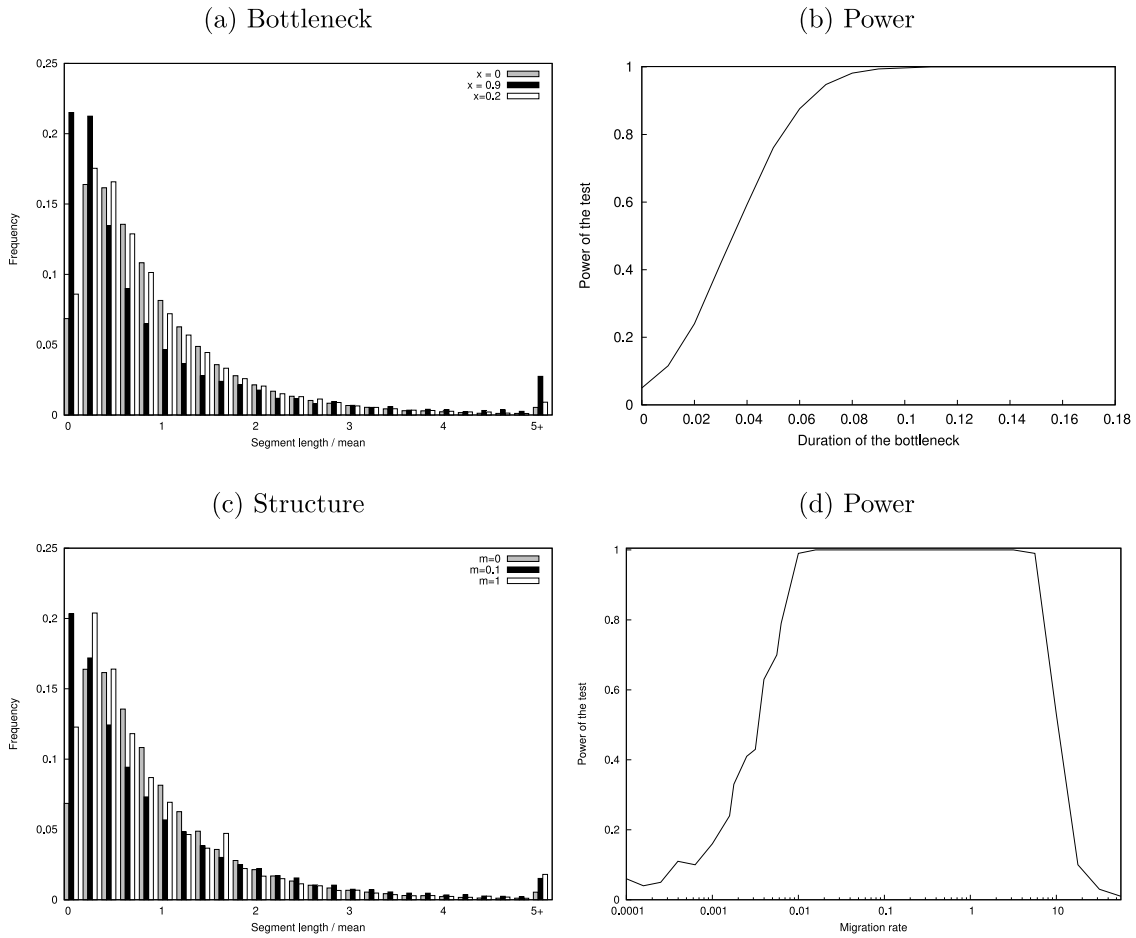


Fig. 9. Distribution of L_c' and power of the f_c test under two alternative scenarios. (a) In the bottleneck scenario, the population size is constant equal to N except during the period $[1-x, 1]$ (measured in units of N generations backwards from the present) during which it equals $N/10$, with $x = \{0, 0.2, 0.9\}$ ($x = 1$ corresponds to decline without recovery). (b) Power of the f_c test on the bottleneck population as a function of the duration of the bottleneck, $x \in [0, 0.18]$. (c) In the island-mainland scenario, the population size is constant equal to N (island) and receives migrants at individual rate $m = \{0, 0.1, 1\}$ from a population of size $10N$ (mainland). (d) Power of f_c test as a function of m, the migration rate from the mainland to the island.

One interesting advantage of using the f and f_c tests is their efficiency in computing time, such that they can scale up to a very large sample of long genomes. For example, the chopping of the entire human chromosome 1 (1,636,975 SNPs) for a sample of 10 unphased genomes takes 16 s on a laptop (with an Intel Core i7 processor running macOS High Sierra). This very short computing time is an interesting asset of this test compared to other methods based on variations (e.g. in SNP density) induced by recombination events (Li and Durbin, 2011; Palamara et al., 2012; MacLeod et al., 2013; Harris and Nielsen, 2013; Browning and Browning, 2015). The main choice that influences the computation time of the chopping algorithm is the number of sites considered for the chopping window. An increase in the chopping window size will increase the number of sites to test for incompatibility.

In the theoretical assessment of the f_c test, we have made the assumption that the recombination rate is constant along the genome and that entire genomes are aligned. Let us discuss the limits of these assumptions.

First, the recombination rate is known to vary along the genome, especially in regions of high recombination known as recombination hot-spots. It could be possible to integrate these variations via the knowledge of the recombination map. Indeed, if the recombination rate is twice higher in a given region of the genome, MRF blocks will be twice smaller, so we can correct this distortion by multiplying all MRF block lengths by 2.

Another issue of the test based on MLD block lengths is the need of whole genome data. For normalization of the MLD block distribution, the average length of a block is needed. If the whole distribution of MLD block is not available, it can compromise the estimate of the average length, and so can compromise the test based on the normalized distribution. The test requires genome data with good SNP quality for all the individuals.

The f_c test is a genome-wide approach that can detect population decline that started even very recently, down to orders of $\tau = 0.01N_0$, where N_0 is the current (effective) population size. This corresponds to very recent times, in particular when considering endangered populations. For example, there are approximately 600 mature mountain Gorilla individuals alive (IUCN Red List of 31 July 2018). Assuming that the current effective population size is a third of the mature individuals, $N_0 \approx 200$, the f_c test will detect decline as recent as $0.01 * 200 = 2$ generations ago. Great apes populations (Bonobos, Chimpanzees, Orangutans) have been sequenced (Prado-Martinez et al., 2013) and are actively re-sequenced (Gordon et al., 2016). The coverage used to sequence the data currently available is not high enough to apply our test. To infer MLD blocks, the sequenced DNA tracts need to be uninterrupted. Using these whole-genome data in higher quality, we will be able to confirm their decline thanks to the f_c test.

Giant pandas have a 'vulnerable' conservation status in the Red List. Recently they have seen their population increased (around

500 mature individuals, from IUCN website 2019). Applying the f_c test on some genomes of theirs (Zhao et al., 2013) sequenced in higher quality, will give some precise information on their demography. As the test is influenced by duration and strength of a bottleneck, the strength and the date of the increase in the population size impact the result of the test. Applying the f_c test to mammals with approximately known demography will be interesting to verify the method. However, the real asset of this test is its possible application to a much wider range of organisms. Whole-genome data start to become more and more common for non-model, non-vertebrate organisms like honeybee (Wallberg et al., 2014), as well as organisms with no conservation status such as mimicry butterflies (Zhang et al., 2017).

The chopping algorithm detects incompatibilities among trees along the genome. All the sites in a MLD block are compatible with one topology. We developed the algorithm to detect a recent change in population size. However, its use is not limited to population demography. Conflicting genealogies are also present in phylogenetic inference (Maddison, 1997). This algorithm could be used to partition the genome according to compatible trees before estimating the trees.

Recombination and mutation events leave a joint imprint on genomes which depends notably on the demography of the population. Their frequency and locations carry information about the past history of this population (decline, bottleneck, structure etc.). Using MLD breakpoints to chop genomes gives insights into this history and may be used to gain further information on other aspects impacting the frequency of recombination events through time and along the genome (e.g. hitch-hiking due to selection).

Acknowledgments

E.K. is funded by the PhD program ‘Interfaces pour le Vivant’ of Sorbonne Université, France. G.A., A.L. and E.K. thank the *Center for Interdisciplinary Research in Biology*, France and the *Fondation François Sommer*, France for funding.

References

Adams, A.M., Hudson, R.R., 2004. Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. Genetics 168 (3), 1699–1712.

Alonso-Blanco, C., Andrade, J., Becker, C., Bemm, F., Bergelson, J., Borgwardt, K.M., Cao, J., Chae, E., Dezwaan, T.M., Ding, W., Ecker, J.R., Exposito-Alonso, M., Farlow, A., Fitz, J., Gan, X., Grimm, D.G., Hancock, A.M., Henz, S.R., Holm, S., Horton, M., Jarsulic, M., Kerstetter, R.A., Korte, A., Korte, P., Lanz, C., Lee, C.-R., Meng, D., Michael, T.P., Mott, R., Muliyati, N.W., Nägele, T., Nagler, M., Nizhynska, V., Nordborg, M., Novikova, P.Y., Picó, F.X., Platzer, A., Rabanal, F.A., Rodriguez, A., Rowan, B.A., Salomé, P.A., Schmid, K.J., Schmitz, R.J., Seren, Ü., Sperone, F.G., Sudkamp, M., Svardal, H., Tanzer, M.M., Todd, D., Volchenboum, S.L., Wang, C., Wang, G., Wang, X., Weckwerth, W., Weigel, D., Zhou, X., 2016. 1,135 genomes reveal the global pattern of polymorphism in arabidopsis thaliana. Cell 166 (2), 481–491.

Ball, F., Stefanov, V.T., 2005. Evaluation of identity-by-descent probabilities for half-sibs on continuous genome. Math. Biosci. 196 (2), 215–225. http://dx.doi.org/10.1016/j.mbs.2005.04.005.

Barnosky, A.D., Matzke, N., Tomiya, S., Wogan, G.O.U., Swartz, B., Quental, T.B., Marshall, C., McGuire, J.L., Lindsey, E.L., Maguire, K.C., Mersey, B., Ferrer, E.A., 2011. Has the earth’s sixth mass extinction already arrived? Nature 471 (7336), 51–57.

Beichman, A.C., Huerta-Sanchez, E., Lohmueller, K.E., 2018. Using genomic data to infer historic population dynamics of nonmodel organisms. Annu. Rev. Ecol. Evol. Syst. 49 (1), 433–456.

Browning, S.R., Browning, B.L., 2010. High-resolution detection of identity by descent in unrelated individuals. Am. J. Hum. Genet. 86 (4), 526–539.

Browning, B.L., Browning, S.R., 2013. Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics 194 (2), 459–471.

Browning, S.R., Browning, B.L., 2015. Accurate non-parametric estimation of recent effective population size from segments of identity by descent. Am. J. Hum. Genet. 97 (3), 404–418.

Carmi, S., Wilton, P.R., Wakeley, J., Pe’er, I., 2014. A renewal theory approach to IBD sharing. Theor. Popul. Biol. 97, 35–48. http://dx.doi.org/10.1016/j.tpb.2014.08.002.

Ceballos, G., Ehrlich, P.R., Dirzo, R., 2017. Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines. Proc. Natl. Acad. Sci. 114 (30), E6089–E6096.

Chapman, N.H., Thompson, E.A., 2003. A model for the length of tracts of identity by descent in finite random mating populations. Theor. Popul. Biol. 64 (2), 141–150.

Chikhi, L., Rodríguez, W., Grusea, S., Santos, P., Boitard, S., Mazet, O., 2018. The IICR (inverse instantaneous coalescence rate) as a summary of genomic diversity: insights into demographic inference and model choice. Heredity 120 (1), 13–24.

Donnelly, K.P., 1983. The probability that related individuals share some section of genome identical by descent. Theor. Popul. Biol. 23 (1), 34–63. http://dx.doi.org/10.1016/0040-5809(83)90004-7.

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V.C., Foll, M., 2013. Robust demographic inference from genomic and SNP data. PLoS Genet. 9 (10), e1003905.

Fu, Y.X., 1995. Statistical properties of segregating sites. Theor. Popul. Biol. 48 (2), 172–197.

Gibbs, R.A., Boerwinkle, E., Doddapaneni, H., Han, Y., Korchina, V., Kovar, C., Lee, S., Muzny, D., Reid, J.G., et al., 2015. A global reference for human genetic variation. Nature 526 (7571), 68–74.

Gordon, D., Huddleston, J., Chaisson, M.J.P., Hill, C.M., Kronenberg, Z.N., Munson, K.M., Malig, M., Raja, A., Fiddes, I., Hillier, L.W., Dunn, C., Baker, C., Armstrong, J., Diekhans, M., Paten, B., Shendure, J., Wilson, R.K., Haussler, D., Chin, C.-S., Eichler, E.E., 2016. Long-read sequence assembly of the gorilla genome. Science 352 (6281).

Griffiths, R., Marjoram, P., 1997. An ancestral recombination graph. In: Progress in Population Genetics and Human Evolution. Springer, pp. 257–270.

Grusea, S., Rodríguez, W., Pinchon, D., Chikhi, L., Boitard, S., Mazet, O., 2019. Coalescence times for three genes provide sufficient information to distinguish population structure from population size changes. J. Math. Biol. 78 (1–2), 189–224.

Gusev, A., Lowe, J.K., Stoffel, M., Daly, M.J., Altshuler, D., Breslow, J.L., Friedman, J.M., Pe’er, I., 2009. Whole population, genome-wide mapping of hidden relatedness. Genome Res. 19 (2), 318–326.

Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H., Bustamante, C.D., 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 5 (10), e1000695.

Harris, K., Nielsen, R., 2013. Inferring demographic history from a spectrum of shared haplotype lengths. PLoS Genet. 9 (6), e1003521.

Hayes, B.J., Visscher, P.M., McPartlan, H.C., Goddard, M.E., 2003. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. Genome Res. 13 (4), 635–643.

Hey, J., Wakeley, J., 1997. A coalescent estimator of the population recombination rate. Genetics 145 (3), 833–846.

Hill, W.G., Robertson, A., 1968. Linkage disequilibrium in finite populations. Theor. Appl. Genet. 38 (6), 226–231.

Hollenbeck, C.M., Portnoy, D.S., Gold, J.R., 2016. A method for detecting recent changes in contemporary effective population size from linkage disequilibrium at linked and unlinked loci. Heredity 117 (4), 207–216.

Hudson, R.R., Kaplan, N., 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111 (1), 147–164.

Kelleher, J., Etheridge, A.M., McVean, G., 2016. Efficient coalescent simulation and genealogical analysis for large sample sizes. PLoS Comput. Biol. 12 (5), e1004842.

Kingman, J.F.C., 1982. The coalescent. Stochastic Process. Appl. 13 (3), 235–248.

Lapierre, M., Lambert, A., Achaz, G., 2017. Accuracy of demographic inferences from the site frequency spectrum: the case of the yoruba population. Genetics 206 (1), 439–449.

Lewontin, R.C., ichi Kojima, K., 1960. The evolutionary dynamics of complex polymorphisms. Evolution 14 (4), 458–472.

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

MacLeod, I.M., Larkin, D.M., Lewin, H.A., Hayes, B.J., Goddard, M.E., 2013. Inferring demography from runs of homozygosity in whole-genome sequence, with correction for sequence errors. PLoS Comput. Biol. 30 (9), 2209–2223.

MacLeod, I.M., Meuwissen, T.H.E., Hayes, B., Goddard, M.E., 2009. A novel predictor of multilocus haplotype homozygosity: comparison with existing predictors. Genet. Res. 91 (6), 413–426.

Maddison, W.P., 1997. Gene trees in species trees. Syst. Biol. 46 (3), 523–536.

Marjoram, P., Wall, J.D., 2006. Fast “coalescent” simulation. BMC Genet. 7, 16.

Marth, G.T., Czabarka, E., Murvai, J., Sherry, S.T., 2004. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. Genetics 166 (1), 351–372.

Mazet, O., Rodríguez, 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.

Mazet, O., Rodríguez, W., Grusea, S., Boitard, S., Chikhi, L., 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.T., Cardin, N.J., 2005. Approximating the coalescent with recombination. Philos. Trans. R Soc. London [Biol.] 360 (1459), 1387–1393.

Díez-del Molino, D., Sánchez-Barreiro, F., Barnes, I., Gilbert, M.T.P., Dalén, L., 2018. Quantifying temporal genomic erosion in endangered species. Trends Ecol. Evol. 33 (3), 176–185. http://dx.doi.org/10.1016/j.tree.2017.12.002.

Palamara, P.F., Lencz, T., Darvasi, A., Pe'er, I., 2012. Length distributions of identity by descent reveal fine-scale demographic history. Am. J. Hum. Genet. 91 (5), 809–822.

Patin, E., Siddle, K.J., Laval, G., Quach, H., Harmant, C., Becker, N., Froment, A., Régnault, B., Lemée, L., Gravel, S., et al., 2014. The impact of agricultural emergence on the genetic history of African rainforest hunter-gatherers and agriculturalists. Nature Commun. 5 (1).

Prado-Martinez, J., Sudmant, P.H., Kidd, J.M., Li, H., Kelley, J.L., Lorente-Galdos, B., Veeramah, K.R., Woerner, A.E., O'Connor, T.D., Santpere, G., et al., 2013. Great ape genetic diversity and population history. Nature 499 (7459), 471–475.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., Sham, P.C., 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81 (3), 559–575.

Régnier, C., Achaz, G., Lambert, A., Cowie, R.H., Bouchet, P., Fontaine, B., 2015. Mass extinction in poorly known taxa. Proc. Natl. Acad. Sci. 112 (25), 7761–7766.

Ringbauer, H., Coop, G., Barton, N.H., 2017. Inferring recent demography from isolation by distance of long shared sequence blocks. Genetics 205 (3), 1335–1351.

Rodrigues, A.S.L., Pilgrim, J.D., Lamoreux, J.F., Hoffmann, M., Brooks, T.M., 2006. The value of the IUCN Red List for conservation. Trends Ecol. Evolut. 21 (2), 71–76.

Rodríguez, W., Mazet, O., Grusea, S., Arredondo, A., Corujo, J.M., Boitard, S., Chikhi, L., 2018. The IICR and the non-stationary structured coalescent: towards demographic inference with arbitrary changes in population structure. Heredity 121 (6), 663–678.

Sánchez-Bayo, F., Wyckhuys, K.A.G., 2019. Worldwide decline of the entomofauna: A review of its drivers. Biol. Cons. 232, 8–27.

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

Sheehan, S., Harris, K., Song, Y.S., 2013. Estimating variable effective population sizes from multiple genomes: A sequentially Markov conditional sampling distribution approach. Genetics 194 (3), 647–662.

Stam, P., 1980. The distribution of the fraction of the genome identical by descent in finite random mating populations. Genet. Res. 35 (2), 131–155.

Stefanov, V.T., 2000. Distribution of genome shared identical by descent by two individuals in grandparent-type relationship. Genetics 156 (3), 1403–1410.

Terhorst, J., Kamm, J.A., Song, Y.S., 2017. Robust and scalable inference of population history from hundreds of unphased whole genomes. Nature Genet. 49 (2), 303–309.

Tiret, M., Hospital, F., 2017. Blocks of chromosomes identical by descent in a population: Models and predictions. Plos One 12 (11), e0187416.

van der Valk, T., Díez-del Molino, D., Marques-Bonet, T., Guschanski, K., Dalén, L., 2019. Historical genomes reveal the genomic consequences of recent population decline in eastern gorillas. Curr. Biol. 29 (1), 165–170.e6. http://dx.doi.org/10.1016/j.cub.2018.11.055.

Wallberg, A., Han, F., Wellhagen, G., Dahle, B., Kawata, M., Haddad, N., Simões, Z.L.P., Allsopp, M.H., Kandemir, I., De la Rúa, P., Pirk, C.W., Webster, M.T., 2014. A worldwide survey of genome sequence variation provides insight into the evolutionary history of the honeybee *Apis mellifera*. Nature Genet. 46 (10), 1081–1088.

Watterson, G.A., 1975. On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. 7 (2), 256–276. http://dx.doi.org/10.1016/0040-5809(75)90020-9.

Wiuf, C., Hein, J., 1999. The ancestry of a sample of sequences subject to recombination. Genetics 151 (3), 1217–1228.

Wright, S., 1931. Evolution in mendelian populations. Genetics 16 (2), 97–159.

Zhang, W., Westerman, E., Nitzany, E., Palmer, S., Kronforst, M.R., 2017. Tracing the origin and evolution of supergene mimicry in butterflies. Nature Commun. 8 (1).

Zhao, S., Zheng, P., Dong, S., Zhan, X., Wu, Q., Guo, X., Hu, Y., He, W., Zhang, S., Fan, W., Zhu, L., Li, D., Zhang, X., Chen, Q., Zhang, H., Zhang, Z., Jin, X., Zhang, J., Yang, H., Wang, J., Wang, J., Wei, F., 2013. Whole-genome sequencing of giant pandas provides insights into demographic history and local adaptation. Nature Genet. 45 (1), 67–71.

2.2.3 Additional information

2.2.3.1 Cost of simulation

For each change in sample size or null model, one needs to redo the simulations to compute a p-value. It is not possible to compute the MLD p-value analytically (although it may be possible for MRF blocks). The simulation takes some CPU time (it can be quite short, see below) but it is also interesting to change the null model to fit the population characteristics.

The CPU time needed to calculate the p-value depends on the way genomes are simulated. Using msprime (Kelleher et al. 2016) and a simple constant population model, simulating a sample of 10 genomes with at least 10^5 MRF blocks takes 47 sec on a personal computer (with an Intel Core i7 processor running macOS High Sierra). The 10,000 samples of 10 genomes needed for the p-value, take 47*10,000s, slightly less than 5.5 days. All samples of genomes must be independent, so they can easily be distributed in different jobs. With 20 jobs running in parallel, it would take a little more than 6.5 hours. For $n = 100$, it is even quicker (as MRF blocks are shorter), taking 18sec to simulate 100 genomes with $\geq 10^5$ MRF blocks, so 50 hours to calculate the p-value or 2.5 hours with 20 jobs in parallel.

2.2.3.2 Impact of the number of blocks and sampled individuals

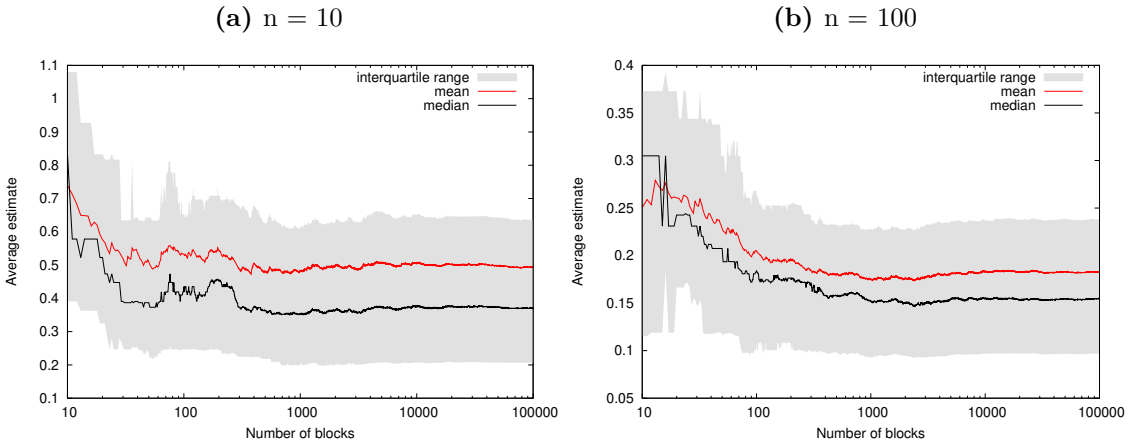


Figure 2.5: Average estimate of the mean (red), the median (black), the interquartile range (grey) of the L_c distribution in function of the number of MLD blocks for 10 sampled individuals (2.5a) and 100 sampled individuals (2.5b), ($\mu = 10, \rho = 1$).

The limiting parameter for applying the test is the number of blocks. The number of blocks will affect the distribution estimate (Fig 2.5) that we used to apply our test. Fewer blocks lowers the quality of the empirical distribution of block length. The relevant number of blocks depends on the number of sampled individuals, as it is easier to estimate for 100 sampled individuals (Fig 2.5b) than for 10 sampled individuals (Fig 2.5a). The number of sampled individuals has

an effect on the length of the blocks. The smaller the sample size the longer the average length of blocks (Fig 2.6).

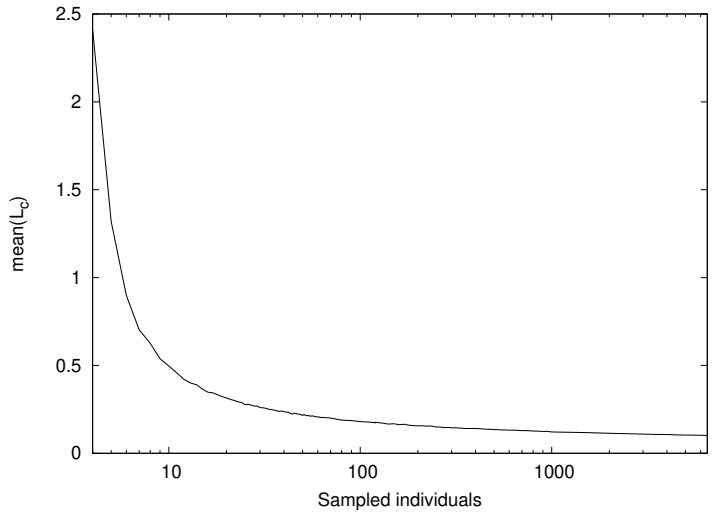


Figure 2.6: Mean L_c per number of sampled individuals (n) for a simulated genome with at least 10^5 MLD blocks, ($\mu = 10, \rho = 1$).

Interestingly, this negative correlation relates to total tree lengths T increasing with the logarithm of the number of sampled individuals n: $T \propto log(n)$.

When there are recombinations between every pair of consecutive SNPs, we cannot detect all of them and thus the fraction of recombination events detected will be low (see Fig 4.a of the paper); however the distribution of detected events (and so of MLD breakpoints) will always depend on the demographic scenario. As the number of blocks also depends on the mutation rate / recombination rate (μ/ρ) ratio (see Fig 4.a of the paper), it is difficult to recommend a sample size without knowing this parameter.

The main difference between considering 1,000 or 6,000 sampled individuals is the CPU time needed for computing the p-value ($\approx \times 6$). Using too many sampled individuals will not impact the result of the test but the CPU time needed to perform it. In that regard, it can be interesting to use a smaller sample size and re-sample different individuals to do the test multiple times. It will use the information of all the sampled individuals without extra CPU time.

As the number of blocks is the really limiting parameter, we recommend to choose the number of individuals as a trade off between the number of MLD blocks and the CPU time needed for the p-value.

2.3 Inférence combinant SFS et blocs MLD

2.3.1 Introduction

Dans cette section, nous cherchons à améliorer notre méthode de détection de déclin de population décrite dans Testing for population decline using maximal linkage disequilibrium blocks. Notre but est d'inférer les différents paramètres décrivant les déclins de populations afin d'attribuer des statuts de conservation aux espèces. En effet, une méthode statistique capable de détecter un déclin démographique récent à partir de données génomiques pourrait permettre d'attribuer un statut de conservation à tout type d'espèce (cf Contexte).

En introduction, nous avons vu que la démographie a un fort impact sur la déformation du SFS par rapport au modèle neutre, et qu'il est donc communément utilisé pour effectuer des inférences démographiques. Toutefois, il souffre de problèmes d'identifiabilité de modèles : de nombreux scénarios pouvant produire le même SFS. Il peut donc être utile de combiner le SFS à une autre statistique. Comme nous l'avons vu dans le chapitre précédent, la distribution des longueurs de blocs MLD est également particulièrement informative concernant la démographie passée d'une population. Il est donc intéressant de développer une méthode d'inférence s'appuyant sur cette distribution. Cependant, la méthode décrite dans le chapitre précédent est difficilement applicable en l'état puisqu'elle repose sur une normalisation par la valeur moyenne des longueurs de blocs MLD et qu'elle n'est pas estimable à partir des données disponibles actuellement. En effet, un génome n'étant jamais séquencé d'un bout à l'autre, la distribution des longueurs de blocs MLD n'est jamais complète et leur longueur moyenne n'est pas estimable. Plus un bloc est long, plus la probabilité qu'une partie du bloc soit inclue dans une zone non couverte par le séquençage est importante, le bloc n'est alors pas mesurable. Les très longs blocs sont rarement mesurables, la longueur moyenne des blocs MLD est donc sous-estimée. La normalisation par la longueur moyenne des blocs MLD ne serait pas nécessaire si les taux de mutation et de recombinaison étaient connus : il serait alors possible de comparer une distribution de longueurs de blocs MLD à une distribution constante sous H_0. Malheureusement, ces taux sont généralement inconnus. Il est donc nécessaire de les inférer. L'utilisation de la distribution de longueurs de blocs MLD demandant de nombreux paramètres à inférer, il serait également utile de la combiner avec une autre statistique.

Nous cherchons donc ici à combiner deux statistiques : le SFS et la distribution de longueurs de blocs MLD pour inférer le déclin récent de population. Nous nous intéressons à l'inférence d'un changement brutal de taille de population à un temps τ d'une intensité κ. Dans un premier temps, nous considérons le cas de taux de mutation et de recombinaison connus et inférons seulement le temps et l'intensité du déclin. Nous nous intéressons à l'impact des valeurs de temps et d'intensité de déclin ainsi qu'aux valeurs des taux de mutation et de recombinaison sur la qualité de l'inférence. Finalement, nous inférons les quatre paramètres suivants : temps du déclin, intensité du déclin, taux de mutation et taux de recombinaison dans

différentes conditions.

2.3.2 Matériel et méthode

2.3.2.1 Modèle

L'ensemble de l'étude a été faite à l'aide de simulations effectuées grâce à notre simulateur de SMC' (Marjoram and Wall 2006) que j'ai implémenté en C.

Nous considérons ici un scénario de déclin soudain de population. On suppose dans chaque simulation qu'au temps présent, la taille de population est égale à N, un déclin soudain de taille de population d'intensité κ survient à un temps τ dans le passé, exprimé en unités de N générations. La population était de taille κN avant le temps de changement de taille de population (dans le passé). Le nombre d'individus échantillonnés dans la population est n, on prendra $n = 10$.

2.3.2.2 Inférence

Pseudo-Vraisemblance

SFS En considérant les sites comme étant indépendants, la pseudo-vraisemblance du SFS se calcule de la manière suivante :

$$PsL(s_1, \ldots, s_{n-1}) = \frac{s!}{s_1! \ldots s_{n-1}!} \prod_{i=1}^{n-1} \left(\frac{E[T_i]}{E[T_{tot}]} \right)^{s_i}$$

avce $s_1, \ldots s_{n-1}$ le SFS observé, $s = \sum_i^{n-1} s_i$ et $E[T_i]$ l'espérance de la somme des longueurs de branches supportant i feuilles et $E[T_{tot}]$ l'espérance de la somme de toutes les longueurs de branches.

La fraction $\frac{s!}{s_1! \ldots s_{n-1}!}$ étant difficile à calculer et constante pour un observé donné, elle n'est pas utilisée lors de notre recherche de maximum de vraisemblance. Les espérances de T_i et T_{tot} sont obtenues par simulation.

Distribution des longueurs de blocs MLD Afin d'étudier la distribution des longueurs de segments MLD x, il est nécessaire de regrouper les longueurs de blocs par catégorie de taille. La largeur de chaque catégorie de taille est choisie en fonction de la loi de Freedman–Diaconis :

$$\text{largeur catégorie} = 2\frac{\text{EI}(x)}{\sqrt[3]{n}},$$

avec EI l'espace interquartile de la distribution observée et n le nombre d'observations. La distribution est décrite par 25 catégories de taille. Les blocs MLD sont considérés indépendants, leurs longueurs également. La pseudo-vraisemblance se calcule donc de la manière suivante :

$$PsL(x_1, \ldots, x_{25}) = \frac{x!}{x_1! \ldots x_{25}!} \prod_{i=1}^{25} E\left[\frac{\#\text{MLD}_i}{\#\text{MLD}_{tot}} \right]^{x_i},$$

avec x_i le nombre de blocs MLD observés possédant une longueur de la catégorie i, et $E\left[\frac{\#\text{MLD}_i}{\#\text{MLD}_{tot}}\right]$ l'espérance du ratio du nombre de MLD dans la catégorie i sur le nombre de MLD totaux. Cette espérance est obtenue par simulation.

La fraction $\frac{x!}{x_1!\dots x_{25}!}$ étant également difficile à calculer et constante pour un observé donné, elle n'est pas utilisée lors de notre recherche de maximum de vraisemblance.

Pseudo-vraisemblance jointe Nous considérons le SFS et la distribution de longueurs de blocs MLD comme indépendants, leur pseudo-vraisemblance jointe est donc :

$$PsL\left((s_1,...,s_{n-1}),(x_1,...,x_{25})\right) = \frac{s!}{s_1!...s_{n-1}!}\prod_{i=1}^{n-1}\left(\frac{E[T_i]}{E[T_{tot}]}\right)^{s_i}\frac{x!}{x_1!...x_{25}!}\prod_{i=1}^{25}E\left[\frac{\#\text{MLD}_i}{\#\text{MLD}_{tot}}\right]^{x_i}.$$

Les log-vraisemblances seront utilisées pour les inférences de paramètres.

Méthodes d'estimation des paramètres Toutes les estimations sont faites à partir du logarithme en base 10 des paramètres. Les 4 paramètres sont inférés avec des méthodes différentes :

μ est estimé avec une version modifiée de l'estimateur de Watterson (Watterson 1975) : $\theta_{\text{modified } w} = \frac{S}{E[T_{tot}]}$ avec S le nombre de mutations observées et $E[T_{tot}]$ l'espérance de la somme des longueurs de branches de l'arbre de coalescence, obtenue par simulation suivant le scénario considéré.

ρ est estimé en maximum de vraisemblance sur la distribution des blocs MLD, utilisant le logarithme de la pseudo-vraisemblance décrite précédemment. L'optimisation de l'estimation se fait avec une méthode d'ascension de gradient. Le gradient est calculé à partir d'un point initial $\rho_{initial} = \mu$. Le gradient est suivi jusqu'à ce que la vraisemblance n'augmente plus, le maximum est alors encadré dans un intervalle. Cet intervalle est réduit de manière constante en utilisant la méthode du nombre d'or.

Notre méthode reposant sur des simulations, il existe une erreur numérique qui bruite beaucoup la surface de vraisemblance. La surface de vraisemblance est plate en son sommet, le bruit domine et engendre un gradient non nul. Pour cela nous effectuons plusieurs fois l'inférence à partir du dernier ρ estimé et l'inférence s'arrête quand le même maximum est trouvé quatre fois d'affilée.

τ & κ sont estimés conjointement, par maximum de vraisemblance utilisant la pseudo-vraisemblance jointe du SFS et de la distribution de longueurs de blocs MLD. L'optimisation est faite de la même manière que pour ρ, en deux dimensions. À partir de différents points de départ (décrits ci-dessous) pour éviter les parties

planes de la surface de vraisemblance, des gradients sont calculés, et sont suivis jusqu'à ce que la vraisemblance diminue. L'intervalle contenant le maximum de vraisemblance est réduit grâce à la méthode du nombre d'or. L'inférence s'arrête également quand le même maximum est trouvé quatre fois d'affilée. Cependant, le paysage de vraisemblance, plus complexe en deux dimensions, possèdent une surface très plate à son maximum, nous effectuons alors une marche aléatoire à partir du dernier couple de points inféré pour mieux échantillonner le paysage.

1. Initialisation. On part des points τ_0, κ_0 inférés. On appelle (τ', κ') le couple de valeurs testé dans le cadre de notre marche aléatoire et (τ_t, κ_t) le couple de valeurs que prend réellement la marche au temps t.
2. Itérations (pour chaque temps t) :
   - Tirage aléatoire des nouveaux points τ', κ' à une distance des points précédents $|\sqrt{\log(\tau') + \log(\kappa')} - \sqrt{\log(\tau_t) + \log(\kappa_t)}| < 0.1$.
   - Calcul du taux d'acceptation à partir de la log-vraisemblance jointe : $\alpha = lnPsL(\tau_t, \kappa_t) - lnPsL(\tau', \kappa')$.
   - Décision : Si $\alpha > 0$, on accepte les nouveaux points $\tau_{t+1} = \tau'$ et $\kappa_{t+1} = \kappa'$. Sinon tirage aléatoire d'un nombre uniformément $u \in]0, 1]$, si $\log(u^{200}) \leq \alpha$ accepter les nouveaux points $\tau_{t+1} = \tau'$ et $\kappa_{t+1} = \kappa'$, sinon rejeter les nouveaux points et $\tau_{t+1} = \tau_t$ et $\kappa_{t+1} = \kappa_t$. Il faut utiliser une condition très stricte : $\log(u^{200}) \leq \alpha$, car les valeurs de vraisemblance sont très élevées et varient peu en proportion.

Le τ final est la moyenne des τ après 2000 itérations de la marche aléatoire. De même pour le κ final.

Algorithme d'inférence Il existe deux versions de l'algorithme d'inférence des paramètres, une où on estime les deux paramètres τ, κ (on fait l'hypothèse que μ et ρ sont connus) et une où on estime les quatre paramètres μ, ρ, τ, κ.

Inférence de τ, κ :

1. Initialisation : estimation de τ et κ avec seulement la distribution de SFS. Afin d'éviter les zones plates de la surface, quatre optimisations avec des couples initiaux différents sont effectuées :
   - $\log(\tau_0) = -2.5, \log(\kappa_0) = 2$
   - $\log(\tau_0) = 0.8, \log(\kappa_0) = 2$
   - $\log(\tau_0) = -2.5, \log(\kappa_0) = -1.8$
   - $\log(\tau_0) = 0.8, \log(\kappa_0) = -1.8$.

   Le couple avec la meilleure vraisemblance est conservé.
2. Estimation de τ et κ à partir des deux distributions (SFS et longueurs de blocs MLD)
3. si $\sigma_{-1} = \emptyset$ ou $|\sigma_{-1} - \sigma| > \varepsilon$ avec ε la précision choisie et σ_{-1} la norme à l'itération précédente, $\sigma_{-1} = \sigma$ et on retourne à l'étape 2
   sinon l'algorithme s'arrête et renvoie les valeurs estimées.

Inférence de μ, ρ, τ, κ : La méthode d'inférence suit l'algorithme suivant :

1. Initialisation : estimation de τ et κ avec seulement la distribution de SFS. Comme nous n'utilisons pas le nombre de mutations pour calculer la pseudo-vraisemblance du SFS, il n'est pas nécessaire de connaître la valeur de μ et ρ pour inférer (τ, κ). Afin d'éviter les zones plates de la surface, quatre optimisations avec des couples initiaux différents sont effectuées :
   - $\log(\tau_0) = -2.5, \log(\kappa_0) = 2$
   - $\log(\tau_0) = 0.8, \log(\kappa_0) = 2$
   - $\log(\tau_0) = -2.5, \log(\kappa_0) = -1.8$
   - $\log(\tau_0) = 0.8, \log(\kappa_0) = -1.8$.

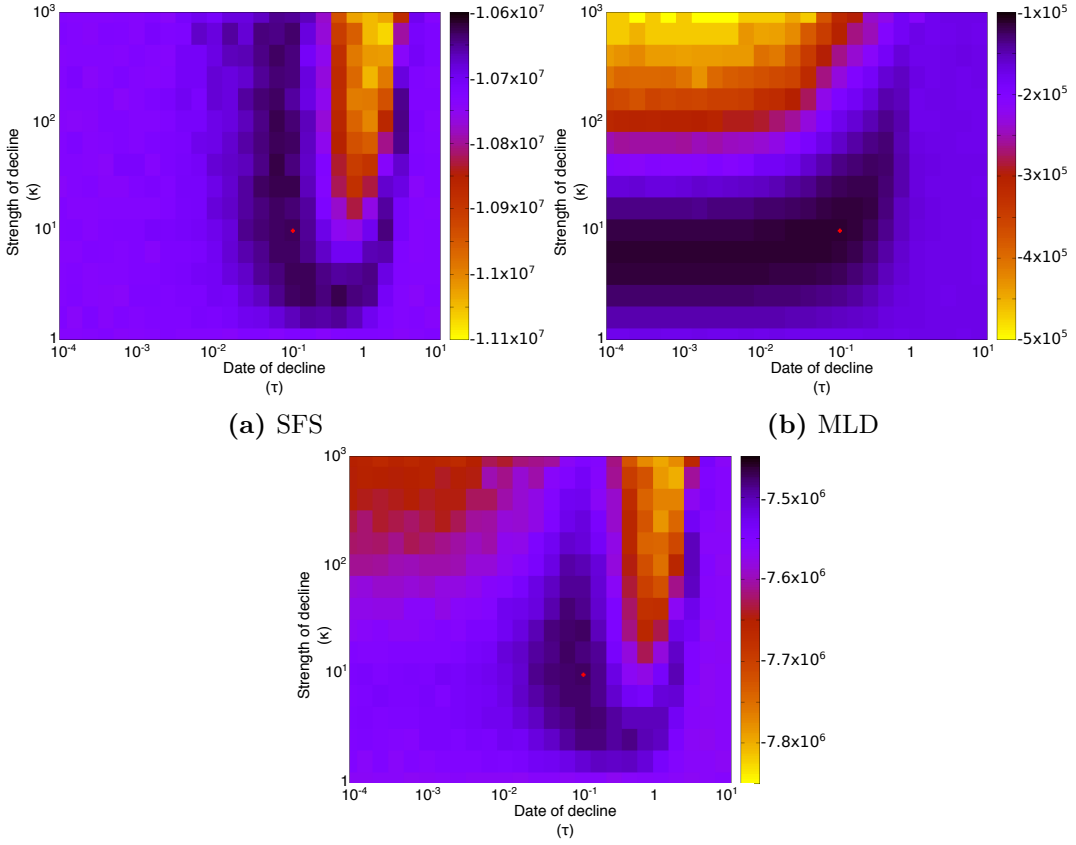
   Le couple avec la meilleure vraisemblance est conservé.
2. Estimation de μ à l'aide des valeurs estimées des paramètres τ, κ estimés,
3. Estimation de ρ à l'aide des valeurs estimées des paramètres μ, τ, κ estimés,
4. Estimation de τ et κ à partir des deux distributions (SFS et longueurs de blocs MLD), suivant les paramètres μ, ρ estimés,
5. Mesure de la norme $\sigma = |\theta|$ où $|\cdot|$ est la norme euclidienne et $\theta = (\log \mu, \log \rho, \log \tau, \log \kappa)$,
6. si $\sigma_{-1} = \emptyset$ ou $|\sigma_{-1} - \sigma| > \varepsilon$ avec ε la précision choisie et σ_{-1} la norme à l'itération précédente, $\sigma_{-1} = \sigma$ et on retourne à l'étape 2
   sinon l'algorithme s'arrête et renvoie les valeurs estimées.

Évaluation des paramètres estimés Pour évaluer notre méthode d'inférence et l'estimation des paramètres, nous simulons avec msprime (Kelleher et al. 2016) un génome d'au moins 500,000 SNPs. Sur ce génome 1,000 estimations de paramètres sont faites et comparées aux valeurs simulées.

2.3.3 Résultats

2.3.3.1 Inférence de (τ, κ), considérant μ et ρ connus

Paysage de vraisemblance Sous la condition que le couple (μ, ρ) soit connu, il est possible de visualiser le paysage de vraisemblance dans le plan (τ, κ) (Fig 2.7). Les paysages de vraisemblance des deux statistiques (SFS et distribution des longueurs de blocs MLD) ont des formes bien distinctes. En effet, le SFS a un paysage très marqué par la date de déclin τ (Fig 2.7a) alors que la distribution des MLD a un paysage très marqué par l'intensité du déclin κ (Fig 2.7b). Les informations portées par le SFS et par la distribution des longueurs de blocs MLD sont donc complémentaires. La vraisemblance jointe des deux distributions permet d'obtenir un paysage de vraisemblance moins plat, centré autour du couple (τ, κ) recherché (Fig 2.7).



(a) SFS **(b)** MLD

(c) SFS & MLD

Figure 2.7: Paysages de vraisemblance d'un scénario de déclin arrivant à un temps $\tau = 0.1$ et d'une intensité $\kappa = 10$ (point rouge), en fonction de la distribution utilisée : le SFS (a), celle des blocs MLD (b) et la combinaison des deux (c). En abscisse la date de déclin τ, en ordonnée l'intensité du déclin κ. Le minimum se situe dans les zones chaudes et le maximum dans les zones mauve foncé.

Inférences L'utilisation de la distribution jointe du SFS et de la distribution des longueurs de bloscs MLD permet d'estimer (τ, κ) précisément.

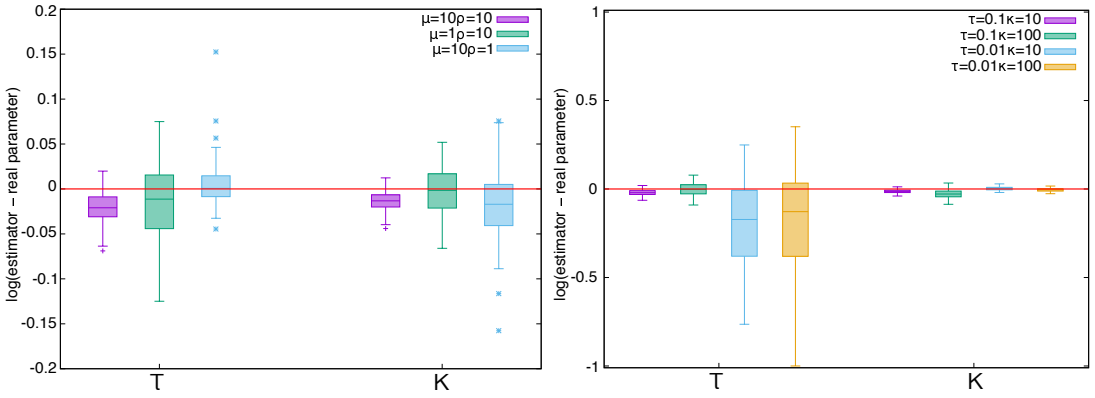
L'estimation des paramètres (τ, κ) varie avec le ratio μ/ρ (Fig 2.8a). En effet, pour un même scénario, la précision des valeurs inférées dépend du ratio μ/ρ :

— Pour un ratio $\mu/\rho = 1$, les valeurs de τ et de κ ont tendance à être sous-estimées.

— Pour un ratio $\mu/\rho = 0.1$, les estimations sont plus variables. L'intensité du déclin κ est en moyenne bien estimée.

— Pour un ratio $\mu/\rho = 10$, la date de déclin est en moyenne bien estimée avec peu de variabilité. L'intensité du déclin est, quant à elle, plus variable et peut être sous-estimée.

La capacité de notre méthode à estimer les paramètres τ et κ dépend également de la valeur de ces paramètres (Fig 2.8b). Avec un taux de mutation et un taux de recombinaison connus, l'estimation de l'intensité du déclin est très précise si elle est



(a) μ, ρ changeant **(b)** τ, κ changeant

Figure 2.8: Inférence de τ, κ, pour (a) différents couples μ, ρ : $\mu = 10, \rho = 10$ (mauve), $\mu = 1, \rho = 10$ (vert) et $\mu = 10, \rho = 1$ (bleu) (avec $\tau = 0.1$ et $\kappa = 10$), ainsi que pour (b) des couples τ, κ différents : $\tau = 0.1, \kappa = 10$ (mauve), $\tau = 0.1, \kappa = 100$ (vert), $\tau = 0.01, \kappa = 100$ (bleu) et $\tau = 0.01, \kappa = 100$ (orange) (avec $\mu = 10$ et $\rho = 10$). En abscisse le paramètre estimé, en ordonnée le $\log_{10}$ de la différence entre l'estimation et la valeur simulée.

forte ($\kappa = 10$) ou très forte ($\kappa = 100$). La date de déclin est difficilement estimable pour des valeurs très proches du présent ($\tau = 0.01$), peu importe l'intensité du déclin. Elle est facilement estimable pour des dates un peu plus lointaines, mais toujours proches du présent ($\tau = 0.1$) (Fig 2.8b).

2.3.3.2 Inférence des quatre paramètres

Les taux de mutation et de recombinaison étant rarement connus, nous allons donc estimer les quatre paramètres suivants : taux de mutation (μ), taux de recombinaison (ρ), date (τ) et intensité (κ) du déclin.

Globalement, les estimations des paramètres sont moins précises que pour les inférences avec (μ, ρ) connus.

Le scénario le mieux inféré est celui d'un déclin récent assez fort ($\mu = 10, \rho = 10, \tau = 0.1, \kappa = 10$).

L'estimation de τ est toujours centrée autour de sa valeur simulée. La variance de l'estimation dépend de la valeur du paramètre, quand la date est très proche du présent ($\tau = 0.01$), l'estimation est plus variable (Fig 2.9).

Les estimations de μ, ρ et κ ont des comportements liés. Pour certains scénarios ($\mu = 10, \rho = 10, \tau = 0.01, \kappa = 100$) et ($\mu = 10, \rho = 10, \tau = 0.1, \kappa = 100$), μ et ρ sont sur-estimés et κ sous-estimé. Alors que dans un autre ($\mu = 10, \rho = 10, \tau = 0.01, \kappa = 10$), le phénomène inverse va se produire, μ et ρ sont sous-estimés et κ sur-estimé (Fig 2.9).

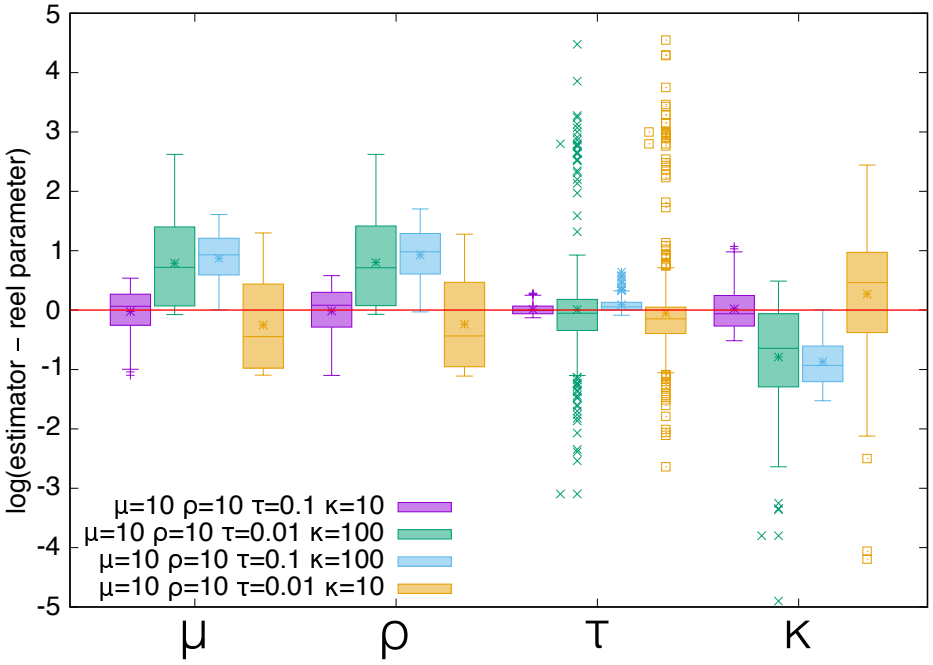


Figure 2.9: Inférence de μ, ρ, τ et κ pour $\mu = 10$, $\rho = 10$ et des couples τ, κ variables : $\tau = 0.1, \kappa = 10$ (mauve), $\tau = 0.01, \kappa = 100$ (vert), $\tau = 0.1, \kappa = 100$ (bleu) et $\tau = 0.01, \kappa = 10$ (orange). En abscisse le paramètre estimé, en ordonnée le $\log_{10}$ de la différence entre l'estimation et la valeur simulée.

2.3.4 Discussion

Inférence de la date et de l'intensité du déclin à partir du SFS et de la distribution des longueurs de blocs MLD L'intensité du déclin a un fort impact sur le nombre de mutations observées. Cependant, cet observable n'est pas pris en compte dans notre calcul de vraisemblance mesurée sur le SFS. L'intensité du déclin n'a donc pas un fort effet sur la vraisemblance du SFS. La date de déclin a toujours un impact sur le nombre de branches affectées par le changement de taille de population. Elle affecte particulièrement le SFS. La longueur des blocs MLD dépend de la taille de la population : la distribution des blocs MLD est donc fortement dépendante de l'intensité du déclin. Le temps du déclin a un plus faible impact sur la taille des blocs, ce qui affecte moins le paysage de vraisemblance. Chaque distribution a donc sa spécialisation. Il est intéressant de combiner les deux pour inférer la date et l'intensité d'un déclin.

Le scénario démographique, et donc les valeurs de (τ, κ), vont jouer sur la déformation du SFS et sur la distribution des blocs MLD. Une intensité de déclin plus importante entraîne une déformation plus importante, elle est facilement remarquable et donc inférable. À l'inverse, une date de déclin plus récente diminue la proportion des évènements de coalescence dans la population de taille actuelle, il est donc plus difficile de faire la différence entre les deux et de déterminer les paramètres de ce déclin.

Inférence de la date et de l'intensité d'un déclin, du taux de mutation et du taux de recombinaison Comme on a pu le voir, certains paramètres sont liés. Tout cela peut s'expliquer par notre méthode d'inférence. Une surestimation de μ entraîne un plus grand nombre de mutations par génération et par unité de taille de population que ce qui se produit réellement, ce qui a comme effet de sous-estimer κ. En effet, l'estimation de κ dépend de $\kappa\mu N$ (taux de mutation populationnel avant le déclin). Pour compenser la surestimation de μ, κ est sous-estimé. Plus l'écart est grand entre les deux tailles de population (plus κ est grand), plus μ peut être surestimé et donc κ sous-estimé. La surestimation de μ dépend donc de la valeur de κ. Un autre effet de la surestimation de μ est la surestimation de ρ. L'estimation de ρ dépend de la distribution des blocs MLD, dont la longueur dépend du ratio μ/ρ. Afin d'inférer le ratio réel une surestimation de μ entraîne une surestimation de ρ.

Comparaison entre les deux La différence dans la qualité d'estimation des paramètres du déclin entre notre inférence des deux paramètres ou des quatre paramètres s'explique simplement par le fait que nous utilisons les mêmes distributions : le SFS et la distribution des longueurs de blocs MLD pour estimer deux paramètres de plus. Nous perdons donc en précision pour les inférences des paramètres du déclin.

Le ratio μ/ρ a un effet important sur la détectabilité des évènements de recombinaison, et donc sur la longueur des blocs MLD. Plus il y a d'évènements de mutation par évènement de recombinaison plus les évènements de recombinaison sont détectables par le test des 4-gamètes, qui sert à découper les blocs MLD (voir Testing for population decline using maximal linkage disequilibrium blocks). Connaître les taux de mutation et de recombinaison nous laisse seulement deux paramètres à inférer, mais nous renseigne également sur ce ratio μ/ρ, qui influence la précision des estimations.

Conclusion L'utilisation jointe du SFS et de la distribution de blocs MLD permet l'inférence du taux de mutation, du taux de recombinaison ainsi que de la date et de l'intensité du déclin. Le SFS et la distribution de blocs MLD ne sont pas informatifs sur les mêmes caractéristiques. Le SFS est plus informatif sur la date de déclin, alors que la distribution des blocs de MLD est plus informative sur l'intensité du déclin. C'est bien la conjugaison des deux qui permet une inférence plus fiable.

En estimant le taux de mutation et le taux de recombinaison, la normalisation par la longueur moyenne utilisée dans Testing for population decline using maximal linkage disequilibrium blocks pour inférer le déclin n'est plus nécessaire. Il serait donc possible d'utiliser la distribution « brute » des longueurs de blocs MLD. Cependant, de nombreux facteurs, comme la qualité des données ou des histoires démographiques complexes, influencent la forme des distributions observées.

2.4 Comment expliquer les distributions de longueurs de blocs MLD observées ?

Le travail théorique a permis de montrer qu'en utilisant à la fois le SFS et la distribution des longueurs de blocs MLD il était possible d'inférer, avec une bonne précision, la date ainsi que l'intensité d'un déclin sans connaître au préalable les taux de mutation et de recombinaison de la population considérée. Naturellement, nous avons souhaité appliquer notre méthode à des données issues de populations réelles. C'est à ce moment que nous avons été confronté à des problèmes que nous n'avions pas anticipé. Cette section relate les différentes pistes que nous avons exploré afin d'expliquer les distributions observées. Bien que les résultats ne soient pas satisfaisant, la distribution de longueurs de blocs MLD reste une source d'information affectée par différents facteurs. N'étant pas capable à l'heure actuelle d'utiliser la distribution des longueurs de blocs MLD pour inférer l'histoire évolutive d'une population, cette section est la dernière utilisant directement cette information. Par la suite, nous avons décider de changer d'approche pour inférer la démographie récente d'une population.

2.4.1 Introduction

Dans cette section, nous allons étudier les distributions de longueurs de blocs MLD de populations aux histoires démographiques différentes. Notre objectif est de prendre en compte les différents facteurs influençant les longueurs de blocs MLD afin d'expliquer les distributions de longueurs de blocs MLD observées. Ces facteurs peuvent être techniques, comme la qualité de séquençage, de génotypage. En effet, les chromosomes n'étant pas séquencés d'un bout à l'autre, la distribution des longueurs de blocs sera biaisée en fonction des parties séquencées. Pour qu'un bloc MLD soit mesurable, il faut que tous les sites entre deux évènements de recombinaison détectés soient séquencés. De plus, le test des quatre-gamètes permettant le découpage des blocs MLD s'appuie sur le génotype des différents sites. Si tous les sites ne sont pas génotypés, ce déficit d'information peut jouer sur la détection des évènements de recombinaison, et donc sur la longueur des blocs. Comme vu dans la partie Testing for population decline using maximal linkage disequilibrium blocks, d'autres facteurs influencent la distribution des longueurs de blocs MLD. Il est donc nécessaire de prendre en compte les facteurs évolutifs comme la démographie et la structure de population.

Afin de pouvoir comparer les distributions entre elles, nous avons choisi des populations dont les longueurs de taille de blocs MLD devraient être de même ordre de grandeur, c'est-à-dire avec des taux de mutation, taux de recombinaison, taille de chromosome et taille efficace de population de même ordre de grandeur. Nous avons sélectionné deux populations dont la démographie est connue : une population en déclin, celle du gorille, et une population en croissance, une population humaine, celle des Yorubas.

Nous avons, dans un premier temps, comparé les distributions observées à des distributions issues de génomes simulés sous un modèle constant de population, en prenant en compte les facteurs techniques pouvant influer sur la distribution de longueurs de MLD : couverture de séquençage, génotypage, taux de recombinaison variable. Nous avons finalement considéré des scénarios plus complexes de démographie, de structure de population et d'un mélange de démographie et de structure.

2.4.2 Matériel et méthode

2.4.2.1 Données

Gorilla Nous avons sélectionné 23 génomes diploïdes de Gorille des plaines de l'Ouest (*Gorilla gorilla*) non-apparentés issus du Great Ape Project (Prado-Martinez et al. 2013).

Yoruba Nous avons sous-échantillonné 23 génomes diploïdes humains de la population des Yorubas issues du 1000 Genomes Project (Consortium 2015).

Un alignement de génome est accessible publiquement pour chacune de ces populations. Nous considérons seulement le chromosome 1 de chacune de ces espèces.

Taille efficace Comme expliqué dans l'introduction, nous avons besoin de populations aux tailles efficaces similaires. Les gorilles ont une taille efficace estimée autour de [3-5].10^4 (Prado-Martinez et al. 2013) et les Yorubas ont une taille efficace estimée autour de [3-4].10^4 (Lapierre et al. 2017).

Description des distributions Les informations sur la qualité et la couverture de séquençage proviennent des publications et analyses des auteurs des publications. Ceux-ci rendent les données accessibles publiquement.

Population	#SNP	#Breakpoints	#MLD	$\overline{L_{MLD}}$
Yoruba	1,076,319	72,411	55,334	2470.8
Gorilla	1,362,119	128,310	14,176	302.5

Table 2.1: Description des données récoltées sur le chromosome 1 de deux populations : les Yorubas et les gorilles. Le tableau comporte le nombre de SNP, le nombre d'évènements de recombinaison détectés (breakpoints), le nombre de blocs MLD et la longueur moyenne de ces blocs ($\overline{L_{MLD}}$).

2.4.2.2 Qualité des données

Génotypages L'ensemble des SNP est génotypé pour les Yorubas. Ce n'est pas le cas pour la population de gorilles. Comme indiqué dans Kerdoncuff et al. 2020, seulement 66% des sites sont génotypés pour tous les individus.

Couverture de séquençage La qualité de séquençage est assez différente entre les génomes des Yorubas et ceux des gorilles. Pour le chromosome 1 des gorilles, il y a plus d'interruptions de séquençage que de SNPs séquencés. Plus de 100,000 évènements de recombinaison sont détectés, mais seulement 14,176 blocs sont mesurables (Tab 2.1). Pour un chromosome de la même taille, 55,334 blocs sont mesurés chez les humains.

2.4.2.3 Méthodes

Simulations Afin de comparer les différentes facteurs influençant la distribution de blocs MLD, nous simulons des génomes à l'aide du logiciel *msprime* (Kelleher et al. 2016). Nous souhaitons qu'ils soient le plus ressemblant aux données. Pour cela :

- Le nombre de génomes échantillonnés est identique aux données.
- Les distributions de qualité de génotypage sont reproduites à l'identique.
- Les motifs de couverture de séquençage sont identiques.
- La taille des génomes séquencés est similaire à celle des données. Cependant, dans un souci de temps de simulation de très long génomes, les génomes sont simulés par morceaux, découpés suivant la couverture de séquençage des données. Si une partie non séquencée est supérieure à 20,000pb, nous ne simulons pas cette partie. Une autre simulation est lancée pour la suite du génome. Toutes les zones non séquencées de moins de 20,000pb sont donc séquencées, mais non considérées dans l'analyse.
- Le taux de mutation est estimé avec une version modifiée de l'estimateur de Watterson (Watterson 1975) en fonction du scénario choisi (voir Section précédente Inférence combinant SFS et blocs MLD).
- Le taux de recombinaison est estimé à partir du nombre de blocs MLD en fonction du scénario choisi. En effet, le nombre de blocs MLD dépend de la détectabilité d'un évènement de recombinaison et du nombre d'évènements de recombinaison ayant lieu. La détectabilité d'un évènement de recombinaison dépend du ratio taux de recombinaison sur taux de mutation (ρ/μ) et du scénario démographique. Nous utilisons cette relation pour inférer ρ. Le nombre d'évènements de recombinaison est :$\overline{L_T} L_G \rho$, avec $\overline{L_T}$ la moyenne de la longueur totale d'un arbre de coalescence suivant un scénario démographique et L_G la longueur de la séquence. Pour faciliter notre estimation, nous étudions la relation entre le ratio $\frac{\rho}{\mu}$ et $\frac{\#MLD}{\#évènements}\frac{\rho}{\mu}$ (Fig 2.10). De cette manière :

$$\frac{\#MLD}{\#évènements}\frac{\rho}{\mu} = \frac{\#MLD}{\overline{L_T} L_G \rho}\frac{\rho}{\mu} = \frac{\#MLD}{\overline{L_T} L_G \mu},$$

  avec $\#MLD$ et L_G obtenus à partir des données. $\overline{L_T}$, dont dépend l'estimation de μ, varie avec le scénario choisi, est obtenu par simulation. Le ratio $\frac{\rho}{\mu}$ est donc obtenu à partir de cette relation. μ étant estimé précédemment, on obtient ρ à partir du ratio. La fraction $\frac{\#MLD}{\overline{L_T} L_G \mu}$, dont toutes les variables

sont mesurables ou estimables, est utilisée à la place de $\frac{\#MLD}{\#évènements}\frac{\rho}{\mu}$ pour l'estimation de ρ.

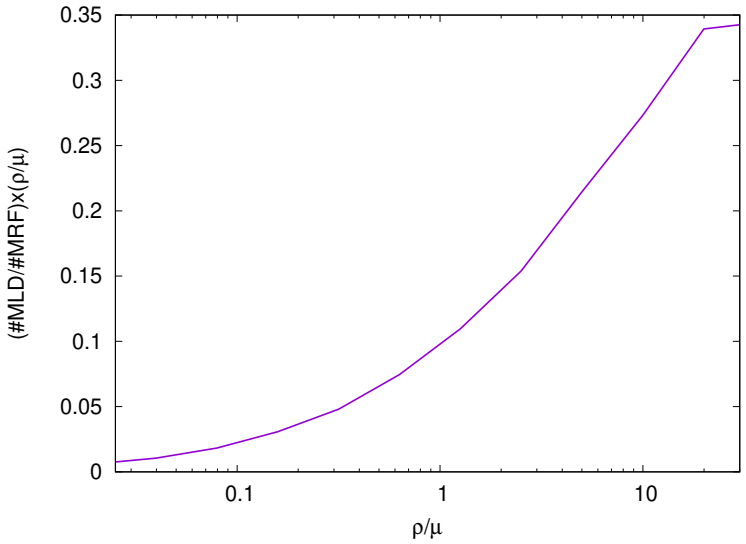


Figure 2.10: Produit de la détectatilité (nombre de blocs MRF sur nombre de blocs MLD) et du ratio entre taux de recombinaison et taux de mutation par rapport au ratio entre taux de recombinaison et taux de mutation.

Scénarios étudiés Nous considérons trois scénarios différents, tous composés de seulement deux paramètres à inférer. Nous nous intéressons à un scénario démographique (deux paramètres), un scénario de structure (deux paramètres) et un scénario combinant démographie (un paramètre) et structure (un paramètre). Le scénario combinant démographie et structure ne contient pas les scénarios de démographie seule et de structure seule. Les paramètres sont inférés en sélectionnant le maximum de vraisemblance parmi les couples de paramètres testés. Les scénarios sont définis de la manière suivante :

Démographie Le scénario démographique considéré est un changement brusque de taille de population à un temps τ, mesuré en unités de N générations dans le passé, d'une intensité κ. La population était κ fois plus grande dans le passé : $\kappa > 1$ est une décroissance, $\kappa < 1$ est une croissance.
L'amplitude des paramètres testés est : $\kappa \in [0.01 : 100], \tau \in [0.001 : 10]$, avec un pas de de 0.1 en échelle logarithmique.

Structure Le scénario de structure considéré est un modèle continent-île avec un continent κ fois plus grand que l'île. Les migrants vont du continent vers l'île à un taux m. Les individus sont échantillonnés dans l'île.
L'amplitude des paramètres testés est : $\kappa \in [0.01 : 100], m \in [0.01 : 100]$, avec un pas de de 0.1 en échelle logarithmique.

Démographie et structure Le scénario combinant démographie et structure considéré est un modèle continent-île avec un continent grandissant à un taux exponentiel g, avec un taux de migration m des migrants allant du continent vers l'île. Les individus sont échantillonnés dans l'île.
L'amplitude des paramètres testés est : $g \in [0.01 : 100], m \in [0.01 : 100]$, avec un pas de de 0.1 en échelle logarithmique.

Pseudo-vraisemblance Pour chaque scénario étudié (possédant ses propres paramètres), les taux de mutation et de recombinaison sont estimés. Des génomes sont ensuite simulés, à partir desquels une vraisemblance est mesurée.

Afin de sélectionner les paramètres du modèle étudié les plus ressemblants aux données, la pseudo-vraisemblance de la distribution de longueurs de blocs MLD est calculée de la même manière que dans la Section 2.3.

Pour chaque type de scénario étudié (démographie, structure...), le couple de paramètres expliquant le mieux les données est sélectionné par maximum de vraisemblance.

2.4.3 Résultats

2.4.3.1 Comparaison des distributions observées

Distributions brutes Comme dit précédemment, les distributions de blocs MLD des Yorubas et des gorilles devraient être de même ordre de grandeur, voire comparables si leurs démographies passées avaient le même profil historique. Cependant, les moyennes des blocs MLD sont très différentes entre ces deux populations, tout comme leur distributions (Fig 2.11a). Les deux distributions se superposent à peine. La distribution des gorilles prend seulement un cinquième des valeurs prises par la distribution des Yorubas.

Effet de la qualité de séquençage Pour voir si cette différence est due au fort écart dans la qualité de couverture de séquençage, nous avons appliqué la couverture de séquençage des gorilles sur l'alignement des Yorubas. Nous avons donc diminué volontairement la couverture de séquençage des génomes de Yorubas pour comparer deux distributions dans les mêmes états.

Les distributions de longueurs de blocs MLD des gorilles et des Yorubas « mal séquencés » prennent les mêmes gammes de valeurs (Fig 2.11b). Elles ne se superposent pas pour autant. La distribution de longueurs de blocs MLD des gorilles a un excès de segments courts et un déficit de segments longs par rapport à celle des Yorubas. Ce motif peut représenter deux populations de même taille au temps présent : une constante au cours du temps et une présentant un excès de courts segments ayant subi un déclin de population (voir Section Testing for population decline using maximal linkage disequilibrium blocks).

L'absence de couverture parfaite sur le génome a bien sûr un effet sur la distribution de blocs MLD. La distribution des Yorubas a drastiquement changé avec l'application de la couverture de séquencage du gorille. Cependant, la distribution

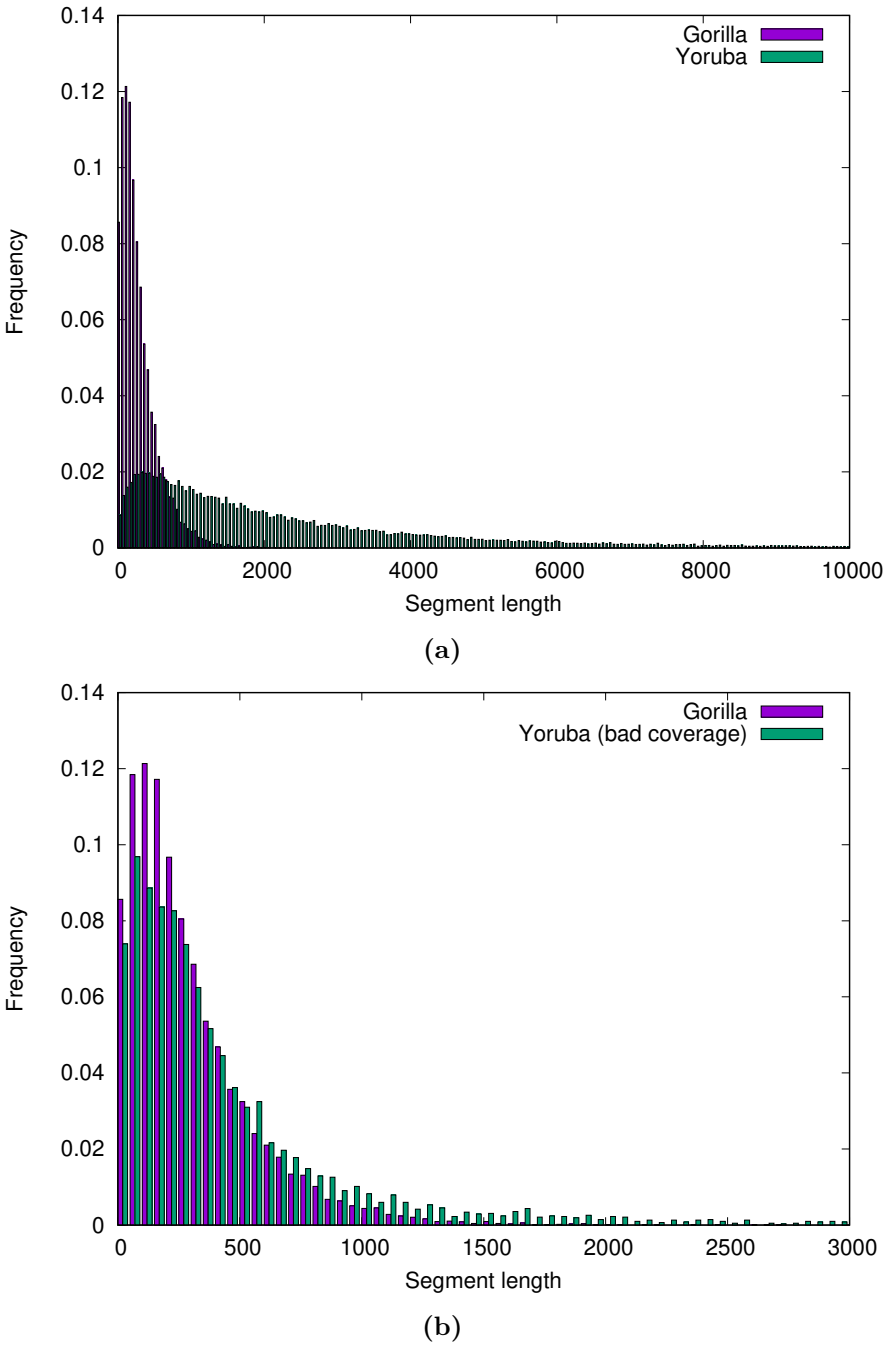


Figure 2.11: Distribution de la longueur des blocs MLD de chromome 1 des gorilles (mauve) et du chromosome 1 des Yorubas (vert) (a) et des Yorubas ayant la même couverture de séquençage que les gorilles (vert) (b).

des Yorubas « mal séquencés » étant toujours différente de celle des gorilles, la disparité de la couverture n'est pas responsable des différences entre les distributions des gorilles et des Yorubas. D'autres facteurs influencent donc ces distributions.

2.4.3.2 Comparaison à une population constante

Comme expliqué dans la partie Méthode, nous avons simulé des génomes possédant les mêmes caractéristiques que les chromosomes 1 des gorilles et des Yorubas, sous un modèle constant de population non structurée. Il est donc possible de comparer chaque distribution X_{obs} avec son équivalent sous H_0 : X_{H_0}. Afin de les comparer, nous divisons chaque distribution observée par son attendue : $\frac{X_{obs}}{X_{H_0}}$. Si la taille était constante alors $\frac{X_{obs}}{X_{H_0}}$ serait uniforme. Le ratio est présenté dans la figure 2.12.

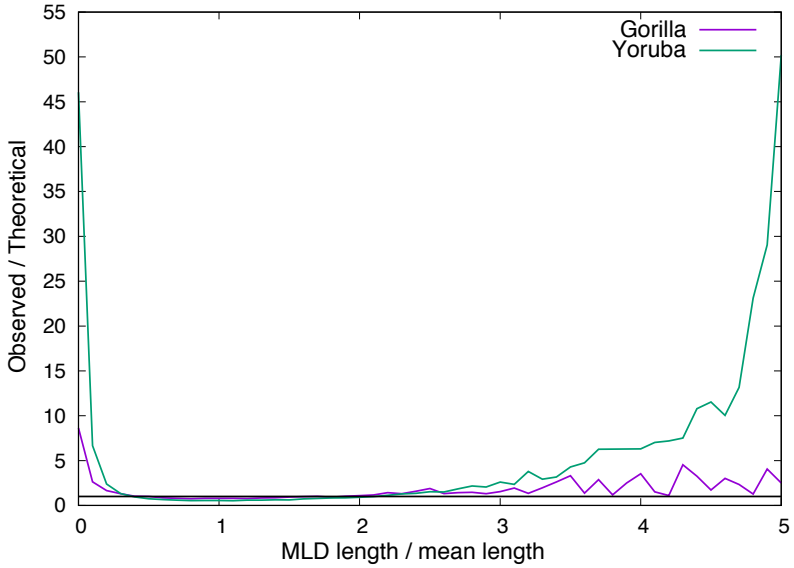


Figure 2.12: Ratio observé sur théorique des longueurs de blocs MLD normalisées pour la distribution des gorilles (mauve) et des Yorubas (vert).

La distribution de longueurs de blocs MLD des gorilles montre un excès de petits et grands blocs, comme attendu dans le cadre d'un déclin (Kerdoncuff et al. 2020). La distribution des Yorubas est, quant à elle, étonnante. Elle montre également un excès de petits et grands blocs, plus important que celle des gorilles. Les Yorubas montrent donc un écart plus important à H_0 (population constante), dans la même direction qu'une population en déclin. Ce phénomène étant très inattendu, nous concentrons maintenant notre étude sur la distribution des longueurs de blocs MLD de la population Yorubas.

2.4.3.3 Taux de recombinaison variable

Le taux de recombinaison étant variable le long du génome (Consortium 2003), cela peut influencer la taille des blocs MLD et la faire dévier de l'attendu, même

sous un modèle standard neutre. Nous nous sommes donc intéressés à la densité des évènements de recombinaison détectés le long du chromosome 1 des Yorubas. Cette dernière est quasi-uniforme (Fig 2.13). La densité des évènements de recombinaison est tout de même plus importante au niveau des télomères. Le centromère est facilement visible puisqu'il constitue une large zone non couverte par le séquencage où aucun évènement de recombinaison n'est détecté. Dans la partie du chromosome suivant le centromère (autour de la position $1.5 * 10^8$), il y a de nombreuses zones non couverts, au même endroit la densité des évènements de recombinaison est très faible (Fig 2.13).

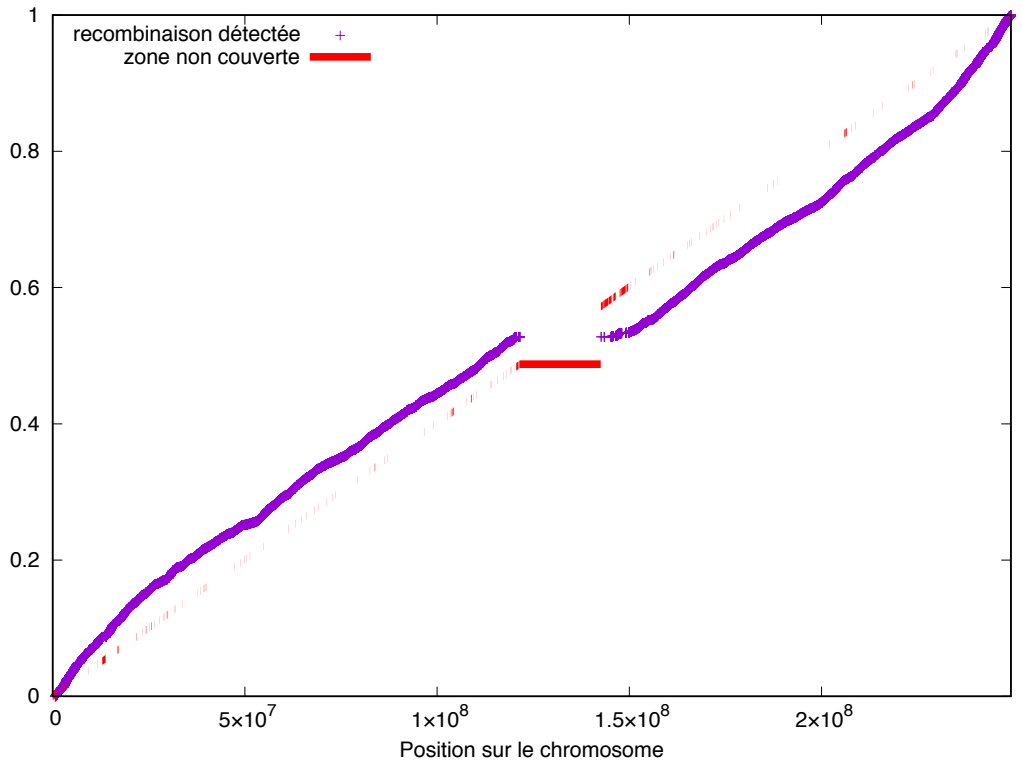


Figure 2.13: Densité des évènements de recombinaison détectés sur le chromosome 1 des Yorubas (mauve) et emplacement des zones non couvertes par le séquençage (rouge). En abscisse, l'emplacement sur le chromosome. En ordonnée, la fréquence cumulée des évènements de recombinaison (en mauve) qui se sont produits avant une certaine position sur le chromosome. Les zones non couvertes par le séquençage (en rouge) sont placées de manière linéaire pour aider à la visualisation.

Nous avons échantillonné les blocs MLD des parties non-interrompues dont la densité est la plus uniforme. La distribution des longueurs de blocs MLD sous-échantillonnés et similaire à la distribution des longueurs de blocs MLD de tout le chromosome. Ce n'est donc pas la variation du taux de recombinaison le long du génome qui explique la surdispersion des tailles de blocs.

2.4.3.4 Facteurs évolutifs expliquant la distribution de longueurs de blocs MLD des génomes de Yorubas

Scénarios testés Afin d'étudier les facteurs à l'origine de la déformation de la distribution de longueurs de blocs MLD des Yorubas, nous cherchons les écarts au modèle neutre pouvant expliquer la distribution observée. Comme décrit dans la partie Méthode, nous nous intéressons à trois scénarios : un scénario démographique, un scénario de structure et un scénario combinant démographie et structure.

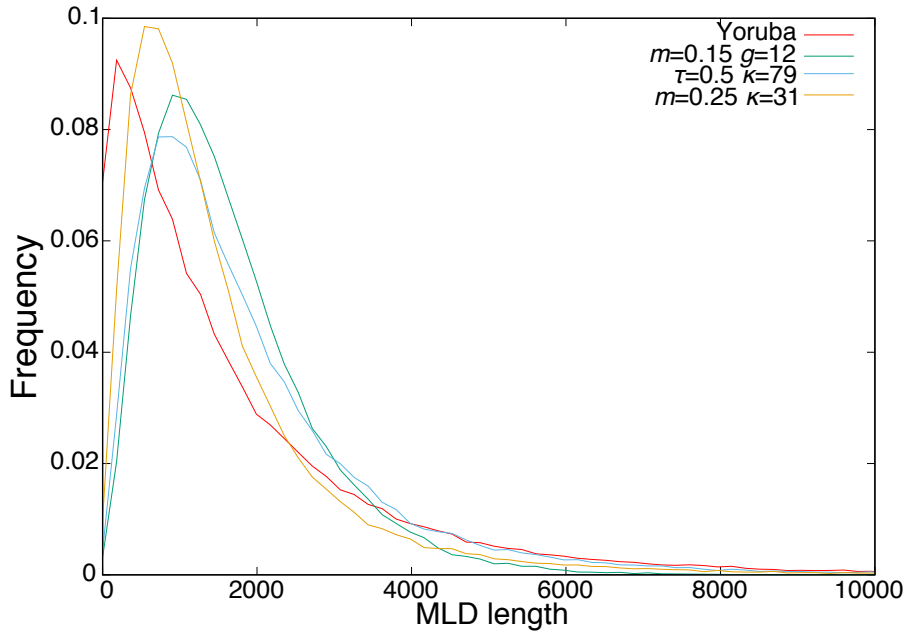


Figure 2.14: Distribution des longueurs de blocs MLD des Yorubas (rouge), de populations de même caractéristiques en croissance (bleu), structurée (orange) et une combinaison des deux (vert).

Résultats La liste des paramètres inférés suivant les différents scénarios est dans la Table 2.2.

Scénario	Paramètre 1	Paramètre 2	Log-Vraisemblance
Démographie (τ, κ)	$\tau = 0.5$	$\kappa = 79$	-242,638
Structure (m, κ)	$m = 0.25$	$\kappa = 31$	-239,468
Démographie (g) et structure (m)	$g = 12$	$m = 0.15$	-247,769

Table 2.2: Couples de paramètres dont la distribution de longueurs de blocs MLD a la meilleure vraisemblance suivant le scénario considéré : démographie, structure ou la combinaison des deux.

Démographie Le scénario démographique possédant la distribution de longueurs de blocs MLD la plus proche de celle des Yorubas est un scénario de décroissance important : la population serait 79 fois plus grande ($k = 79$) avant le déclin qui aurait eu lieu assez récemment ($\tau = 0.5$).

Structure Le scénario de structure avec la meilleure vraisemblance est un scénario où le continent est 31 fois plus grand que l'île ($k = 31$) et le taux de migration entre les deux assez élevé ($m = 0.25$). C'est la distribution qui se rapproche le plus de la distribution observée (Table 2.1).

Démographie et structure Le scénario de démographie et structure se rapprochant le plus de la distribution observée comporte une croissance exponentielle importante (à taux $g = 12$) avec un taux de migration un peu plus faible que celui inféré pour le scénario de structure seule ($m = 0.15$ vs $m = 0.25$ pour la structure seule).

Deux des trois scénarios correspondent à des histoires évolutives divergentes. Le scénario démographique indique une décroissance importante alors que le scénario combinant de la démographie et de la structure de population comporte une décroissance importante. Visuellement, aucun des trois scénarios testés ne produit une distribution en adéquation avec celle observée chez les Yorubas (Fig 2.14). Aucun scénario n'explique la distribution de longueurs de blocs MLD des Yorubas.

2.4.4 Conclusion

Les distributions de longueurs de blocs MLD mesurées sur des alignements de génomes rencontrent fréquemment la difficulté suivante : leurs qualités de génomes (couverture, génotypage) sont insuffisantes pour obtenir un bon échantillonnage de la distribution. La couverture de séquençage peut avoir un fort impact sur la distribution de longueurs de blocs MLD. Certaines longueurs de blocs ne sont pas mesurables, leur longueur dépassant la longueur des segments séquencés. Cependant, même avec une couverture de séquencage sous-optimale, la distribution de longueurs de blocs MLD contient toujours de l'information.

À l'aide de simulations, en prenant en compte les différents problèmes de séquencage des données, il est possible d'étudier et de comparer la probabilité de différents scénarios, aussi bien démographiques que des scénarios de structures de population. Malheureusement, nous ne sommes pas encore en mesure de reproduire des scénarios à l'origine de certaines distributions observées sur des données réelles. Il existe donc des scénarios plus complexes, ou faisant appel à des processus évolutifs non considérés, comme la sélection, qui expliqueraient les distributions de longueurs de blocs MLD.

Chapitre 3

Déséquilibre de liaison au sein et entre les segments non recombinés

Contents

3.1 Impact de la démographie sur les évènements de recombinaison

Ce travail a été réalisé dans le cadre du stage de Master 1 de Ludovic Fourteau intitulé *Caractérisation de l'effet d'un évènement de recombinaison sur la diversité observée au sein d'une population.*

3.1.1 Motivation et méthode

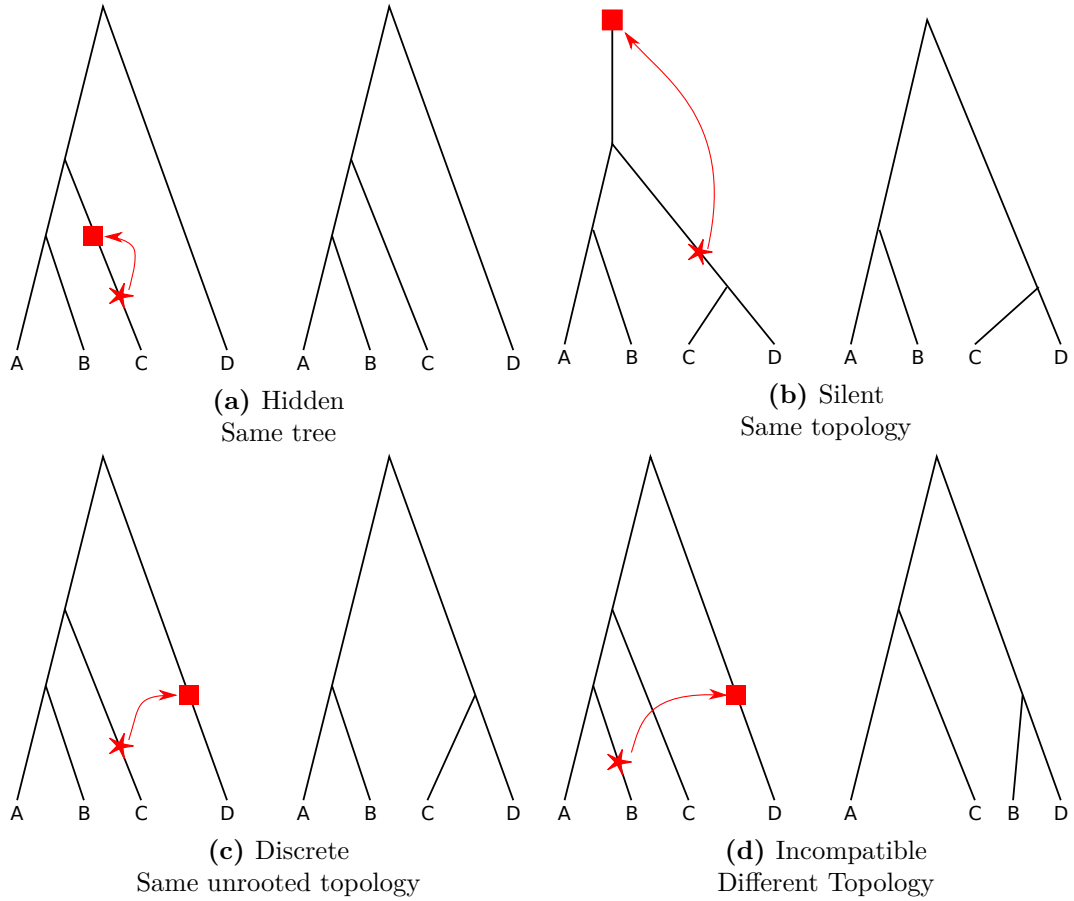
3.1.1.1 Motivation

Dans le chapitre précédent, nous avons vu que la distribution des longueurs de blocs MLD dépend du scénario démographique. La longueur d'un bloc MLD dépend du nombre et de la longueur des blocs MRF le composant. La longueur des blocs MRF dépend de l'arbre de coalescence sous-jacent à la population et donc, de son histoire évolutive. Le nombre de blocs MRF contenus dans un bloc MLD dépend de la détectabilité des évènements de recombinaison. Nous avons vu que la détectabilité des évènements dépendait, notamment, du ratio entre le taux de mutation et le taux de recombinaison. Dans cette partie, nous allons nous intéresser à l'effet de la démographie sur la détectabilité des évènements de recombinaison. Pour cela, nous avons décidé de caractériser plus précisément la relation entre évènement de recombinaison, topologie et longueurs de branches de l'arbre de coalescence.

3.1.1.2 Méthode

A l'aide d'un simulateur de coalescent avec recombinaison (Wiuf and Hein 1999) que j'ai codé en C, nous simulons, selon le scénario choisi, un arbre de coalescence, puis *un* seul évènement de recombinaison arrivant aléatoirement sur l'arbre. L'évènement de recombinaison arrivant sur l'arbre (représenté par une croix rouge dans Fig 3.1) génère une cassure de la lignée à l'emplacement de l'évènement de recombinaison. Nous appellerons cette lignée *lignée recombinante*. Cette lignée coalesce à un temps antérieur à l'évènement de recombinaison dans l'arbre (évènement de coalescence représenté par un carré rouge dans Fig 3.1).

3.1.2 Définition des différents types de recombinaison



(a) Hidden
Same tree

(b) Silent
Same topology

(c) Discrete
Same unrooted topology

(d) Incompatible
Different Topology

Figure 3.1: Les 4 types de recombinaison. Cachée (a), Silencieuse (b), Discrète (c) et Incompatible (d).

En fonction de l'effet d'un évènement de recombinaison sur l'arbre de coalescence, il est possible de définir 4 types de recombinaison (définitions inspirées de Ferretti et al. 2013 et concordantes avec Deng et al. 2021) :

Hidden : Recombinaison invisible. Elle n'a aucun effet sur l'arbre de coalescence : ni sur la topologie ni sur les longueurs de branches. L'évènement de recombinaison et l'évènement de coalescence se situent entre des individus de la même lignée, de la même branche de l'arbre (Fig 3.1a).

Silent : Recombinaison silencieuse. Il n'y a pas de changement au niveau de la topologie de l'arbre. Cependant les longueurs de branches varient ainsi que le temps d'un noeud de l'arbre (Fig 3.1b). Ce type de recombinaison se produit quand la coalescence a lieu entre la lignée recombinante et sa lignée ancêtre.

Discret : Recombinaison discrète. Il y a un changement de topologie (et de longueurs de branches). Pourtant, la topologie non racinée de l'arbre ne change pas. Ce type de recombinaison n'est pas détectable utilisant le test des 4-gamètes (Hudson and Kaplan 1985), mais pourrait l'être en connaissant les allèles

ancestrales. Ce type de recombinaison a lieu quand la lignée recombinante coalesce avec une lignée possédant la même lignée ancêtre (Fig 3.1c).

Incompatible : Recombinaison incompatible. Il y a un changement de topologie (et de longueurs de branches), détectable avec le test des 4-gamètes. Ce type de recombinaison se produit quand la lignée recombinante coalesce avec une lignée ne possédant pas la même lignée ancêtre (et n'étant pas la lignée ancêtre de cette dernière) (Fig 3.1d).

3.1.3 Fréquence des différents types de recombinaison

La fréquence de ces différents types de recombinaison dépend de plusieurs paramètres : le nombre de feuilles de l'arbre ainsi que la topologie et les longueurs de branches de l'arbre. La topologie et les longueurs de branches sont fortement affectées par la démographie.

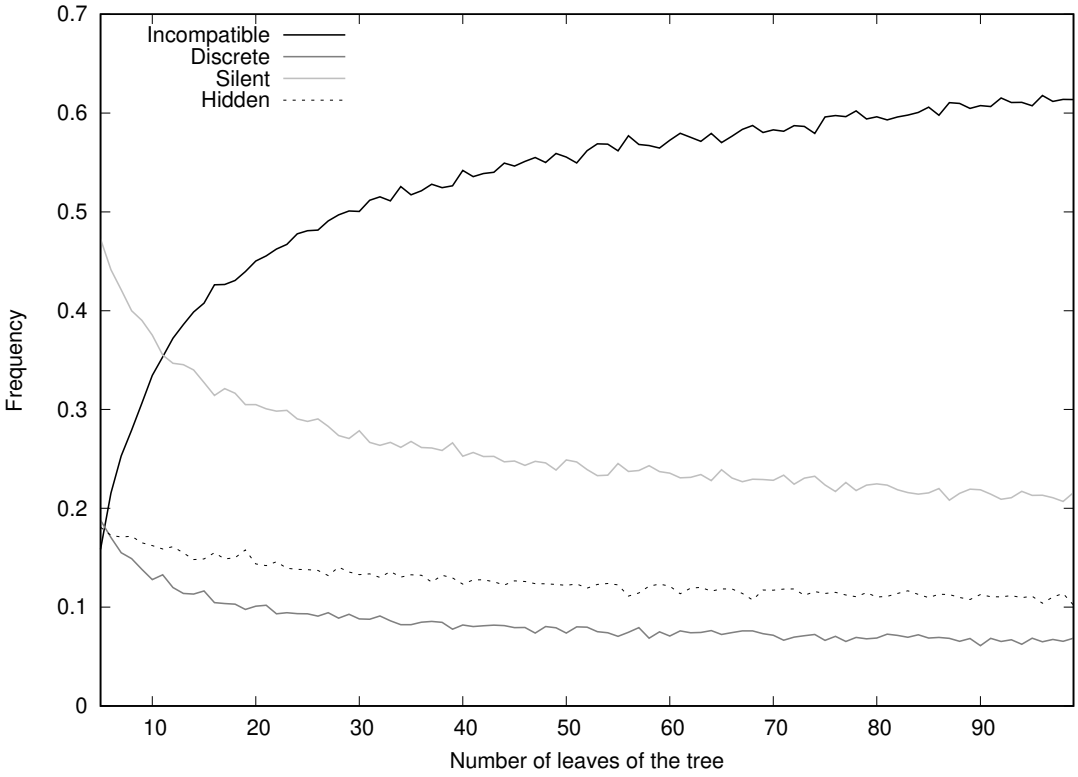


Figure 3.2: Proportion des différents types de recombinaison : recombinaison incompatible (noir), discrète (gris foncé), silencieuse (gris clair) ou cachée (pointillé) en fonction du nombre de feuilles de l'arbre $n \in [5 : 99]$.

3.1.3.1 Nombre de feuilles de l'arbre

En ce qui concerne le nombre de feuilles, si l'arbre en contient seulement 3, il ne peut y avoir de recombinaison incompatible. De même, s'il en contient 4,

peu de configuration vont créer de l'incompatibilité (comme montré dans Fig 3.1). Les proportions d'évènements de recombinaison incompatible vont donc augmenter avec le nombre de feuilles de l'arbre et celles de recombinaison silencieuse vont diminuer (Fig 3.2).

3.1.3.2 Topologie et longueurs de branches de l'arbre - Démographie

Afin d'étudier l'impact de la topologie et des longueurs de branches de l'arbre, nous avons décidé de comparer différents scénarios démographiques : une population constante, un déclin soudain et une croissance soudaine.

Scénario Nous considérons un changement brutal de la taille de la population (N_0 au temps présent) à un temps τ (exprimé en N_0 générations), d'une intensité κ. Au temps présent la population est de taille N_0, elle était de taille κN_0 avant le changement brutal. Si $\kappa = 1$ la population est constante, si $\kappa > 1$, la population était plus grande avant le changement de taille, elle est donc en décroissance. Si $\kappa < 1$, la population était plus petite avant le changement, elle est donc en croissance.

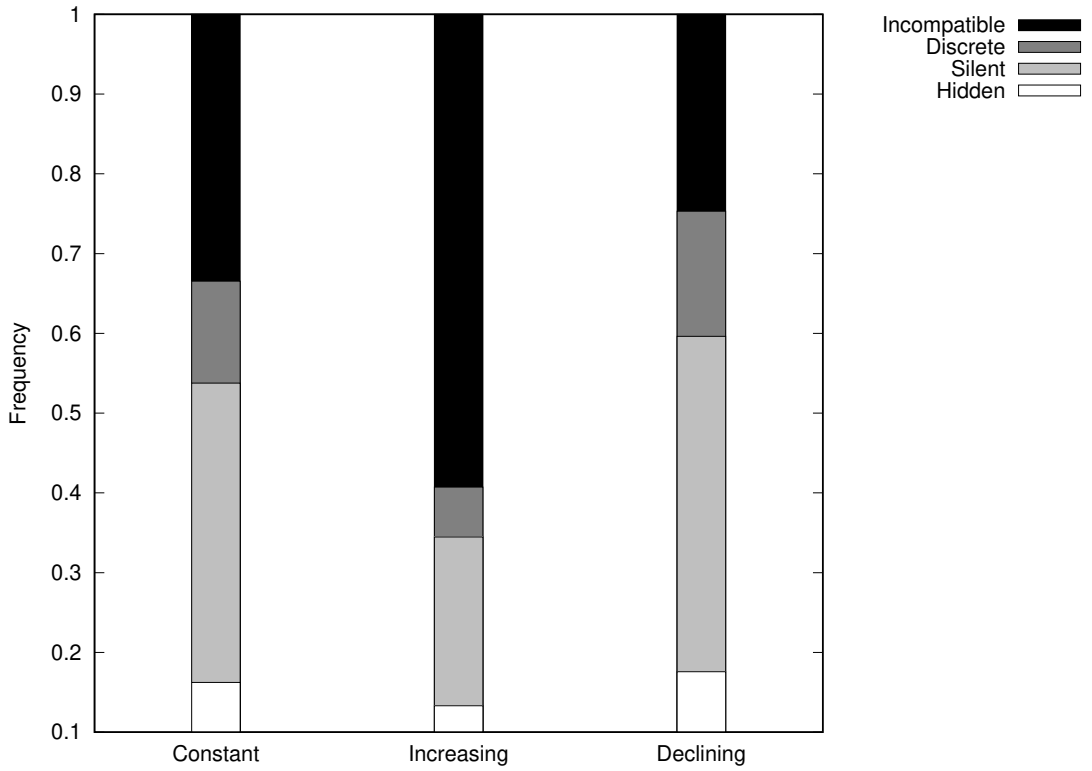


Figure 3.3: Fréquence des 4 types de recombinaison : recombinaison incompatible (noir), discrète (gris foncé), silencieuse (gris clair) ou cachée (blanc), en fonction du scénario démographique : constant (droite), croissance (milieu) ou déclin (gauche). Pour $n = 10$, $\tau = 1$, dans le cas de la croissance $\kappa = 0.1$ et dans le cas de la décroissance $\kappa = 10$.

Observations La démographie d'une population a un impact sur l'arbre de coalescence de cette population : sur sa topologie et ses longueurs de branches. L'arbre d'une population en déclin a des branches ancestrales plus grandes que celui d'une population constante et encore plus que celui d'une population en croissance. Comparativement, il y a plus d'évènements de recombinaison se produisant sur ces branches, qui entraînent des recombinaisons silencieuses (Fig 3.3). A l'opposé, l'arbre d'une population en croissance a des branches terminales (près des feuilles) plus grandes que celui d'une population constante ou en déclin. Il y a donc plus d'évènements de recombinaison se produisant sur ces branches, qui entraînent des recombinaisons incompatibles (Fig 3.3).

3.1.4 Discussion

3.1.4.1 Estimation du taux de recombinaison

Nous avons montré que les types de recombinaison se produisant dépendent du nombre de feuilles de l'arbre et de l'histoire évolutive de la population. Les marques : incompatibilité, déséquilibre de liaison..., laissées par les évènements de recombinaison sur les génomes servent à en estimer le taux de recombinaison. Certains types de recombinaisons, comme les recombinaisons invisibles, ne laissent pas de marque. Cependant, les recombinaisons invisibles sont très peu souvent considérées afin d'estimer le taux de recombinaison et faussent donc l'estimation du taux. Il peut être possible de les prendre en compte comme le propose la correction de Deng et al. 2021 qui étudient, entre autres, les temps d'attente entre deux changements de topologie et/ou de longueurs de branches le long d'alignement multiple. Cependant, les proportions de recombinaison invisible et silencieuse, et donc la distribution des temps d'attente entre deux changements de topologie varie en fonction de la déformation de la topologie et des longueurs de branches d'un arbre. Nous avons seulement considéré ici le cas de changement démographique, mais d'autres phénomènes comme la structure de population, peuvent faire varier les topologies et les longueurs de branches des arbres de la même manière. Il peut être important de prendre en compte l'histoire évolutive d'une population afin d'estimer son taux de recombinaison.

3.1.4.2 Utilisation du taux de recombinaison pour inférer l'histoire de la population

Il est souvent demandé de connaître le taux de recombinaison pour calculer certaines mesures sur le génome et pouvoir inférer l'histoire de la population (notamment pour les logiciels utilisant l'information de l'IICR comme PSMC (Li and Durbin 2011)). Pour estimer un taux de recombinaison ne dépendant pas de la démographie de la population, il est nécessaire de l'estimer sur une population la plus proche possible du modèle standard neutre. L'étude de la démographie d'une population utilisant l'information du taux de recombinaison populationnel nécessite l'étude préalable d'une population de la même espèce proche. Bien qu'un taux de

recombinaison populationnel ne soit pas exact s'il est estimé à partir d'une population à l'histoire évolutive complexe, nous ne connaissons pas l'effet que cela peut avoir sur les inférences démographiques l'utilisant. (Il est toujours possible d'estimer le taux de recombinaison à partir de données de pedigrees, qui ne semblent pas souffrir des mêmes biais.)

3.2 Étude de la distribution du déséquilibre de liaison

3.2.1 Motivation

Dans la section précédente, nous venons de montrer, par simulation, l'existence d'interactions entre la démographie et la recombinaison. Nous souhaitons maintenant aller plus loin et essayer d'utiliser cette interaction afin d'étudier la démographie d'une population. Une statistique souvent utilisée pour étudier l'histoire évolutive d'une population à l'aide des évènements de recombinaison est le déséquilibre de liaison (LD) (Conrad et al. 2006; Patin et al. 2014).

Le LD est à la fois affecté par la démographie et par la recombinaison (McVean 2008), nous avons donc étudié la possibilité de séparer ces deux effets. Pour cela, nous nous intéressons à la distribution du déséquilibre de liaison en l'absence de recombinaison, c'est-à-dire au sein d'un même bloc MRF. Puis entre des sites séparés d'un nombre de recombinaison connu x, donc situés à x blocs MRF d'écart. Les blocs MRF n'étant pas identifiables sur un alignement de séquence, nous menons la même étude sur le déséquilibre de liaison au sein et entre les blocs MLD.

3.2.2 Déséquilibre de liaison et blocs MRF

Nous étudions le déséquilibre de liaison à l'aide de la covariance D. Nous considérons ici la covariance D entre deux sites, chacun composé de deux allèles : A/a et B/b. Nous considérons A et B comme étant les allèles dérivés. Les variables f_A et f_B sont respectivement les fréquences de A et de B et f_{AB} est la fréquence de la co-occurrence AB. La covariance D est donc définie par : $D_{AB} = f_{AB} - f_A f_B$ (Lewontin and ichi Kojima 1960).

Nous considérons d'abord la distribution de cette valeur en fonction de la topologie et des longueurs de branches de l'arbre. C'est-à-dire en mesurant D entre des mutations présentes sur un même arbre, donc sur un même MRF. Puis nous nous intéressons à la distribution de D, entre sites séparés par un ou plusieurs évènements de recombinaison. Pour cela nous étudions la distribution de D calculée sur des paires de mutations séparées par un certain nombre x d'évènements de recombinaison, donc se trouvant à x blocs MRF d'écart. On note D_x^{MRF} la distribution de D entre des sites situés à x blocs MRF d'écart. La distribution de D calculée entre des sites d'un même bloc se note donc D_0^{MRF}. La covariance D calculée entre deux sites se trouvant dans des blocs MRF adjacents se nomme D_1^{MRF} (Fig 3.4).

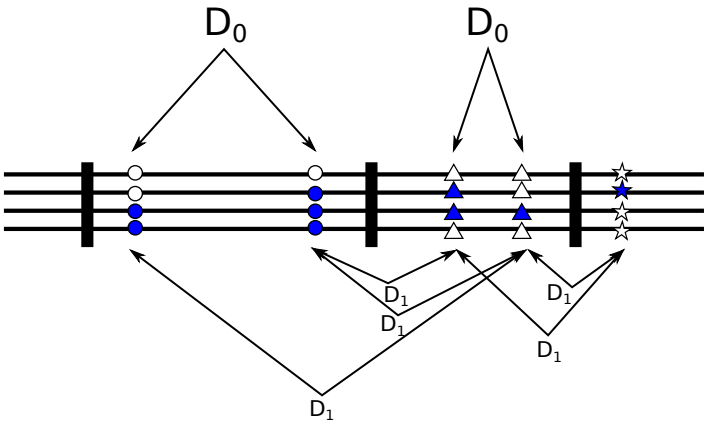


Figure 3.4: Illustration du calcul des distributions D_0^{MRF} et D_1^{MRF}. Les lignes horizontales représentent un alignement de génome, les lignes verticales l'emplacement d'évènement de recombinaison. Les ronds, triangles et étoiles représentent des sites polymorphes comprenant deux allèles : l'allèle ancestral (en blanc) et l'allèle dérivé (en bleu). À partir de cet alignement il est possible de calculer deux valeurs de D_0^{MRF} et 5 valeurs de D_1^{MRF}.

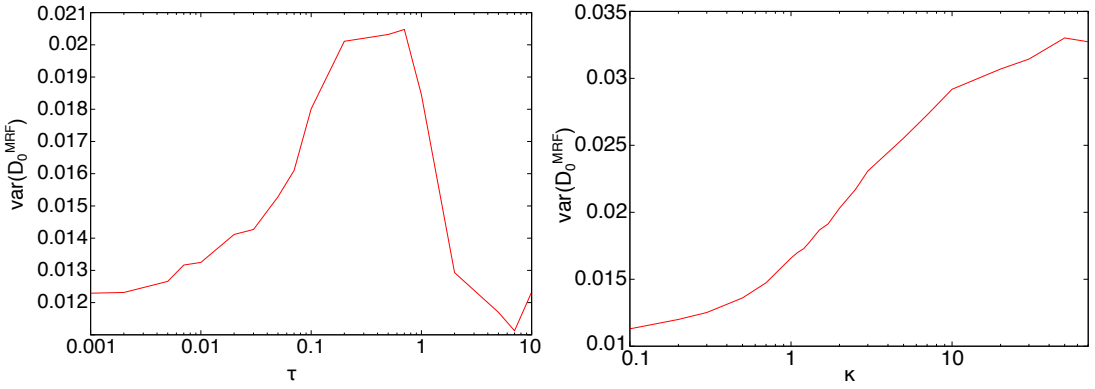
3.2.2.1 Espérance

Dans ce cadre Amaury Lambert s'est intéressé à la moyenne ainsi que la variance de D considérant le cas d'un modèle à une infinité de site ainsi qu'un arbre ultramétrique fixé. Il a ainsi démontré que l'espérance de D était nulle, la variance quant à elle, dépend de la topologie et des longueurs de branches. La démonstration mathématique est présentée dans la note en Annexe 6.1.

3.2.2.2 Variance

Par la suite, nous nous sommes intéressés à la variance de D et à l'impact de la démographie et des évènements de recombinaison sur sa valeur. Cette étude est faite à l'aide de simulations effectuées avec le logiciel msprime (Kelleher et al. 2016). Les scénarios démographiques considérés sont identiques à ceux de la section précédente. Il s'agit de changements brutaux de taille de population. La taille de la population au temps présent N_0 a subi un changement à un temps τ (exprimé en N_0 générations), d'une intensité κ. La population était de taille κN_0 avant le changement de taille.

En fonction de la démographie La variance de D varie selon la topologie et les longueurs de branches de l'arbre et donc selon la démographie (McVean 2008). En effet, D dépend des fréquences alléliques et ces dernières varient avec la topologie et les longueurs de branches, la covariance D est donc affectée. En ce qui concerne la topologie et les longueurs de branches d'un arbre en croissance, les branches terminales sont comparativement plus longues que celles d'une population constante, on observe plus de mutations à faible fréquence. Comme D est mesurée à l'aide du produit de ces fréquences, D a donc des valeurs absolues plus faibles et une variance plus faible. A l'inverse, un arbre en décroissance possède des branches ancestrales



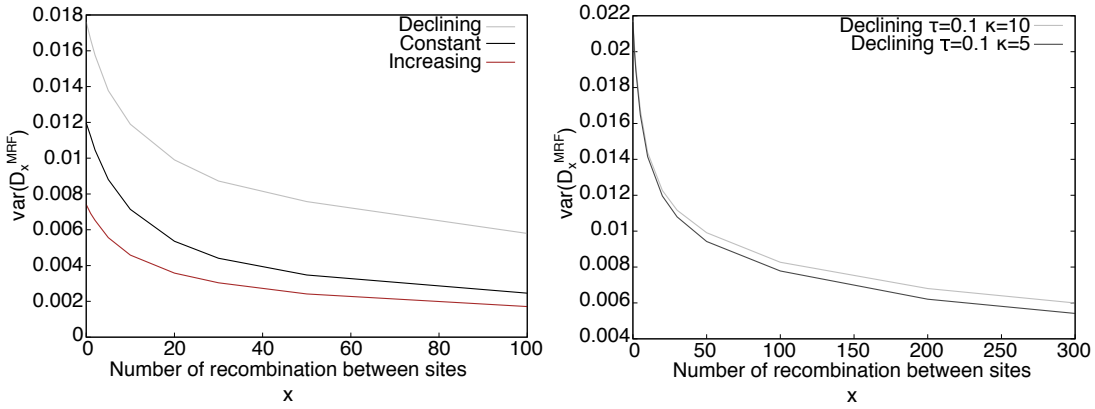
(a) τ variable **(b)** κ variable

Figure 3.5: Variance de D_0^{MRF} dans le cadre d'un changement brutal de taille de la population : considérant (a) une intensité de changement fixée $\kappa = 5$ et une date de changement variable τ ainsi que (b) une date de changement fixée $\tau = 0.5$ et une intensité de changement variable, pour $n = 100$.

plus longues en comparaison avec une population constante. Ces dernières reçoivent donc plus de mutations, la fréquence des mutations étant plus élevée, la valeur absolue de D est plus importante et la variance de la distribution de D est plus élevée. La variance de D dépend donc des caractéristiques décrivant la démographie de la population (exemple pour D_0^{MRF} Fig 3.5). La nature d'un changement de taille de population ainsi que l'intensité de ce dernier aura donc un impact sur la variance de D (exemple pour D_0^{MRF} Fig 3.5b). Plus un changement est important plus la différence sera observable. La temporalité d'un changement de taille de population ayant aussi un effet important sur la topologie et les longueurs de branches d'un arbre de coalescence, elle a aussi un effet sur la variance de D (exemple pour D_0^{MRF} Fig 3.5a).

En fonction du nombre d'évènement de recombinaison La liaison physique entre deux allèles peut être interrompue par des évènements de recombinaison, ce qui a pour effet d'homogénéiser D, et donc sa distribution. La variance de la distribution va donc diminuer avec le nombre d'évènements de recombinaison x séparant les deux mutations (Fig 3.6a).

La cassure de la liaison va dépendre du type d'évènement de recombinaison, qui comme on l'a vu dans la section précédente, va dépendre également de la démographie. En effet, un évènement de recombinaison silencieuse ou invisible n'affectera pas la liaison entre deux allèles. Le nombre d'évènements de recombinaison nécessaire pour que la distribution s'homogénéise, que la variance diminue jusqu'à atteindre la valeur de mutations non liées, va donc dépendre de la démographie. Une population en croissance, pour laquelle la fréquence des évènements de recombinaison changeant complètement la topologie et les longueurs de branches d'un arbre est plus importante, atteint, pour un nombre plus faible d'évènements de recombinaison, la variance correspondant à celle des mutations indépendantes. A



(a) Trois scénarios différents. **(b)** Deux scénarios de décroissance.

Figure 3.6: Variance de D_x^{MRF} pour des scénarios différents : (a) une population constante (noir), une population en déclin ($\tau = 0.1$,$\kappa = 5$) (gris) et une population en croissance ($\tau = 0.5$,$\kappa = 0.2$) (bordeaux), pour $n = 100$ et (b) deux scénarios de décroissance : division de la taille par 5 $\kappa = 5$ (noir) et par 10 ($\kappa = 10$) (gris) à un temps identique ($\tau = 0.1$), pour $n = 100$.

l'inverse, une population en décroissance a besoin de plus d'évènements de recombinaison pour atteindre ce plateau (Fig 3.6a).

L'intensité du changement a un effet sur la topologie et les longueurs de branches de l'arbre et la fréquence des différents types d'évènements de recombinaison, elle a donc aussi un effet sur la variance de D_x^{MRF}. Ainsi, pour deux décroissances arrivant à la même date, la plus forte, celle avec les plus longues branches terminales et donc le plus de recombinaisons silencieuses, a une variance qui décroît moins par évènement de recombinaison (Fig 3.6b).

3.2.3 Déséquilibre de liaison et blocs MLD

Les blocs MRF n'étant pas identifiables sur un alignement de séquence, nous avons décidé d'utiliser les blocs MLD, définis dans le chapitre précédent, et possédant des caractéristiques similaires. Nous nous sommes donc intéressés à la variance de D au sein et entre les blocs MLD que nous appelons D^{MLD}. Nous appelons D_x^{MLD} la covariance D entre sites à x blocs MLD d'écarts. De la même manière que pour les blocs MRF, D_0^{MLD} correspond à la covariance D entre des sites au sein du même MLD et D_1^{MLD} pour des sites dans des blocs MLD adjacents.

Avant propos Une des différences majeures entre les blocs MRF et les blocs MLD est l'importance du ratio entre le taux de mutation et le taux de recombinaison (μ/ρ). Ce dernier a un effet sur la détectabilité d'un évènement de recombinaison et donc sur le nombre de blocs MRF par bloc MLD. Cela va avoir un impact sur la variance à l'intérieur des blocs MLD.

Plus le ratio μ/ρ augmente, plus il y a d'évènements de mutation par évènement de recombinaison et plus ces derniers seront détectés. Donc plus le ratio est grand,

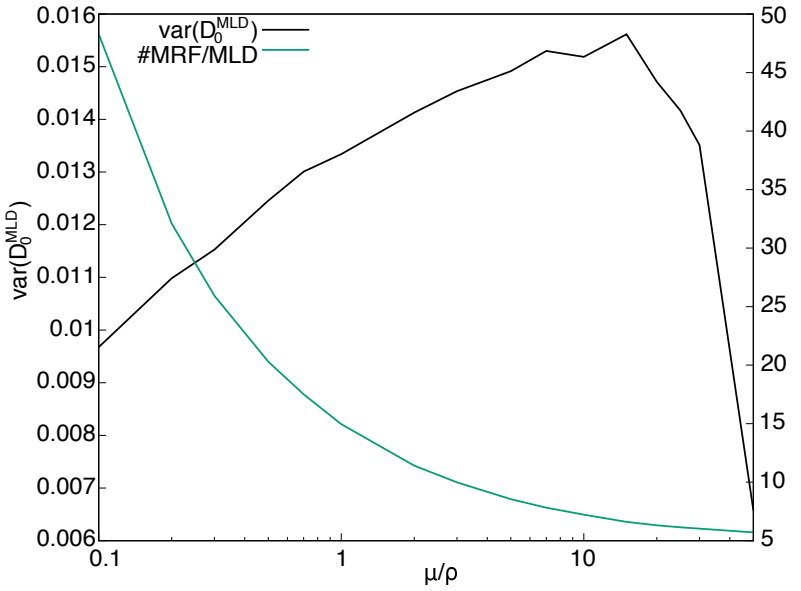


Figure 3.7: Variance de D_0^{MLD} au sein des blocs MLD (noir) et variation du nombre de bloc MRF par MLD (vert) en fonction du ratio entre le taux de mutation et le taux de recombinaison, pour $n = 100$.

moins il y a de blocs MRF par bloc MLD. Lorsque le nombre de blocs MRF par bloc MLD est élevé, il y a de multiples arbres au sein d'un même bloc MLD, dont certains sont incompatibles entre eux. La liaison complète entre les sites peut donc être brisée.

Lorsque le nombre de blocs MRF par bloc MLD diminue, cela affine le découpage des arbres. Il y a de moins en moins de topologies différentes au sein d'un même bloc MLD, ce qui a comme effet d'augmenter la variance. La variance de D^{MLD} va donc augmenter avec le ratio μ/ρ, jusqu'à un certain point (exemple pour D_0^{MLD} Fig 3.7). En effet, quand le ratio est très élevé, le nombre de mutations par arbre est très important. L'échantillonnage de la distribution de D^{MLD} est de plus en plus précis, ce qui a pour effet de réduire la variance, la distribution échantillonnée convergeant vers la vraie distribution.

Pour la suite de l'étude nous avons choisi d'utiliser le ratio $\mu/\rho = 10$.

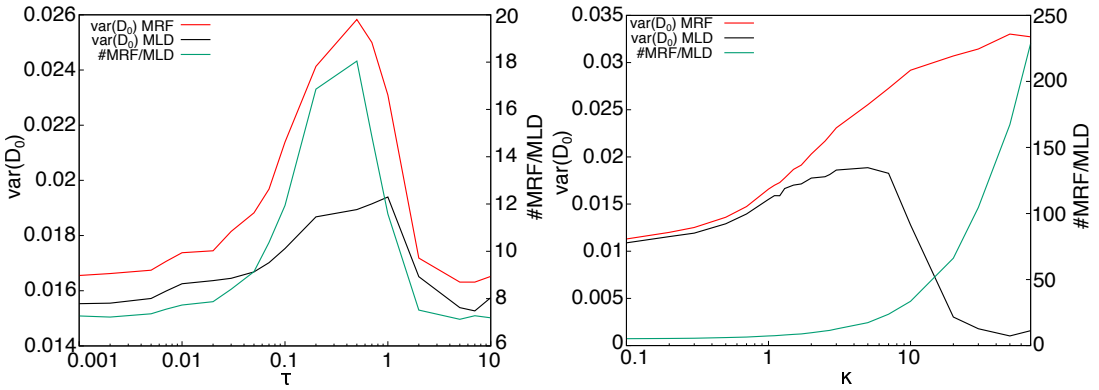
3.2.3.1 Au sein d'un bloc

La variance de D^{MLD} au sein des blocs MLD (D_0^{MLD}) dépend à la fois de la variance de D_0^{MRF} (au sein des blocs MRF) et du nombre de blocs MRF contenu dans chaque bloc MLD. Le nombre de blocs MRF par bloc MLD dépend de la détectabilité des évènements de recombinaison, qui dépendent eux-mêmes du ratio μ/ρ comme vu précédemment, mais aussi du type d'évènement de recombinaison, qui dépend de la démographie 3.8).

Dans le cadre d'un changement brutal de taille de population, la variance de D_0^{MLD} augmente avec l'intensité du déclin (plus le déclin est fort plus la variance est élevée), jusqu'à un point de saturation (Fig 3.8b). En effet, pour des intensités

fortes ($\kappa > 10$), les arbres sont très déformés. Il y a beaucoup de recombinaisons silencieuses, le nombre de blocs MRF par bloc MLD augmente ce qui a pour effet de diminuer la variance de D_0^{MLD}.

L'effet de la date du changement de taille de population sur la variance de D_0^{MLD} est très similaire à celui sur la variance de D_0^{MRF}. La variation est cependant moins forte car atténuée par le nombre de bloc MRF contenu dans chaque bloc MLD, ce nombre augmentant, la variance de D_0^{MLD} diminue (Fig 3.8a).



(a) τ variable **(b)** κ variable

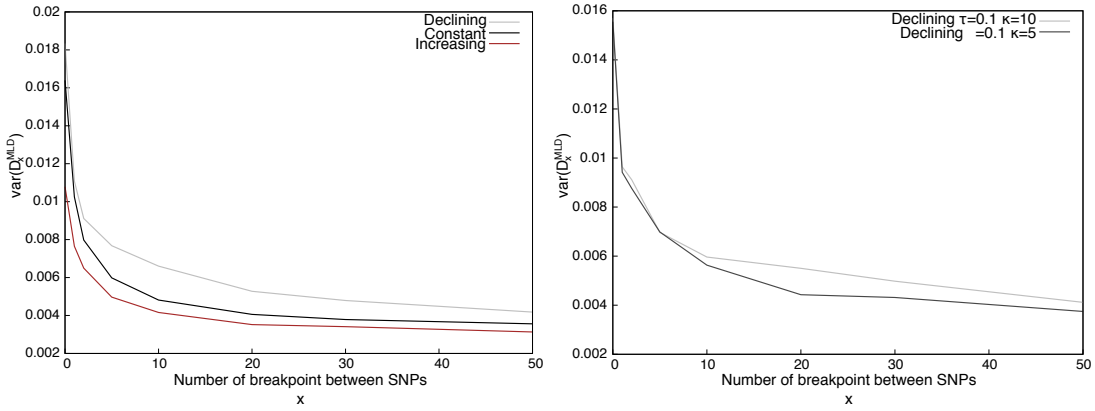
Figure 3.8: Variance de D_0^{MRF} par bloc MRF (rouge) et D_0^{MLD} par bloc MLD (noir) et variation du nombre de blocs MRF par bloc MLD (vert), dans le cadre d'un changement brutal de taille de la population : considérant (a) une intensité de changement fixée $\kappa = 5$ et une date de changement variable τ ainsi que (b) une date de changement fixée $\tau = 0.5$ et une intensité de changement variable, pour $n = 100$.

3.2.3.2 Entre les blocs

Les blocs MLD étant définis à l'aide des topologies incompatibles provoquées par des évènements de recombinaison cassant la liaison entre les sites, la différence de variance entre des sites compris dans le même MLD ou entre deux MLD adjacents est assez importante. La variance diminue beaucoup entre la distribution de D_0^{MLD} et de D_1^{MLD}. Les blocs MRF étant séparés par tous les types de recombinaison même celles n'ayant aucun effet sur les liaisons entre sites, la diminution de la variance en fonction du nombre de recombinaisons incompatibles séparant deux sites (ou le nombre de blocs MLD les séparant) est plus importante que dans le cas des blocs MRF (Fig 3.9a). Cependant on retrouve la même dynamique, une diminution de la variance dépendant de la nature du changement de taille de population (Fig 3.9a) et de l'intensité de ce changement (Fig 3.9b).

3.2.3.3 Conclusion

La variance de D est plus élevée dans le cas d'une population décroissante et plus basse dans le cas d'une population en croissance. La liaison entre les sites diminue en fonction du nombre d'évènements de recombinaison les séparant, mais



(a) Trois scénarios différents. **(b)** Deux scénarios de décroissance.

Figure 3.9: Variance de D_x^{MLD} en fonction du nombre d'évènements de recombinaison détectable séparant les allèles pour des scénarios différents : (a) une population constante (noir), une population en déclin ($\tau = 0.1$, $\kappa = 5$) (gris) et une population en croissance ($\tau = 0.5$, $\kappa = 0.2$) (bordeaux), pour $n = 100$ et (b) deux scénarios de décroissance : division de la taille par 5 ($\kappa = 5$) (noir) et par 10 ($\kappa = 10$) (gris) à un temps identique ($\tau = 0.1$), pour $n = 100$.

la force de cette diminution dépend de la topologie et des longueurs de branches de l'arbre (dans notre cas : du scénario démographique dont la population dépend).

Il est possible d'étudier ces différents phénomènes en considérant la variance de D au sein et entre les blocs MRF. Mais également en considérant la variance de D^{MLD} au sein et entre les blocs MLD. La distribution de D^{MLD} est mesurable sur un alignement multiple, à partir du moment où la qualité des données est suffisante pour découper des blocs MLD.

3.3 Inférences utilisant D_0 et D_1

Dans ce chapitre, nous définissons deux nouvelles statistiques utiles pour étudier l'histoire évolutive d'une population : D_0 la distribution de D au sein d'un bloc MLD et D_1 la distribution de D calculé entre des sites se situant dans des blocs MLD adjacents, correspondant respectivement à D_0^{MLD} et D_1^{MLD} dans la section précédente. Nous étudions leur sensibilités et les comparons à celle du SFS, à la fois dans le cadre d'un scénario simple de changement brutal de taille de population, et dans le cadre de l'étude de trois jeux de données portant sur une population d'humains de l'ethnie Yoruba, une population de gorilles et une population de bonobos.

Ce chapitre a été écrit en anglais pour faciliter le transfert vers une publication.

3.3.1 Introduction

Intraspecific patterns of genetic diversity are the result of evolutionary processes, e.g., selection, structure, demography. Conversely, one can use present-day diversity to infer evolutionary history and past demographic variations.

Classically, demographic inference methods have relied on the frequencies of genetic polymorphisms, that are often summarized by their distribution genome-wide, a distribution known as the Site Frequency Spectrum (SFS) (Fu 1995). The SFS can be abstracted in a vector noted $\boldsymbol{\xi}$ which component ξ_i is the number of polymorphic sites where the derived variant is at frequency i/n in a sample of size n. This variant can be a SNV or an indel. The genome SFS is thus a summary statistic of the genetic diversity segregating in multiple individuals, averaged over the whole genome. Because the SFS is strongly distorted by the demographic history of the species (Adams and Hudson 2004; Marth et al. 2004), it is used to infer parameters of demographic models, either by likelihood maximisation (Gutenkunst et al. 2009; Excoffier et al. 2013) or by distance minimization (Beaumont et al. 2002; Lapierre et al. 2017). Methods based on the SFS can suffer from non-identifiability issues as they cannot always discriminate between similar demographic scenarios (Lapierre et al. 2017; Rosen et al. 2018).

SFS disregards any source of correlations between the frequencies of the different polymorphic sites. Within a single chromosome, all loci are however physically linked on the same DNA molecule, which breaks the assumption of independence between the variants located at different sites on the same chromosome. If there were no recombination, the fate of an allele at a given locus would be entirely linked to all other variants in the same chromosome. However, homologous recombination mixes the paternally inherited and the maternally inherited chromosomes during meiosis, producing chimeric chromosomes in the gametes. Recombination therefore results in a decoupling of the loci located on both sides of the recombination breakpoint. Statistically, the more recombination occurs between two loci throughout the generations, the more independent they are. One common statistical measure of the dependence between the frequencies of two variants located at two sites is *linkage disequilibrium* (LD).

Mathematically, LD has been quantified by the statistic $D = f_{AB} - f_A f_B$, where f_A is the frequency of the A allele at the first locus, f_B the frequency of the B allele at the second locus and f_{AB} the frequency of the joint haplotype with both alleles A and B at both loci (Lewontin and ichi Kojima, 1960). The statistic D is can also be seen as the covariance of the indicator random variables acknowledging the presence of allele A at locus one and of allele B at locus two.

LD contrasts information of the joint frequency of the alleles (f_{AB}) with information from the marginal frequencies of both alleles (f_A and f_B). Although the latter relate to the classical SFS, the former relates to the joint SFS of two different sites, that be coined 'two-locus frequency spectrum' (Hudson 2001) or 'two-Sites Frequency Spectrum' or 2-SFS (Ferretti et al., 2018). In case of no recombination, the expectations (Ferretti et al., 2018) and the variances (Klassmann and Ferretti

2018) of the 2-SFS have been derived exactly in a closed formula. For intermediate recombination, results for small sample size were obtained (Golding 1984; Ethier and Griffiths 1990), or numerical solutions relying on simulations (Hudson 2001). Analytical results do not exist for the general case and one can imagine that it becomes even more difficult when the population size is not constant.

The distribution of LD carries both the information of the allele frequency of each allele and the joint distribution of allele frequencies. Consequently, like the SFS and the 2-SFS, the distribution of D is shaped by evolutionary history, and therefore can be potentially used to study the history of a population (Pritchard and Przeworski 2001). As it contains more information, LD is likely more informative than the SFS alone for inferring demographic history (Ragsdale and Gutenkunst 2017).

On a side note, it may also be possible to infer evolutionary history using the expected 2-SFS as a function of the genetic distance between two sites as is done for studies using LD. However, the recombination map is often not available so that the genetic distance is impossible to assess wit certainty.

In this work, we explore ways to use the information of the LD inside Maximal Linkage Disequilibirum blocks that are chromosomal regions with no detectable recombination in a given sample (Kerdoncuff et al. 2020). We also compute the LD between adjacent MLD blocks that are separated by one or few recombination events. We compare results obtained using linkage patterns vs. using SFS. To do so, we have developed new neutrality tests based on LD alone or combined with SFS. We further have used them to infer the parameters of a putative change in population size using simulations and maximum likelihood. Finally, we have analyzed the demography of three populations from different species: humans, gorillas and bonobos.

3.3.2 Materials and methods

3.3.2.1 Summary statistics

Linkage Disequilibrium For measuring LD between two loci, we only consider sites with two alleles, noted A/a for the first locus and B/b for the second one. Alleles A and B are the alleles with smaller frequency without any consideration to the unknown ancestral/derived state. The covariance definition provided above has been generalized to unphased data (Rogers and Huff, 2009) to:

$$D = 2(x_{AB}x_{ab} - x_{Ab}x_{aB}) \tag{3.1}$$

with x_{AB} a naïve estimate of f_{AB} assuming that the double heterozygote genotype has equal probability of the two possible phasing configurations (see Ragsdale and Gravel 2019 for details).

Maximal Linkage Disequilibrium (MLD) blocks MLD blocks (Kerdoncuff et al., 2020) are runs of single nucleotide polymorphisms compatible with a single

tree, i.e., where no recombination can be detected using the 4-Gamete Test (4GT) (Hudson and Kaplan 1985). They are separated by breakpoints inferred using the 4GT that detect incompatibilities between pairs of sites. We used an efficient algorithm to chop an alignment of 4 or more chromosomes into MLD (see Kerdoncuff et al. 2020) that minimizes the number of breakpoints needed to explain all incompatibilities. Importantly, the algorithm can be applied to both phased and unphased genomes.

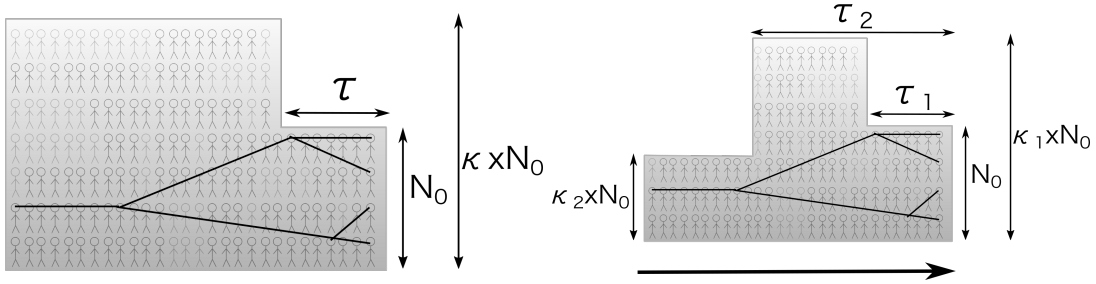
LD and MLD blocks In principle, D can be computed for any pair of sites using eq 3.1. In this work, we compute D for different categories of pairs of sites depending on the number of breakpoints between the two sites. We thus define D_k the measure of linkage disequilibrium for 2 sites separated by k breakpoints. In particular, in this study, we have chosen to focus on the distribution of D_0 and D_1 that are:

D_0: the D values of all pairs of sites located within the same MLD block,

D_1: the D values of all pairs of sites located in two adjacent MLD blocks.

Site Frequency Spectrum We also compute the folded-SFS that is the distribution of the Minor Allele Frequencies (MAF), disregarding the ancestral/derived status.

3.3.2.2 Demographic scenarios



(a) One change. **(b)** Two changes: "Podium".

Figure 3.10: Demographic models used for inference. A "one change" model (a), composed of two parameters: τ the date of the change and κ the strength of the change. A "Podium" model (b), composed of four parameters: τ_1 the date of the first change, τ_2 the date of the second change, κ_1 the strength of the first change and κ_2 the strength of the second change.

No change (H_0) In this scenario, the population size is constant and we have $N_t = N_0$ regardless of the time t in the past, measured in the coalescent time scale, that is N_0 generations per unit of time.

One change (H_1) In this second scenario, we assume a single abrupt change of population size (Fig 3.10a) at time τ. We denote by κ the strength of the change defined as: $\kappa = N_\infty/N_0$. When $\kappa = 1$, there was no change. When $\kappa > 1$, the population size has been suddenly reduced τN_0 generations in the past, whereas $\kappa < 1$ models cases of population growth.

Two changes (H_2) We finally consider two changes of population size described by four parameters: the population size has first changed $\tau_1 N_0$ generations in the past but had also changed $\tau_2 N_0$ generations ago, with $\tau_1 < \tau_2$. The strengths of both changes are respectively set by $\kappa_1 = N_{\tau_1<t<\tau_2}/N_0$ and $\kappa_2 = N_\infty/N_0$. This last scenarios generated more complex cases as there could increase or decrease in population size in two steps. For example, it can generate an ancient increase and a recent decrease in population size, creating a 'podium'-like demography (Fig 3.10b).

3.3.2.3 Inference methods

We simulated, using msprime (Kelleher et al. 2016), n haploid genomes and computed a folded-SFS normalized to 1 (diving each bin of the SFS by the total number of polymorphic sites) as well as the D_0 and the D_1 distributions, also normalized to 1. The three theoretical distributions are considered as probability distributions, numerically evaluated. Thus for any set of parameters, we compute the composite pseudo-likelihood of an observed genome by assuming that the frequency of a polymorphic site is an independent draw from the theoretical SFS. Similarly, the likelihood of the LD of a pair of sites located within the same MLD bloc is given by the theoretical distribution of D_0, and similarly for sites in adjacent MLD blocks using the D_1 distribution. For an observed SFS with counts $(\xi_1, \xi_2, ..., \xi_n)$, $S = \sum_{i=1}^{n} \xi_i$, and a set of parameters θ that generates a theoretical distribution $(p_1, p_2, ..., p_n)$, the likelihood is computed as a multinomial sampling:

$$\mathbf{L}(\xi_1, \xi_2, ..., \xi_n|\theta) = \binom{S}{\xi_1, \xi_2, ..., \xi_n} \prod_{i=1}^{n} p_i^{\xi_i}$$

A similar likelihood function can be defined for sampling pairs of sites with either an observed D_0 or D_1 distribution using the theoretical expectations.

As we wish to combine the information of the three distributions, we have to correct for the great difference in number of sites versus number of pairs of sites. Therefore, to combine the information of the three distributions, each log-likelihood is divided by its number of observations. The combined log-likelihood is simply obtained adding the three normalized log-likelihoods:

$$\ln L_{\text{combined}} = \frac{1}{\#\text{SNP}} \ln L_{SFS} + \frac{1}{\#D_0 \text{ pairs}} \ln L_{D_0} + \frac{1}{\#D_1 \text{ pairs}} \ln L_{D_1}.$$

One change In the case of the "one change" scenario, we simulated genomes for different values of the pair (τ,κ) uniformly spread in a log-scaled grid of $\tau \in [0.001 : 10]$ and $\kappa \in [0.01 : 100]$. Likelihood is measured for each pair of parameters and the pair (τ, κ) with the highest likelihood is retained.

Two changes As likelihood maximization on a 4-dimensional grid was computationally too costly, we decided to rely on random sampling of the parameter space. We sampled 100,000 values of the 4-tuple $(\tau_1, \tau_2, \kappa_1, \kappa_2)$. For each replicate, the four parameters were drawn log-uniformly in $\tau_1 \in [0.001 : 0.01]$, $\tau_2 \in [0.1 : 10]$, $\kappa_1 \in [0.031 : 20]$ and $\kappa_2 \in [0.031 : 20]$. From this random sample of the parameter space, we retained the ones with the highest likelihoods (the best 1%) evaluated either for each summary statistic (SFS, D_0, D_1) and for the combination of the three. We analyzed the posterior parameter distributions to infer the demographic scenarios. For each parameter and each distribution, we selected for each parameter estimate the mode of the likelihood distribution for that parameter (at this stage the mode is inferred visually).

3.3.2.4 Neutrality test

We developed a novel neutrality test based on the distribution of the likelihood under a constant population scenario considering only the SFS, D_0 or D_1 or considering the combination of the three.

As is done classically in statistics, we defined Z_n as an average of n realized random variables X_j:

$$Z_n = \frac{1}{n} \sum_{j=1}^{n} X_j.$$

The central limit theorem states that, when the X_j are independent and identically distributed, for large n, Z_n converges in distribution to their common expectation $\mu = E(X_1)$ and will be distributed around μ like a Normal Gaussian variable with variance: $Var(X_1)/n$.

Here, we define X_j as the log-likelihood of one of the three statistics at the j-th site or pair of sites. The variance of the average should be decreasing like $1/n$ if all sites (pairs of sites) were independent. Because they are not rigorously independent, we empirically found that the variance decreases like $1/n_e$, where $n_e = n^{1+\alpha}$ with α slightly negative. In this case, the variance becomes:

$$Var(Z_n) = \frac{C}{n^{1+\alpha}}$$

The values of μ, C and α depend on the summary statistics considered (SFS, D_0, D_1) so does the combination of the three. Values were estimated numerically by fitting simulations of constant size. All the values of the corresponding distributions can be found in Table 3.1.

Parameter	SFS	D_0	D_1	Combined
μ	-3.3672	-2.1771	-2.1259	-7.6702
α	-0.1053	-0.0748	-0.0725	-0.0823
C	$e^{2.9195}$	$e^{7.5459}$	$e^{7.9202}$	$e^{10.097}$

Table 3.1: Parameters describing the distribution of each Z_n used for the neutrality test in function of the summary statistic used (SFS, D_0, D_1 and the combination of the three). μ is the mean of the distribution, α and C parameterize a functional form of the variance. All the parameters are issued from simulations.

The value Z_n that is an average of log-likelihoods being distributed as a normal random variable with known mean and variance, it can be easily tested whether the observed Z_n is outside the 95% variation interval. In this case, we reject the constant population size hypothesis (H_0) when $Z_n > \mu + 1.96\sqrt{Var}$ or $Z_n < \mu - 1.96\sqrt{Var}$, with μ and Var calculated from the parameters of Table 3.1.

Power of the test To assess the power of the test, we simulated m replicates (with $m > 1,000$) under H_1 (a single change in population size) and test the value of a Z_n against the one expected under H_0. The power of a test is the frequency at which it rejects H_0 under H_1. Here, rejecting H_0 indicates that we detect a change in population size by the test. A power of 1 means that the test detects the change for all replicates. If α is the Type I error rate chosen to design the test (here $\alpha = 0.05$), then a power equal to α indicates that the test cannot distinguish between H_0 and H_1. The power of the test indicates when the test can be used to detect a change in population size.

3.3.2.5 Data

Population	n	Phased	SNP	MLD	D_0 pairs	D_1 pairs	Source
Yoruba	50	Yes	1,373,597	98,643	10,507,041	10,793,390	Consortium
Gorilla gorilla	23	No	837,832	14,176	111,270	35,362	Prado-Martinez et al.
Pan panicus	10	No	490,495	1,353	10,422	2,461	Prado-Martinez et al.

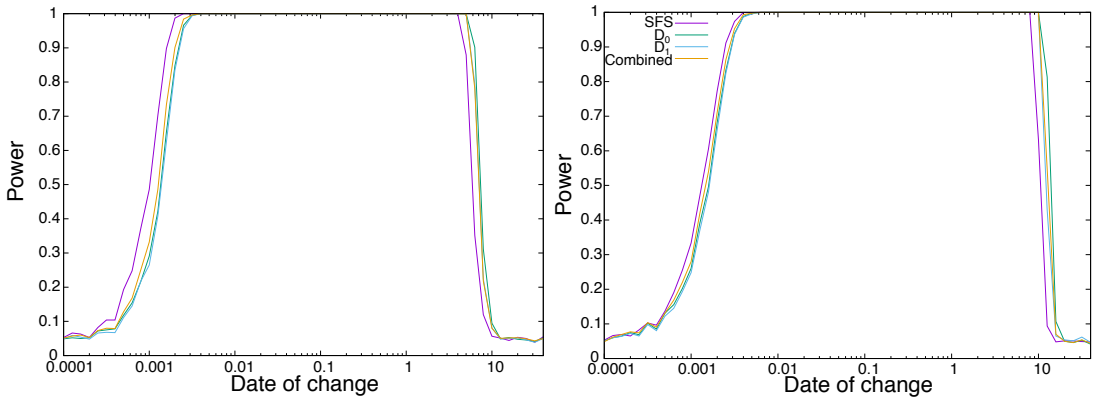
Table 3.2: Data description. For each sample of the three populations (Yoruba, *Gorilla gorilla* and *Pan panicus*), the number n of diploid individuals sampled, the phasing information, the number of SNPs, the number of MLD blocks, the number of D_0 and D_1 pairs measured and the source of the data.

We collected 3 datasets from public databases, as described in Table 3.2. We sub-sampled 50 individuals of the Yoruba population from the 100 Genomes Project (Consortium 2015), 23 unrelated western lowland gorillas (*Gorilla gorilla*) and 10 wild bonobos (*Pan panicus*) from the Great Ape Genome Project (Prado-Martinez

et al. 2013). We used the first chromosome for each species and applied the callable mask given by the authors for our study.

3.3.3 Results

3.3.3.1 Theoretical analysis

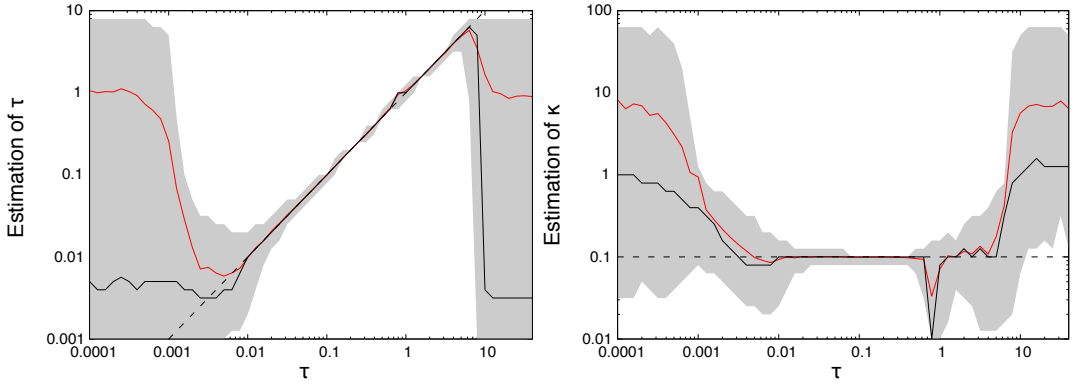


(a) Increasing population $\kappa = 0.1$ **(b)** Decreasing population $\kappa = 10$

Figure 3.11: Power of detection of the test for each scenario: Increasing population with $\kappa = 0.1$ (3.11a) and Declining population with $\kappa = 10$ (3.11b) considering different summary statistics: the SFS (purple), D_0 (blue), D_1 (green) and the combination of the three (orange).

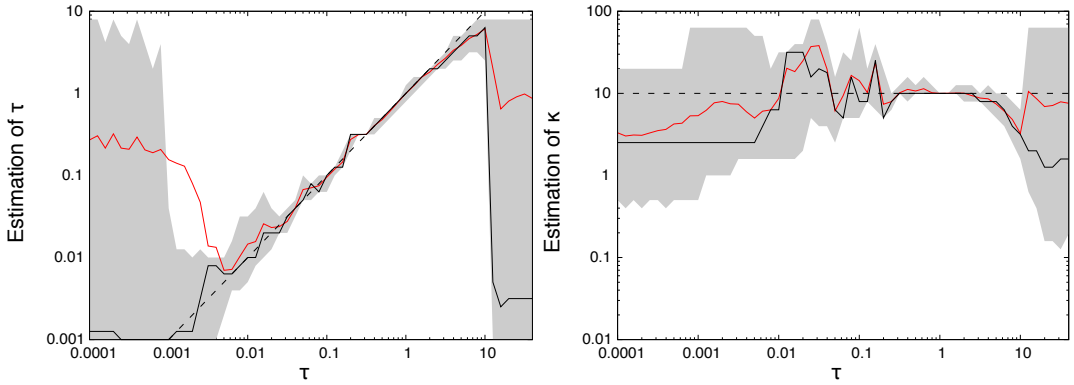
Power of the test The range of dates τ where the change is detectable is quite similar among the statistics used, in both the case of an increase and a decline in population size (Fig 3.11). However, the statistics D_0 and D_1 (and their combination) detect the change for a small range of large values when the SFS has lost the signal. The range of τ depends on the scenario chosen: the test detects an increase in population size (Fig 3.11b) earlier than a decrease (Fig 3.11b). Inversely, it detects a decrease later than an increase. For an increase in population size, the SFS-based test detects earlier changes than the other distributions, and the difference is less remarkable for a decrease in population.

Estimation The combined distribution has a good sensitivity to detect a change in population size for a large range of τ values. The estimation of parameters using the log-likelihood of the combined distribution is also good for a large range of τ (Fig 3.12). Both τ (Fig 3.12b and 3.12d) and κ (Fig 3.12a and 3.12c) are correctly inferred using the combined log-likelihood, in both a scenario of a decline (Fig 3.12d and 3.12c) and an increase (Fig 3.12b and 3.12d). The estimation of τ is accurate for a similar range as the sensitivity to detection. The estimation of κ can have some inaccuracies for the most recent τ values.



(a) Estimation of τ (for $\kappa = 0.1$)

(b) Estimation of κ (for $\kappa = 0.1$)



(c) Estimation of τ (for $\kappa = 10$)

(d) Estimation of κ (for $\kappa = 10$)

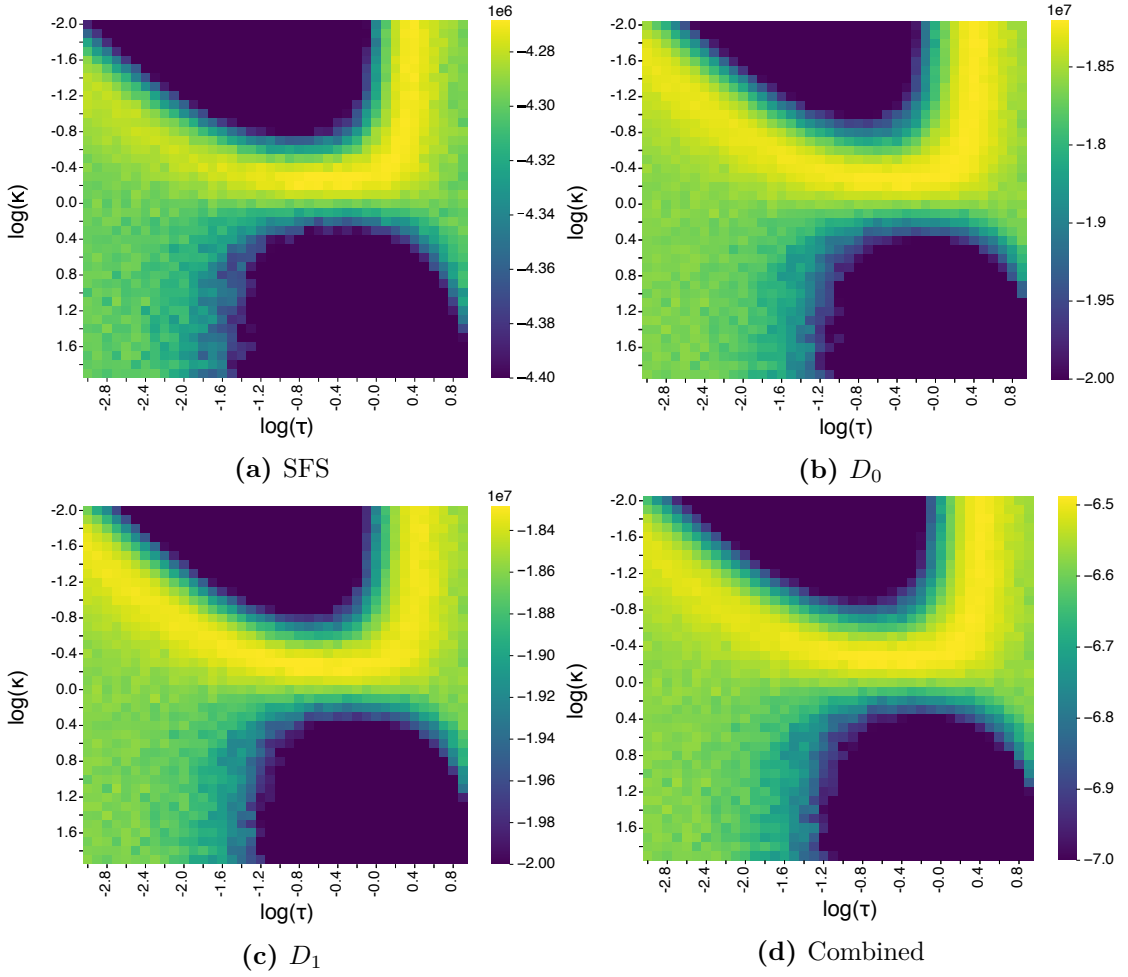
Figure 3.12: Estimation of parameters τ (left panels) and κ (right panels) using the combination of the three distributions (SFS, D_0 and D_1), in function of τ and the demographic scenario: an increasing population ($\kappa = 0.1$, upper panels) and a declining population ($\kappa = 10$, lower panels). The dashed line is the true value of the parameter, the red line is the mean of the estimation, the black line is the median and the grey area is the interquartile range.

3.3.3.2 Application - Inferences

On change The best pairs (τ, κ) inferred from the different distributions are gathered in Table 3.3.

Population	SFS		D_0		D_1		Combined	
	τ	κ	τ	κ	τ	κ	τ	κ
Yoruba	0.25	0.5	2	0.19	0.5	0.63	0.2	0.5
Gorilla gorilla	2	0.02	0.03	63	0.5	5	0.08	4
Pan panicus	1	0.013	3.2	0.013	1.3	3.2	0.5	0.8

Table 3.3: Best inferences of τ and κ for the different populations (Yoruba, *Gorilla gorilla* and *Pan panicus*) for each statistic considered (SFS, D_0, D_1 of the combination of the three).

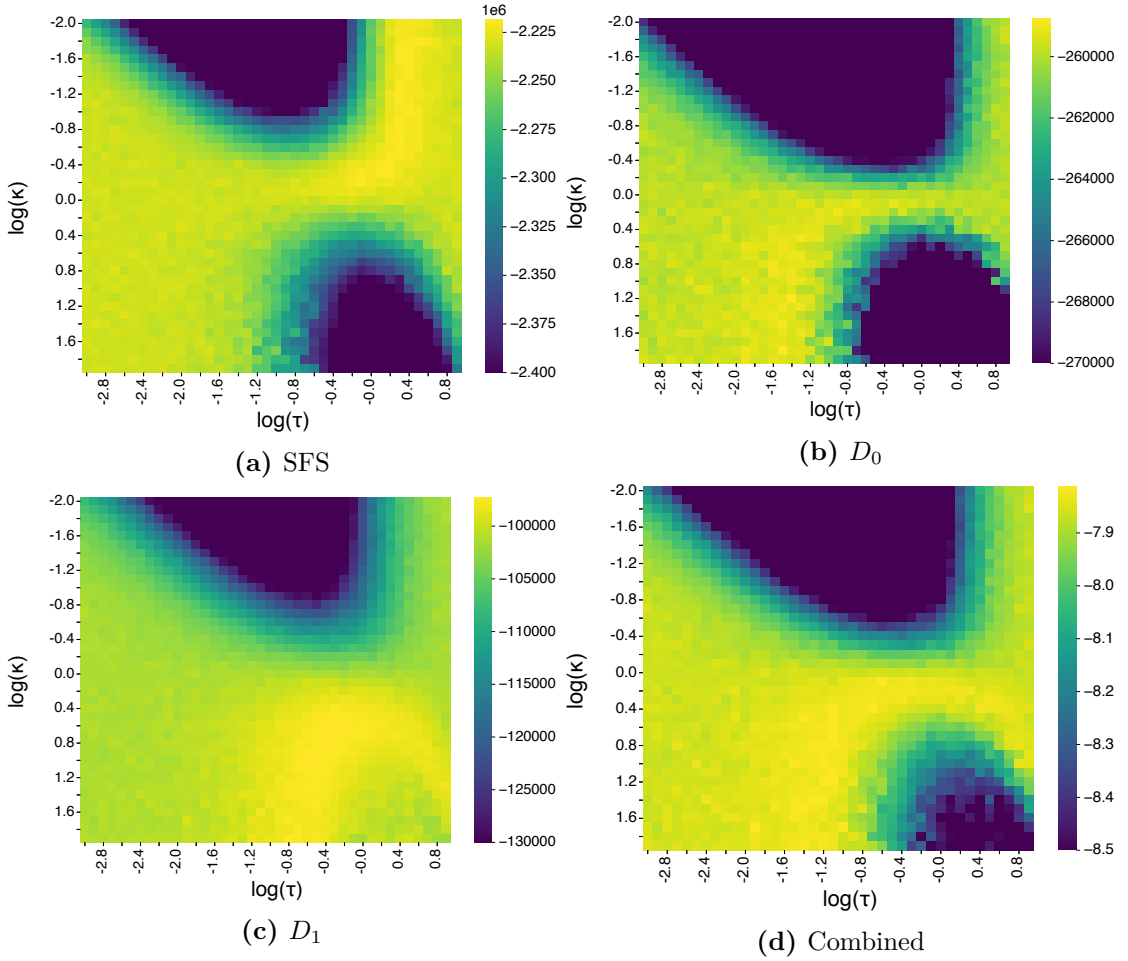


(a) SFS **(b)** D_0 **(c)** D_1 **(d)** Combined

Figure 3.13: Pseudo-Likelihood landscapes of 4 summary statistics: SFS (a), D_0 (b), D_1 (c) and the combination of the three (d) in function of τ (x-axis) and κ (y-axis) (log-scaled) measured on the first chromosome of the Yoruba genomes.

Yoruba Considering the Yoruba population, the four distributions (Fig 3.13) indicate the same scenario: an increasing population, as expected in the literature (Lapierre et al. 2017). The parameters inferred still vary from 0.2 to 2 for τ and from 0.19 to 0.63 for κ. However, looking at the likelihood landscapes, the four distributions support, with some variance, the same pattern. The SFS landscape (Fig 3.13a) seems to be the most accurate in that case.

Gorilla For the gorilla population, parameter estimations point to different scenarios (Table 3.3): the SFS point to an increasing population and the the two D-based distributions (and so the combination of the three) point to a decreasing population. Studying the likelihood landscapes (Fig 3.14), it is difficult to agree on a single scenario. As said previously the SFS points to an increasing population at an ancient time (Fig 3.14a), the D_0 distribution points to a decreasing population at a recent time (Fig 3.14b) and the D_1 distributions at an older time (Fig 3.14b).

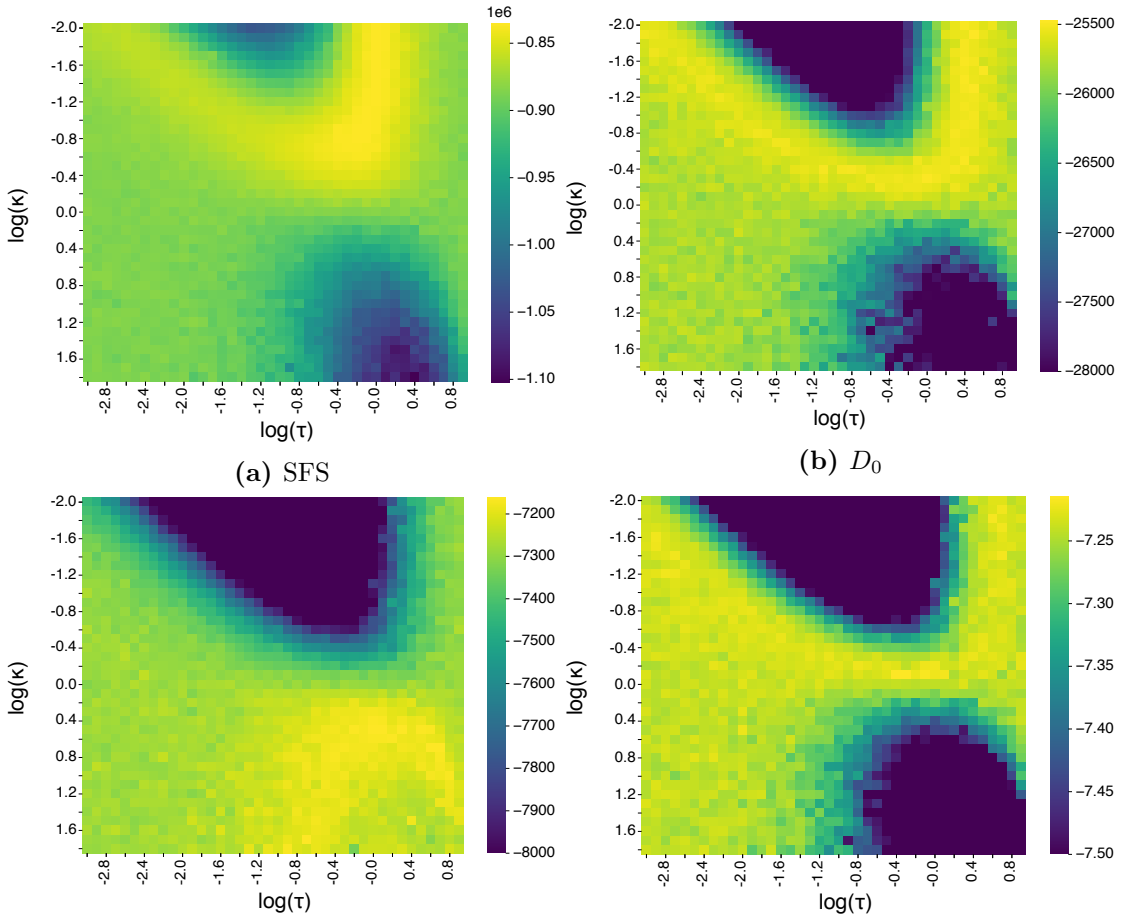


(a) SFS **(b)** D_0 **(c)** D_1 **(d)** Combined

Figure 3.14: Pseudo-Likelihood landscapes of 4 summary statistics: SFS (a), D_0 (b), D_1 (c) and the combination of the three (d) in function of τ (x-axis) and κ (y-axis) (log-scaled) measured on the first chromosome of the Gorilla genomes.

The combination of the three distributions suggests a decline at a time situated between the values estimated by D_0 and D_1 respectively (Fig 3.14d).

Pan panicus The Bonobo population shows us another configuration. Estimations point to an increasing population for the SFS and the D_0 distribution (and also the combination), interestingly the D_1 distribution does not point to the same scenario as the D_0, but rather to a declining population (Table 3.3). The likelihood landscapes reflect the same configuration (Fig 3.15). However, the signal seems more confused on the combined distribution landscape (Fig 3.15d): the landscape seems nearly flat. As for the gorilla population, these not consistent likelihood landscapes show us a misspecification problem. The bonobo is also endangered with extinction and the use of the D_1 distribution helps us see it.

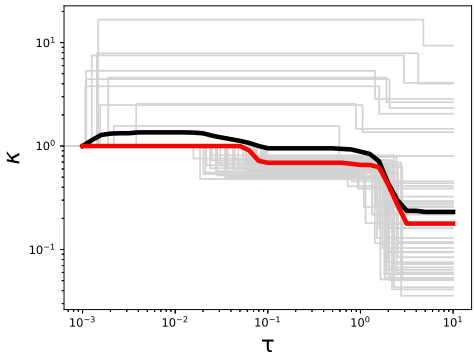


(a) SFS **(b)** D_0

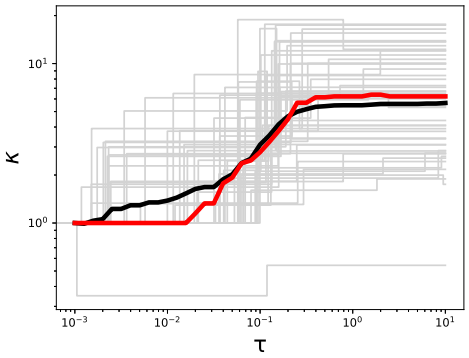
(c) D_1 **(d)** Combined

Figure 3.15: Pseudo-Likelihood landscapes of 4 summary statistics: SFS (a), D_0 (b), D_1 (c) and the combination of the three (d) in function of τ (x-axis) and κ (y-axis) (log-scaled) measured on the first chromosome of the *Pan panicus* genomes

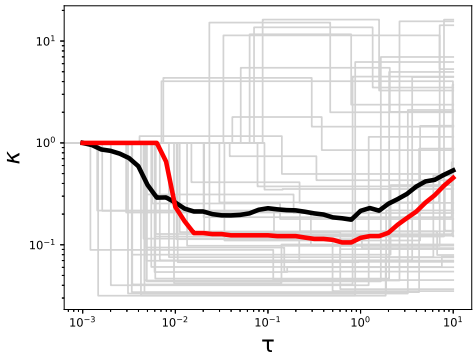
Two changes Demographic scenarios of the 50 best simulations based on combined likelihood, for each population are illustrated in Fig 3.16.



(a) Yoruba



(b) Gorilla



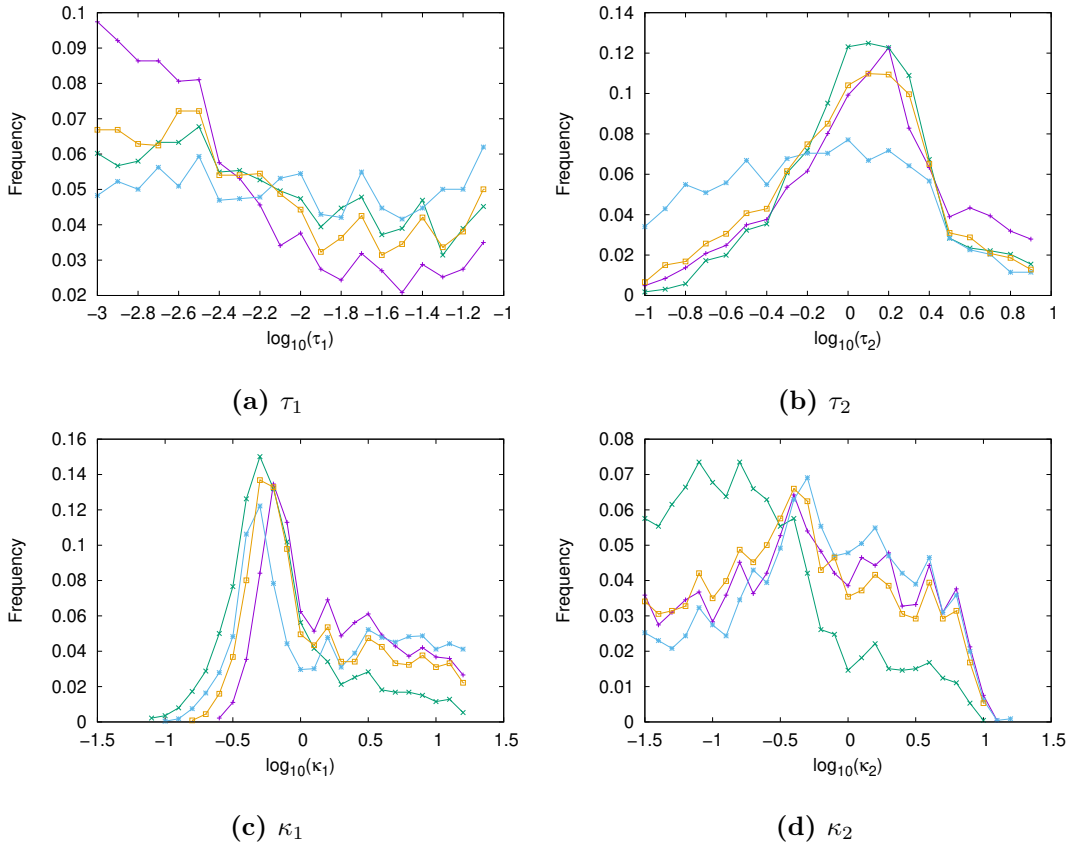
(c) Pan panicus

Figure 3.16: Demographic scenarios of the 50 best simulations (grey) based on combined likelihood, with median population sizes (red) and geometric mean population sizes (black). The x-axis is the date of change (τ) and the y-axis the strength of changes (κ). For (a) the Yoruba population, (b) the *Gorilla* population and (c) the *Pan panicus* population.

Yoruba The scenario with two changes inferred for the Yoruba population is a two-step increase in population size for all the summary statistics used (Table 3.4). Globally, all the summary statistics show similar parameter distributions (Fig 3.17). Concerning the dates of change, the SFS points towards a more recent first date of change than the other distributions (Fig 3.17a) and the D_1 distributions has a flatter mode for the inference of the second date of change compared to the others (Fig 3.17b). The κ_1 distributions have all the same shape (Fig 3.17c) and only the D_0 distribution points towards a stronger increase for κ_2 (Fig 3.17c). The best scenarios based on the combined likelihood are, in majority, increase in population size (Fig 3.16a).

Summary statistics	τ_1	τ_2	κ_1	κ_2
SFS	0.001	1.58	0.63	0.4
D_0	0.0031	1.26	0.5	0.16
D_1	0.0031	1.26	0.5	0.5
Combined	0.0031	1	0.5	0.4

Table 3.4: Inferred values of τ_1, τ_2, κ_1 and κ_2 for the Yoruba population, in function of the summary statistic considered (SFS, D_0, D_1 of the combination of the three).



(a) τ_1 **(b)** τ_2

(c) κ_1 **(d)** κ_2

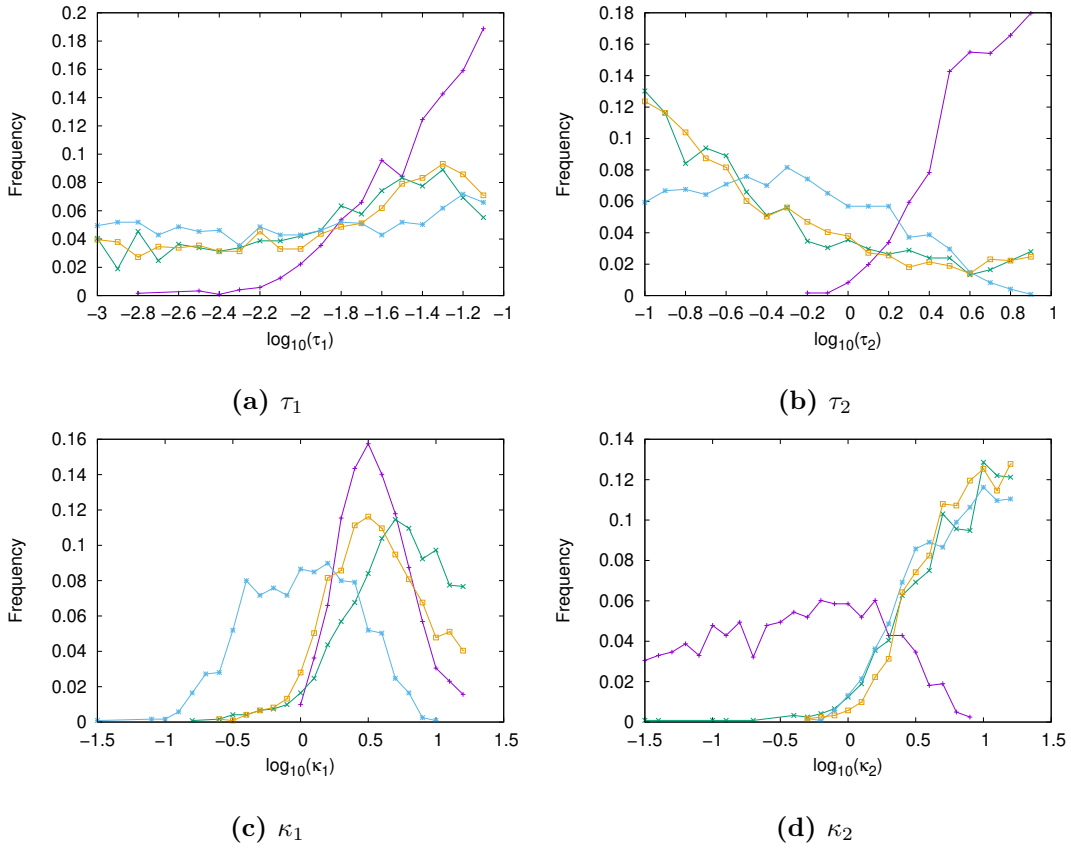
Figure 3.17: Distribution of the parameters from the 1% best-likelihood simulations for τ_1 (a), τ_2 (b), κ_1 (c) and κ_2 (d). For different summary statistics calculated on the chr1 of yorubas: SFS (purple), D_0 (green), D_1 (blue) and the combination of the three (orange).

Gorilla Two different scenarios are inferred for the gorilla population: a two-step decrease in population size for the D_0, D_1 and the combined distributions and a 'podium'-like scenario with $\kappa_2 < \kappa_1$ for the SFS, also inferred from the SFS in McManus et al. (2015) (Table 3.6).

The parameters distributions are more divergent than the Yoruba's ones. For τ_1, the SFS parameters distribution indicates the same range of values than the other distributions but with a stronger signal (Fig 3.18a). The parameter τ_2 seems

Summary statistics	τ_1	τ_2	κ_1	κ_2
SFS	0.079	7.943	3.16	1.58
D_0	0.05	0.01	5	10
D_1	0.063	0.5	1.58	10
Combined	0.05	0.01	3.16	20

Table 3.5: Inferred values of τ_1, τ_2, κ_1 and κ_2 for the gorilla population, in function of the summary statistics considered (SFS, D_0, D_1 of the combination of the three).



(a) τ_1 **(b)** τ_2

(c) κ_1 **(d)** κ_2

Figure 3.18: Distribution of the parameters from the 1% best-likelihood simulations for τ_1 (a), τ_2 (b), κ_1 (c) and κ_2 (d). For different summary statistics calculated on the chr1 of gorillas: SFS (purple), D_0 (green), D_1 (blue) and the combination of the three (orange).

difficult to infer as the SFS, the D_0 and the D_1 distributions have different shape of parameter distribution (Fig 3.18b). The SFS indicates a more ancient time, the D_0 and the combined distributions a more recent time and the D_1 distribution an intermediate time (Fig 3.18b). Concerning the strengths of the change in populations size, declines are inferred for all the summary statistics. The value of κ_1 changes in function of the summary statistics, the D_0 distribution indicates a high value, the D_1 distribution a small one, and the SFS and the combined dis-

tribution an intermediate value (Fig 3.18c). For κ_2, all the D_0, the D_1 and the combined distributions have the same shape and indicate a higher decline than κ_1 (Fig 3.18d). Nearly all the best scenarios based on the combined likelihood are two-step decreases in population size (3.16b). For κ_2 the SFS has a flatter parameter distribution which indicates a podium-like scenario (Fig 3.18d).

Pan panicus The scenario with two changes is more complex to infer for the bonobo population. For τ_1 all the summary statistics have a flat distribution (Fig 3.19a). For τ_2, the SFS and the D_1 distribution indicate a late change, at the opposite the D_1 distribution indicates a more recent change and the combined distribution carries no information about this date of change (Fig 3.19b). Concerning the strengths of change, the SFS and the D_0 distribution show the same pattern: a decrease for the recent change and an increase for the ancient change (Fig 3.19c and Fig 3.19d). It is a podium-like scenario. The D_1 distribution has a parameter distribution around $\kappa_1 = 1$: an absence of change in population size and a κ_2 distribution indicating a decline in population size (Fig 3.19c and Fig 3.19d). The combined distribution is still no really informative about the strengths of change. Indeed, the best scenarios based on the combined likelihood show really different histories (Fig 3.16c).

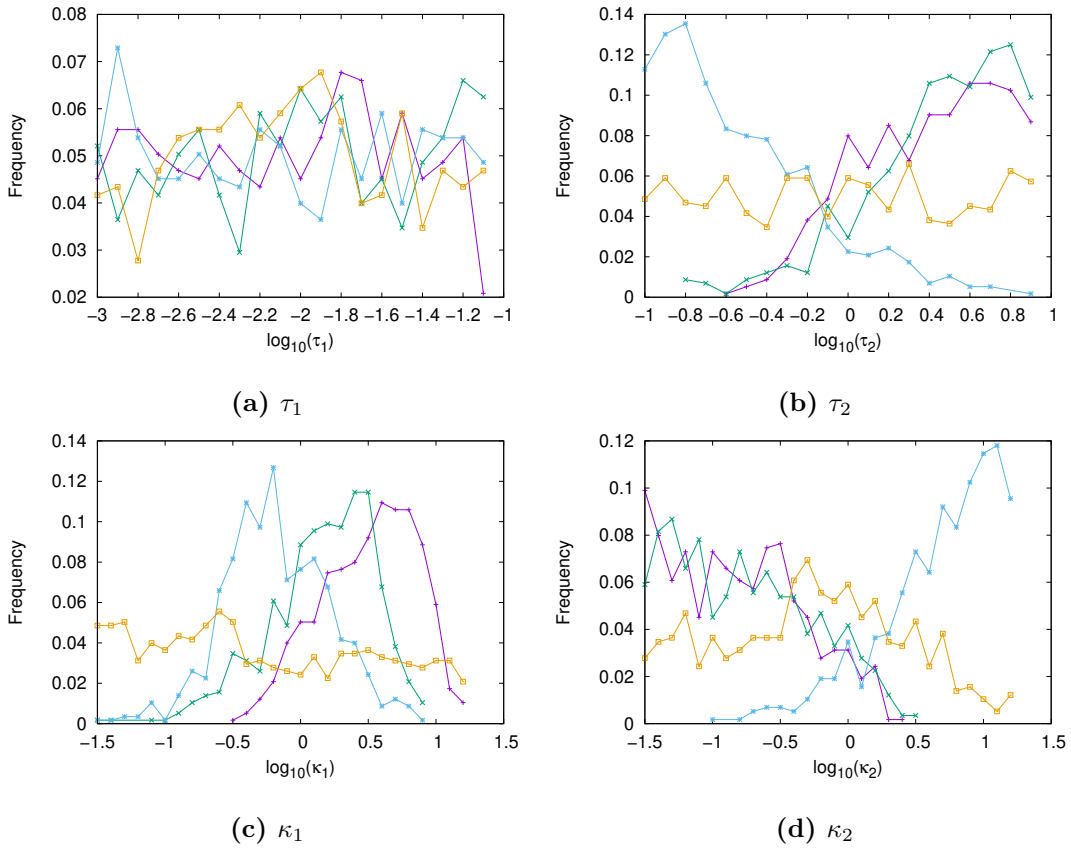
Summary statistics	τ_1	τ_2	κ_1	κ_2
SFS	0.0158	3.98	4.0	0.03
D_0	0.063	6.3	2.8	0.05
D_1	0.0012	0.16	0.63	12.6
Combined	0.0125	2.0	0.4	0.5

Table 3.6: Inferred values of τ_1, τ_2, κ_1 and κ_2 for the bonobos population, in function of the summary statistics considered (SFS, D_0, D_1 of the combination of the three).

3.3.4 Discussion

In this chapter, we introduced two new summary statistics D_0 and D_1 distributions, based on linkage disequilibrium segmented in function of the topology of the underlying coalescent trees, with MLD blocks. D_0 considers the LD inside each MLD block and D_1 the LD along neighbouring MLD blocks.

We compare the distribution of D_0 and D_1 to the SFS when computed on the same multiple alignment sequences and study the combination of the three distributions. We developed neutrality tests based on those distributions that show a great power for detecting both increase and decline in population size. However, considering a simple scenario of only one change in population size, the use of the combined distribution does not drastically increase the power or the efficiency of the demographic inferences.



(a) τ_1 **(b)** τ_2

(c) κ_1 **(d)** κ_2

Figure 3.19: Distribution of the parameters from the 1% best-likelihood simulations for τ_1 (a), τ_2 (b), κ_1 (c) and κ_2 (d). For different summary statistics calculated on the chr1 of bonobos: SFS (purple), D_0 (green), D_1 (blue) and the combination of the three (orange).

However, demography does not impact each summary statistic in the same way. As we saw analyzing real datasets, the three summary statistics can point to the same scenario (in the case of the Yoruba population) or to different scenarios (in the cases of the Gorilla and the Bonobo populations). Even D_0 and D_1 both based on linkage disequilibrium can point to different scenarios (e.g., for Bonobo). The summary statistics based on two-locus statistics seems to be more sensitive to demographic history than single-locus statistics (shown in Ragsdale and Gutenkunst 2017), as a solely SFS-based method always points to increasing population with the chosen datasets.

Misspecification is an important concern for the inference of evolutionary history. It is not possible to differentiate evolutionary history scenarios using only one summary statistic, as shown for different demographic scenarios with the SFS (Lapierre et al. 2017) or to differentiate between structure and demography using IICR (Mazet et al. 2016).

The combination of summary statistics not based exactly on the same information will help determining the evolutionary history of the population.

Chapitre 4

U-shaped genome site frequency spectra : challenging the reference model of molecular evolution ?

Contents

4.1 Résumé de l'article

4.1.1 Motivation

Dans cet article, nous confrontons 45 SFS issus de populations diverses, échantillonnées à différents endroits dans l'arbre de la vie, à d'autres modèles que le modèle standard neutre basé sur le coalescent de Kingman. Ces modèles autorisent la potentielle existence de multifurcations dans la généalogie, des noeuds avec plus de deux descendants. Ils sont appelés *Multiple Merger Coalescents* (MMC), nous considérons deux classes : le Beta-MMC (Schweinsberg 2003) et le Psi-MMC (Eldon and Wakeley 2006). Ils sont tous les deux calibrés avec un seul paramètre (α pour le Beta-MMC et ψ pour le Psi-MMC) permettant aux généalogies d'être comprises entre des bifurcations simples et la radiation de la population. Nous avons intégré la démographie à ces modèles comme une croissance exponentielle dépendant d'un unique paramètre.

Les SFS collectés sont dépliés et ont, pour une grande majorité, une forme en U. Cette forme ne peut être expliquée par un coalescent de Kingman avec de la démographie, mais peut être produite par un MMC. Elle peut être également produite par d'autres processus, comme par exemple une erreur au niveau de l'ancestralité de l'allèle. Ces deux possibilités sont considérées dans notre article, afin d'estimer quels modèles entre le coalescent de Kingman (avec démographie et erreurs d'ancestralité) et les deux MMC (avec également démographie et erreur d'ancestralité) expliquent le mieux les données observées.

4.1.2 Principaux résultats

Nous avons trouvé qu'une large majorité (73%) de SFS était mieux expliqué par un MMC que par le coalescent de Kingman. Le modèle le plus approprié est souvent le Beta-MMC (51%), puis le Kingman (27%) et enfin le Psi-MMC (13%). Dans quelques cas, il n'était pas possible de différencier quel MMC était le plus approprié (9%). Les deux modèles de multifurcation indiquent des résultats similaires.

32 des 45 SFS sont « bien » illustrés par un des trois modèles, cependant 13 restent à expliquer.

La présence de multifurcation et la force de ces multifurcations (représentée par le paramètre de MMC), sont bien dispersées dans l'ensemble de l'arbre de la vie échantillonné. La présence de multifurcation ne semble pas dépendre de l'espèce considérée.

Si un coalescent de Kingman est inféré à la place d'un MMC, le taux d'erreur d'ancestralité sera identique, cependant le taux de croissance sera très différent : il sera très surestimé.

4.1.3 Conclusion

Nous avons pu démontrer que la présence de multifurcations dans les généalogies est beaucoup plus fréquente que ce que l'on pouvait penser. Un modèle permettant les multifurcations serait donc plus approprié à décrire l'ensemble des généalogies des populations. Le MMC pourrait donc être considéré comme un modèle standard neutre, plus inclusif que le coalescent de Kingman.

Il est cependant important de noter que nous n'avons pas réussi à expliquer tous les formes des SFS échantillonnés, il reste donc des mécanismes à comprendre/considérer dont les MMC ne prennent pas en compte dans leur forme actuelle.

4.2 Article

U-shaped genome site frequency spectra: challenging the reference model of molecular evolution?

F Freund*, E Kerdoncuff*, S Matuszewski, M Hildebrandt, J.D. Jensen, L Ferretti, A. Lambert, T. Sackton, G. Achaz

September 30, 2021

Abstract

We gathered a collection of 45 genomic site frequency spectra from a diverse set of species. Most of them display an excess of low and high frequency variants compared to the expectation of the standard neutral model, resulting in U-shaped spectra. We show that models of multiple merger coalescents are however better candidates to predict this U-shape pattern in most of the studied species. This study raises the question of revisiting the standard neutral model by incorporating multiple merger genealogies and thus defining a better reference model that fits the observed patterns of genomic diversity.

1 Introduction

Since the advent of the *neutral theory of molecular evolution* [Kim68, Kim83], patterns of molecular diversity in a population are assumed to result from a mutation-drift equilibrium. New variants are added to the genetic pool by mutations. Selectively neutral variants are lost (or fixed) by genetic drift after some time, while non-neutral ones become so in much smaller time periods. Thus, the latter do not contribute much to the observed genetic diversity across the genome. As new variants appear in very few individuals, sometimes a single, their frequencies distribution is skewed toward low values [Ewe72]. The simplest version of the mutation-drift equilibrium, known as the *Standard Neutral Model* (SNM), further assumes that the population is panmictic and has a fixed constant size. Under these assumptions, genetic diversity of a sample of individuals can be characterized by tracking mutations on a purely bifurcating, random genealogy tree which distribution is given by the Kingman's coalescent [Kin82].

One common metrics used to study the consistency between the SNM and the observed data is the *Site Frequency Spectrum* (SFS), that is the distribution of mutation frequencies, typically computed for a sample of n haploid genomes. Under the SNM, the expected SFS, averaged across the tree space, is given by $E[\xi_i] = \theta/i$, where ξ_i is the number of sites that carry a derived variant of frequency i/n [Fu95]. The θ parameter of the SNM is defined as $\theta = 2pN\mu$, with p the ploidy (typically 1 or 2), N the population size and μ the mutation rate. Given that the genome consists of genomic regions whose ancestries are partially decoupled by recombination, the genome-wide SFS resembles an average SFS ; indeed, the effect of recombination can be approximately seen as changing locus-specific genealogy trees alongside the genome resulting in pseudo-unlinked loci that are partially independent [MC05, MW06, Wal00]. For samples of recombining genomes, an easy visual check for the relevance of the SNM is to represent a transformed SFS ($\phi_i = i\xi_i$), which has a uniform expectation under the SNM [NT08, Ach09, LLA17].

Observed SFS have usually not the hyperbolic shape expected under the SNM. As it is usually assumed that almost all polymorphisms are neutral, global deviations from the expectation of the SNM are generally

interpreted as non-selective effects, such as non-constant demography or population structure; the effect of selection, being often thought to affect only few loci spread across the genome. The relative importance of drift and selection is however still being debated [KH18, JPS$^+$19]. Barring selective sweeps and very strong population size changes, the majority of genealogy trees still stay bifurcating and can be described as variants of Kingman's coalescent, where only the branch lengths are distorted [Hud90]. A standard procedure in population genetics used for inference methods is thus to first statistically test for the SNM (treated as H_0, a null statistical model) and then, when rejected, introduce non-constant demography and/or population structure.

So to test for the generality of the SNM, a relevant starting point is to collect population genomic data sets of different species and test if the SNM fits well with the SFS (*i.e.* whether the transformed SFS is flat). In this article, we show that among a random collection of genome-wide SFS, many show an unexpected excess of low and high frequency variants, resulting in a U-shaped SFS, especially striking in the transformed ϕ SFS representation.

Although demography can only produce monotonically decreasing SFS [SW08], several violations of the SNM can produce U-shaped SFS. Some are classic processes of population genetics like recent migration from non-sampled populations [ME20], population structure [LBL$^+$16], confusion between ancestral and derived alleles [BD03], while other effects are less studied in standard analyses: biased gene conversion [PATE18], positive selection at many targets across the genome [BWSH01], background selection [CGD18] or selection fluctuating through time [HSDB08] (as well as mixtures of e.g. positive and background selection [LBL$^+$16]).

An interesting violation of the SNM that leads to genome-wide U-shaped SFS is the presence of multiple mergers in genealogies, which distributions are given by one of the Multiple Merger Coalescent (MMC) [Sag99, Pit99, DK99, MS01, Sch00]. In a nutshell, MMCs arise when the number of offspring per individual has very high variance across the population. MMC results in nodes with more than two descendants, which are known as polytomies in phylogeny. Although several biological scenarios can lead to the emergence of MMC [TL14], they all originate from the existence of individuals or genomes that concentrate the paternity (forward time) or the ancestry (backward time) of a macroscopic fraction of the population in short time window (up to a single generation). Such effect of concentrations have been reported in various species across all kingdoms of life [Mon16]. Furthermore, support for MMC genealogies has been observed for a large variety of species ranging from bacteria (e.g. for *Mycobacterium tuberculosis* [MASSJ20, MGF20]) to animals (e.g. the nematode *Pristionchus pacificus* [RNW$^+$14], multiple fish species, e.g. [ÁH14, NNY16]) and even in cancer cells [KVS$^+$17].

The three main mechanisms leading to MMC proposed so far are *sweepstake reproduction*, pervasive recurrent *rapid selection* and *genetic draft*. The term sweepstake reproduction has been proposed for species that have rare individuals with a high reproduction rate coupled with high early-life mortality: by chance, a single or few individuals become ancestors of a macroscopic fraction of the population, thus resulting in MMC genealogies (for a review, see [Eld20]). Multiple models featuring recurrent rapid emergences of genotypes with high fitness also result in MMC genealogies, often modelled by the Bolthausen-Sznitman coalescent or related models, e.g. [BD13, NH13, DWF13, BBS13, Sch17]. Genetic draft is the effect of recurrent positive selection on a partially linked neutral locus (repeated hitchhiking) [SH74, Gil00, SD05]. Interestingly, other biological factors can also lead to MMC genealogies: large rapid demographical deviations [BBM$^+$09], seed banks [CCSWB20], extinction-recolonisation in metapopulations [TV09] and range expansions [BHK21] (the latter could be present across a wide range of species). For some scenarios, MMC genealogies can be expressed on biologically sound time scales so that it is possible to incorporate population structure and demography [MHAJ18, KB19, Fre20].

In this study, we randomly collected 45 species (Table 1) spread throughout the tree of life (bacteria, plants, invertebrates and vertebrates) for which genome-wide polymorphic data (with sample size $n > 10$) were available together with an outgroup to assign ancestral and derived alleles. Most of the resulting genomic SFS display a U-shaped SFS or nearly U-shaped. We further show that MMC genealogies augmented with non constant demography and confusion between the ancestral and the derived alleles fit statistically better most observed SFS than Kingman genealogies with demography and confusion errors. We have tested two simple MMC models: Beta-MMC [Sch03] and Psi-MMC [EW06], both tuned by a single parameter that interpolates between pure radiation to a Kingman-like tree. Demography is here given by a single parameter (a simple exponential growth), so is the amount of confusion errors. Using composite-likelihood maximisation [Nie00] on genome-wide data, we were able to statistically disentangle the demography, confusion errors and strength of multiple-mergers. We finally discuss the potential origin of the U-shape and how it challenges the universality of the SNM in population genetics.

2 Materials and Methods

2.1 Coalescent and allele confusion models

We compare the biological observed SFS to the theoretical SFS expected under models of genealogies with mutations. The genealogy models emerge from suitable discrete generation reproduction model. Tracing back a sample's genealogy in a Wright-Fisher model or in a Moran model leads to Kingman's coalescent. MMC emerges when the variance in offspring number is much larger than in the Wright-Fisher or the Moran models. Each coalescent is a (random) tree with n leaves which approximates the genealogy for a sample of size n in a reproduction model with population size N very large ($N \to \infty$). One unit of time in the coalescent tree corresponds to many generations in the underlying reproduction model: for Kingman's coalescent one time unit corresponds to N generations of an haploid Wright-Fisher model or order of N^2 time steps of the haploid Moran model. This correspondence affects how population size changes in the reproduction model are reflected in the coalescent approximation (see definition below, for mathematical justification and details see [GT94, MHAJ18, Fre20]). On the genealogy tree, mutations are placed randomly via a Poisson process with rate $\theta/2$ and interpreted under the infinite sites model.

We compared three different coalescent models: Kingman's n-coalescents, Psi-n-coalescents (also called Dirac-n-coalescents) with parameter $\Psi \in [0, 1]$ and Beta$(2-\alpha,\alpha)$-n-coalescent with $\alpha \in [1, 2]$. The parameters α or Ψ regulate the strength and frequency of multiple mergers, the smaller α or the larger the Ψ, the more frequent coalescence events are multiple mergers and the larger they get. Both MMCs incorporate Kingman's n-coalescent as a special case ($\alpha = 2$ or $\Psi = 0$).

Both MMC coalescent models can be defined for demographic variation in the underlying reproduction model that stays of the same order, *i.e.* where the populations size ratio $\nu_t = N_t/N_0$ of the population size at time t in the past (in coalescent time units) is positive and finite (for large population sizes N). The coalescent merges any k of b (ancestral) lineages present at a time t with rate

$$\lambda_{n,k}(t) = \nu(t)^{-\eta} \int_0^1 x^{k-2}(1-x)^{n-k}\Lambda(dx), \tag{1}$$

where

- Λ is either the Dirac distribution (point mass) in Ψ (Psi-coalescent) or the Beta$(2-\alpha,\alpha)$ distribution (Beta coalescent).

- η is the scaling factor reflecting how many time steps from the discrete reproduction model form one unit of coalescent time. It is the power of N of the scaling factor, e.g. $\eta = 2$ for the Moran model.

We focused on exponentially growing populations, *i.e.* a population size ratio $\nu(t) = \exp(-gt)$ for growth rate $g \geq 0$. As underlying reproduction models, we use modified Moran models [HM13, EW06, MHAJ18]. At each time step, in a population of size N, a single random individual has $U + G$ offspring(s) while $N - U$ random individuals have 1 offspring (leaving $U - 1$ individuals without offspring). As a consequence, the population grows from N to $N + G$ individuals and G is chosen to fit the desired growth rate (for low g, it is 0 for most time steps and 1 for the others).

In a Moran model, we have $U = 2$ and $G = 0$ when the population size is constant. However, for both MMCs, U is set to different values. In both cases, the mean of U does not grow indefinitely as N grows (for all parameters α and Ψ) but its variance does (for $\alpha \neq 2$ and $\Psi \neq 0$).

- In the Psi-n-coalescent (essentially [EW06, MHAJ18]), we have $U = 2$, except when a sweepstake event occurs with a small probability of order $N^{-\gamma}$ $(1 < \gamma < 2)$; in this case, $U = \lfloor N\Psi \rfloor$. In the coalescent time scale, one unit of time corresponds to an order of N^γ time steps; this is the expected time to an event of sweepstake. We choose $\gamma = 1.5$ for $\Psi > 0$, so $\eta = 1.5$. For $\Psi = 0$, we use a standard Moran model with $U = 2$ in every time step ($\eta = 2$).

- In the Beta-n-coalescent [HM13, Fre20], U has distribution $P(U = j) = \lambda_N^{-1} \binom{N}{j} \frac{B(j-\alpha, \alpha+N-j)}{B(2-\alpha,\alpha)}$, where B is the Beta function and λ_N is the normalizing constant. Consequently, although U takes different values at each time step, it is usually small, with a mean of at most $\frac{\alpha}{\alpha-1}$. Rarely, U is large. See Section A.1 for more details. On the coalescent time scale, one unit of time corresponds to an order of N^α time steps. Note that $\alpha = 2$ is the classical Moran model and thus leads to the Kingman's coalescent.

For statistical inference, we treat the site frequency spectrum of s mutations as s independent multinomial draws from the expected SFS (see [Nie00] and [EBBF15, Eq. 11] [MHAJ18, Eq. 14]). This computes an approximate composite likelihood function of the data for any combination of growth rate (g) and coalescent parameter (α or ψ). However, to include the effect of confusing the ancestral allele with the derived allele, we introduced another parameter e. On average, a confusion probability of e lets a fraction e of the derived allele carried by i sequences to be falsely seen as appearing in $n - i$ sequences. Additionally, as described in [Lap17, Section 4.2] or [BD03, p. 1620], as confusion stems from double-mutated sites, e also relates to the number of sites that cannot be polarized when compared with the outgroup as a third allele is observed in the outgroup. We account for these two effects of e by swapping a fraction e of the variants at frequency i/n to $1 - i/n$ and we assume a Jukes and Cantor substitution model [JC+69] to predict for $s_{\neq}$ the number of non-polarizable variants. This leads to a slight variant of [MHAJ18, Eq. 14]. For any coalescent model with a specific set of coalescent, exponential growth and confusion parameters, the pseudolikelihood is:

$$PsL(s_1, \ldots, s_{n-1}, s, s_{\neq}) = \tag{2}$$
$$\frac{s!}{s_1! \cdots s_{n-1}!} \prod_{i=1}^{n-1} \left(\frac{\mathbb{E}[T_i](1-e) + \mathbb{E}[T_{k-i}]e}{\mathbb{E}[T_{tot}]} \right)^{s_i} \underbrace{\binom{s + s_{\neq}}{s_{\neq}} \left(\frac{2e}{1+2e} \right)^{s_{\neq}} \left(\frac{1}{1+2e} \right)^{s}}_{\text{from non-polarizable variants}},$$

where $s_1, \ldots, s_{n-1}$ is the observed site frequency spectrum (so we observe s_i sites with derived allele frequency i/n), $s = \sum_i s_i$ is the total number of polarizable polymorphic sites and $s_{\neq}$ is the number of non-polarizable sites. $E[T_i]$ is the expected sum of lengths of branches that support i leaves in the genealogy and $E[T_{tot}]$ is the sum of all branch lengths. For $e = 0$, we set the term estimated from non-polarizable variants to 1. See Appendix A.3 for details on the derivation.

2.2 Statistical inference

To find the best-fitting parameters, we conduct a grid-search for the highest pseudolikelihood. The expected branch lengths $E[T_i]$ in Eq. (2) are computed as in [MHAJ18], using the approach from [SKS16]. We use the following grids with equidistant steps

Beta: $\alpha \in [1, 2]$ in steps of 0.05, $g \in [0, 25]$ in steps of 0.05, $e \in [0, 0.15]$ in steps of 0.01.

Psi: $\Psi \in [0, 1]$ in steps of 0.05, g, e as for Beta above, complemented with $\Psi \in [0, 0.2]$ in steps of 0.01 (further expanding $g \in [0, 30]$ by steps of 0.05 and $e \in [0, 0.2]$ by steps of 0.01) when Ψ was estimated to be close to 0.

To perform model selection between the three coalescent models, we computed the two following log Bayes factors:

$$BF_1 = \max(\log \max_{\alpha,g,e} PsL, \log \max_{\Psi,g,e} PsL) - \log \max_{\alpha=2,g,e} PsL, \quad BF_2 = \log \max_{\alpha,g,e} PsL - \log \max_{\Psi,g,e} PsL \qquad (3)$$

from the maximum pseudolikelihoods computed for the three models. We inferred a MMC genealogy when $BF_1 > \log(10)$ and further chose a Beta coalescent or a Psi-coalescent when $BF_2 > \log(10)$ or $BF_2 < -\log(10)$ respectively.
For the best fitting parameter combinations either over the full parameter space or restricted to the Kingman coalescent with growth and allele confusion (i.e. fixing $\alpha = 2$ or $\Psi = 0$), we assess the goodness-of-fit of the observed data. First, we graphically compare the observed SFS with the expected SFS, approximated as $(\frac{E[T_1]}{E[T_{tot}]}, \ldots, \frac{E[T_{n-1}]}{E[T_{tot}]})$. Second, we quantify the (lack of) fit of the data by Cramér's V, a goodness-of-fit measure which handles different sample sizes and different number of polymorphic sites. See Section A.4 for details.

2.3 Data

We have collected 45 genome-wide SFS that are described in Tables 1 and 7. The collected SFS come from public data sets or private communications. Supplementary files 1,2 show the shapes of the SFS.

3 Results

3.1 Statistical performance

Using simulations, we first assess the power of the method to retrieve the correct model and then its power to estimate the parameters. See Section A.5 for details of the setup.

The model selection approach based on Bayes factors computed from Eq. (2) identifies the correct multiple merger model in most cases if multiple mergers are not very small and infrequent, which happens for $\alpha \approx 2$ or $\Psi \approx 0$. Not surprisingly, larger sample sizes lead to smaller errors. For all sample sizes investigated, even if a Kingman-based model is misidentified, α (Ψ) is estimated to be close to 2 (to 0) (Table 5). We would like to emphasize that our approach is conservative towards Kingman as we select a Kingman genealogy model if the Bayes factor does not distinctively point towards a MMC model.

We show that parameter estimation within both the Beta- and Psi-coalescent models works well for multi-locus data from non-small samples, especially for the allele confusion rate e and for the coalescent parameter, see Figure 1 and Figures 5–7. The growth rate is only estimated well for low growth rates, see Figs. 10, 13, 16, 19.

Table 1: Data sets, best fitting model and goodness-of-fit grade

Order	Species	n	#SNP	Model	Coal	g_{Model}	e_{Model}	Grade
Vertebrates	*Aptenodytes patagonicus*	20	1,278	Beta	1.25	1.5	0	B
	Athene cunicularia	40	11,268,203	Beta	1.8	1	0.03	B
	Corvus cornix	38	7,167,395	Beta	1.95	1	0	A
	Coturnix japonica	20	5,061,864	Beta	1.45	0.5	0.01	A
	Egretta garzetta	10	9,318,499	Beta	1.75	0	0.02	B
	Ficedula albicollis	24	14,697,230	Ψ	0.01	0.5	0.01	A
	Gorilla gorilla gorilla	54	9,878,547	Beta	1.9	0	0	B
	Homo sapiens	216	19,441,528	Beta	1.85	0	0	A
	Lepus granatensis	20	769	MMC/Ψ	0.12	0	0.03	C
	Nipponia nippon	16	1,140,694	KM	$\varnothing$	0	0.03	D
	Pan paniscus	26	6,293,657	Beta	1.85	1	0	B
	Pan troglodytes ellioti	20	10,009,190	Beta	1.7	0	0	A
	Parus major	54	14,174,305	Beta	1.75	0	0.01	A
	Parus caeruleus	20	866	MMC/Ψ	0.04	0	0.02	B
	Passer domesticus	16	18,501,992	KM	$\varnothing$	0	0	A
	Phylloscopus trochilus	24	33,401,127	KM	$\varnothing$	12.5	0	A
	Taeniopygia guttata	38	53,263,038	Beta	1.75	4	0	A
Invertebrates	*Armadillidium vulgare*	20	23,323	Beta	1.7	0	0.03	C
	Artemia franciscana	20	5,548	Beta	1.65	0	0.03	B
	Caenorhabditis brenneri	20	1,339	Beta	1.5	0	0.06	C
	Caenorhabditis elegans	574	165	KM	$\varnothing$	0	0.06	D
	Ciona intestinalis A	20	480	KM	$\varnothing$	0	0.11	B
	Ciona intestinalis B	20	1,883	Beta	1.65	0	0.02	B
	Culex pipiens	20	5,442	Beta	1.55	0.5	0.01	B
	Drosophila melanogaster	196	4,662,706	Beta	1.65	0.5	0.02	A
	Emys orbicularis	20	515	KM	$\varnothing$	0.5	0	C
	Halictus scabiosae	22	712	MMC/Ψ	0.04	0	0.01	B
	Melitaea cinxia	18	1,695	Beta	1.7	0.5	0.03	B
	Messor barbarus	20	9,651	KM	$\varnothing$	0.5	0	C
	Ostrea edulis	20	939	MMC/Ψ	0.04	0	0.02	B
	Physa acuta	18	4,286	Beta	1.5	0	0.02	B
	Sepia officinalis	18	1,740	KM	$\varnothing$	0	0.02	C
Plants	*Arabidopsis thaliana*	345	10,322,757	Beta	1.6	0	0.07	A
	Zea mays	66	520,310	Ψ	0.01	0	0	A
Bacteria	*Acinetobacter baumannii*	79	78,175	Beta	1.8	0	0.1	B
	Bacillus subtilis	38	105,523	Ψ	0.14	0	0.2	B
	Chlamydia trachomatis	59	9,924	KM	$\varnothing$	0	0.11	D
	Clostridium difficile	11	192	KM	$\varnothing$	15	0.15	D
	Escherichia coli	62	84,222	KM	$\varnothing$	0	0.06	B
	Helicobacter pylori	70	27,498	Ψ	0.01	1	0.2	B
	Klebsiella pneumoniae	156	203,601	KM	$\varnothing$	18.5	0.15	D
	Mycobacterium tubercolosis	33	7,142	Beta	1.05	2.5	0	C
	Pseudomonas aeruginosa	86	90,258	Ψ	0.06	3	0.2	B
	Staphylococcus aureus	152	30,052	Ψ	0.01	1	0.2	B
	Streptococcus pneumoniae	32	49,917	Beta	1.5	0	0.08	C

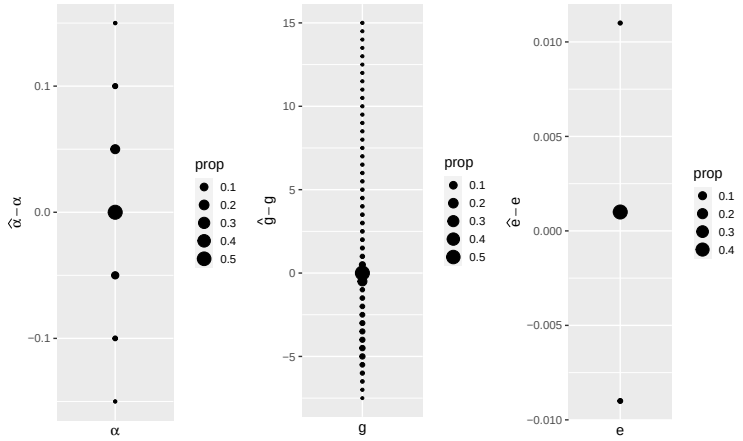


Figure 1: Error for estimating parameters for Beta coalescents with exponential growth and allele confusion) across the parameter grid for (α, g, e). Sample size $n = 100$, 50 independent loci w. 100 mutations on average. 500 simulations per parameter triplet.

3.2 Data analysis

MMC fit better Using the Bayes Factor criterion, we selected the best fitting model for each real SFS (Table 1). We found that a large majority (73%) of the SFS has a better fit with MMC than with Kingman. Globally, the best model selected is more often a Beta-coalescent (51%) than a Kingman (27%) or Psi-coalescent (13%). In a few cases, it is not possible to determine which one is the best fitting MMC (9%).

Goodness of fit. In agreement with the Cramér's V values of the fits and visual inspections (supplementary files 1,2), we designed grade categories: from 'very accurate' fit to 'very poor' fit, as following: A: $V \in [0 : 0.033[$, B: $V \in [0.033 : 0.066[$, C: $V \in [0.066 : 0.1[$ and D: $V \in [0.1 : \infty[$. Importantly, the MMC models fit well to 71% of data sets: 32/45 SFS have grades A or B on Table 3. This demonstrates that not only MMC are better choice on statistical ground but also that they appear as good candidate to explain patterns of diversity for a large majority of species.

The amount of multiple mergers greatly varies 68% (31/45) SFS have $\hat{\alpha} < 1.9$, which indicates that the multiple mergers are not only statistically sounds but that "something" important is not capture by the SNM (Table 6, α estimates of all data sets including best fitting to Kingman). While estimates of α cover the full range $[1, 2]$ and are evenly dispersed across the tree of life (Fig 2), they are skewed towards higher values. Similarly estimates of Ψ are concentrated in $[0, 0.15]$ and are skewed towards 0 and evenly distributed in the tree (Fig 8).

Great differences in g, no differences in e For each SFS, parameters inferred under a Kingman model are highly correlated with parameters inferred under a MMC model (Table 6). The growth parameters are often higher in the Kingman than in the MMC although it depends on the MMC parameter value (Fig 3a). The allele confusion parameters are almost identical between the Kingman model and the MMC (Fig 3b). Thus, choosing a Kingman model instead of an MMC will result in overestimating g, but not e.

Both MMC models points to similar results As expected, the parameters of the two MMC models inferred are highly correlated. The multiple merger parameters α of the Beta-coalescent and Ψ of the

sample size	true model	grid	Fraction model inferred as			
			Kingman	Beta	Psi	MMC
$n = 20$	$\alpha = 2$	yes	0.77	0.22	0.00	0.00
	$\alpha = 1.9$	yes	0.37	0.58	0.02	0.02
	$\alpha = 1.8$	yes	0.06	0.79	0.09	0.08
	$\alpha = 1.625$	no	0	0.82	0.09	0.08
	$\alpha = 1.025$	no		0.99	0.01	0
	$\alpha = 1$	yes		0.99	0	0
	$\Psi = 0.025$	no	0.14	0.59	0.15	0.12
	$\Psi = 0.05$	yes	0.01	0.17	0.70	0.11
	$\Psi = 0.075$	no		0.12	0.82	0.06
	$\Psi = 0.1$	yes		0.04	0.94	0.02
$n = 100$	$\alpha = 2$	yes	0.87	0.13		
	$\alpha = 1.9$	yes	0.12	0.88		
	$\alpha = 1.8$	yes		1		
	$\alpha = 1.625$	no		1		
	$\alpha = 1.025$	no		1		
	$\alpha = 1$	yes		1		
	$\Psi = 0.025$	no		0.92	0.06	0.02
	$\Psi = 0.05$	yes			1	
	$\Psi = 0.075$	no			1	
	$\Psi = 0.1$	yes			1	

Table 2: Model selection via two-step Bayes factor criterion. Based on 2,000 simulations for each true model assuming 100 loci with 50 observed mutations. For each simulation, the coalescent parameter is fixed and the growth parameter g and the allele confusion rate e are randomly chosen ($g \in [0, 11.25]$, $e \in [0, 0.1]$). The column grid shows whether the parameters used for simulation were included in the inference grid. For details on both simulations and inference parameters see Section A.5. Fractions are rounded to two digits.

Psi-coalescent are negatively correlated (Fig 4a, Spearman correlation: $\rho = -0.73$). The growth and misidentification estimated parameters are higly positively correlated (Spearman correlations $\rho = 0.74$ and $\rho = 0.96$). The case of *Clostridium difficile* is a notable exception. The best model inferred is the Kingman, consistent with $\hat{\Psi} = 0$ inferred for the Psi-coalescent, but for the Beta-coalescent $\hat{\alpha} = 1$, the strongest MMC component, is estimated. However, this discrepancy is likely due to statistical noise: the data set is very small (192 mutations in a sample size $n = 11$) and the species has a very low recombination rate.

4 Discussion

Chosen multiple-merger models, alternatives and limitations We chose two commonly used haploid multiple-merger models, the Beta and the Psi coalescents, which are often associated with sweepstake reproduction [EW06, Sch03] but also with Darwinian positive selection. Note that a selective sweep indeed corresponds to a "sweepstake reproduction" event but only for the genomic region around the selected locus. Beta n-coalescent with $\alpha = 1$ is the Bolthausen-Sznitman n-coalescent and emerges in a variety of models with rapid selection. Beta-coalescent also arises for loci in vicinity of a selected locus for the case of genetic draft [BD13, NH13, DWF13, BBS13, Sch17]. Similarly, Psi n-coalescents have been successfully used as

Model/Grade	A	B	C	D	Total
Kingman	2	2	3	5	12
Beta	8	11	4		23
Psi	2	4			6
MMC		3	1		4
Total	12	20	8	5	45

Table 3: Distribution of goodness-of-fit grades of the best-fitting models for the 45 collected SFS. Calculated from Cramér's V, A: $V \in [0 : 0.033[$, B: $V \in [0.033 : 0.066[$, C: $V \in [0.066 : 0.1[$ and D: $V \in [0.1 : \infty[$.

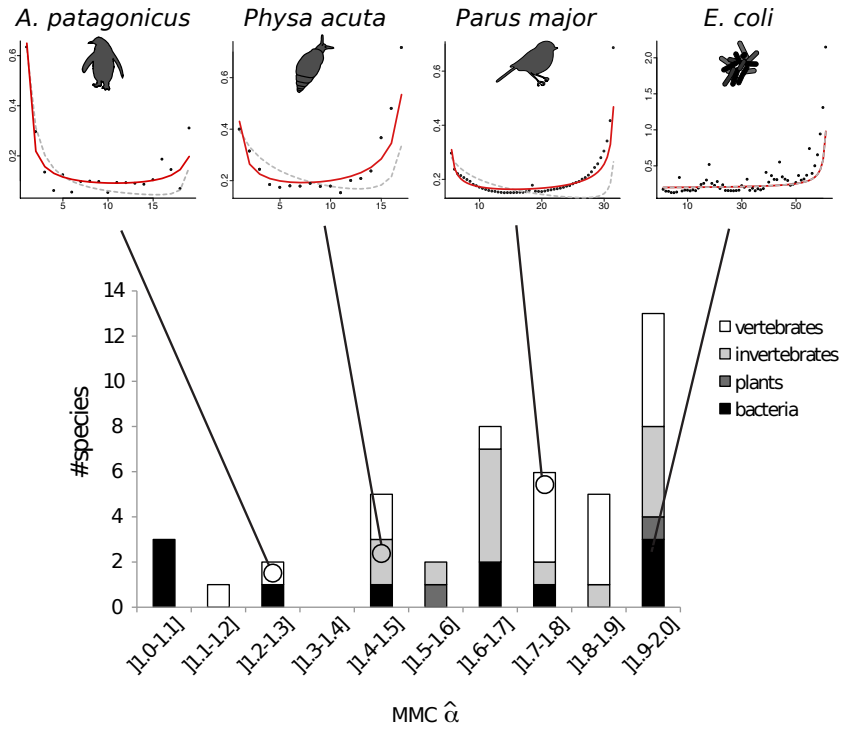
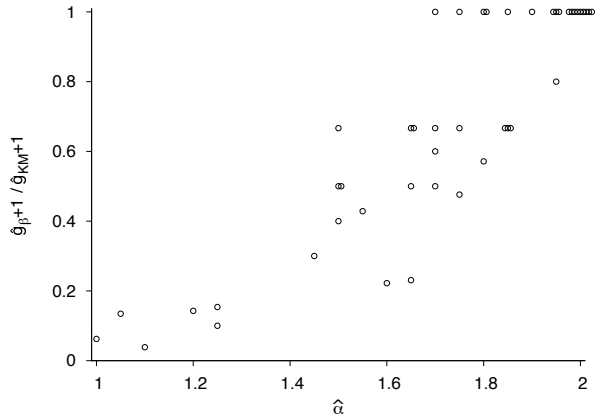
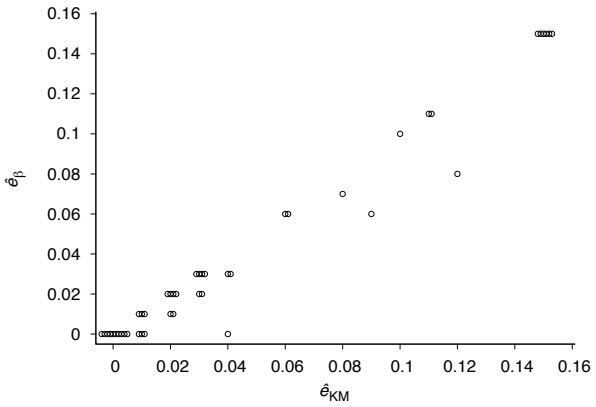


Figure 2: Distribution of α in function of the order of the species. The four top panels represent transformed ϕ-SFS for four species from different order: two vertebrates *Aptenodytes patagonicus* (left) and *Parus major* (center right) an invertebrate *Physa acuta* (center left), and a bacteria *Escherichia coli* (right). Black dots are the observed values, grey dotted lines are the best fits under the Kingman's coalescent model and red lines are the best fits under a Beta-coalescent model.

proxy models for the detecting regions under positive selection [HJ20]. Beta coalescents have also recently been linked to range expansions [BHK21]. While Beta and Psi coalescents models are linked to several biological properties potentially present in a considerable number of species, these are not the only MMCs used to model biological populations. For instance, in the modified Moran models presented above one can let the Ψ to be random, which leads to another more general class of MMC that also belong to the family of Λ-coalescents [HM13], that are also good candidate for modeling sweepstake reproduction. Other alternative models exist mimicking more closely recurrent selective sweeps [DS05] or appear as variants of Psi and Beta coalescents but for diploid reproduction [BCEH16, BLS18, KB19].

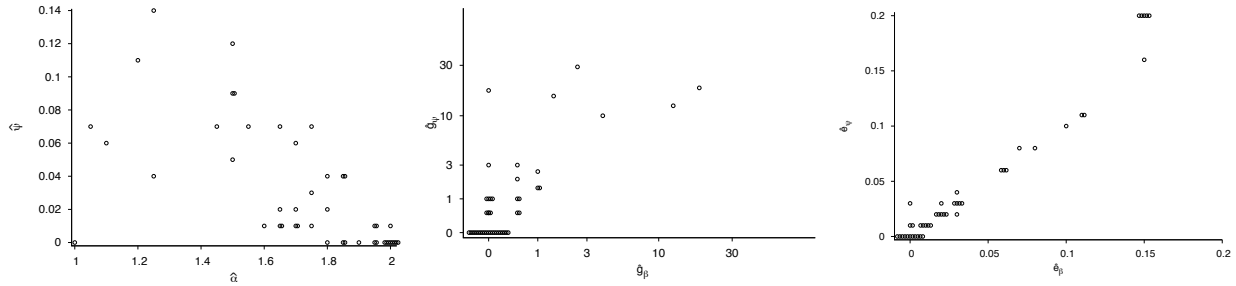


(a) Kingman model and Beta-coalescent model $\hat{g}+1$ ratio in function of $\hat{\alpha}$.



(b) Kingman model $\hat{e}$ in function of the corresponding Beta-coalescent allele confusion parameter.

Figure 3: Comparison of parameters between Kingman model and Beta-coalescent model.



(a) α/Ψ (b) g (c) e

Figure 4: Comparison of Beta-coalescent (x-axis) and Psi-coalescent (y-axis) parameters inferred from each SFS.

We settled for two simple classes of coalescent processes since these already interpolate between the two extreme tree shapes that are a purely bifurcating Kingman tree ($\Psi = 0$ or $\alpha = 2$) and a star-shaped tree ($\Psi = 1$ or $\alpha = 0$). Alternative multiple merger models with a strong enough multiple merger coalescent could potentially be be (mis)identified as Beta- or Psi-coalescents and not Kingman's coalescent, as shown for haploid and diploid versions of Beta coalescents in [FSJ21]. Our method thus should be still able to detect multiple merger signals even if caused by processes that lead to different multiple merger coalescents. Generally, our results show that even a simple extension of the standard model tuned by a single and interpretable parameter leads to considerably better fit of observed genetic diversity, pointing to a need to revise our reference null model in molecular evolution (that barely fit data) by expanding the SNM to MMCs. Nevertheless, assessing further which MMC models are best fitting for biological populations is informative, as discussed in [MGF20]. Our inference approach is based on computing $E(T_i)$ from Eq. (2) via the method from [SKS16], so it can easily be extended to incorporate most multiple merger models (any Λ- or Ξ-coalescent) and demographies by replacing the Markov transition rate matrix of the coalescent and the population size profile ν.

For the assessment of the quality of our inference method, we used a simplified approach where unlinked loci are assumed to be independent. This is not always true for MMC models (see [BBE13] and Section A.7), especially for Psi-coalescents caused by strong sweepstake reproduction events with Ψ well above 0. Thus,

the real error rates of our techniques may be higher than anticipated by our simulation study. However, this increase is balanced out for data sets featuring considerably more SNPs and/or bigger sample sizes than our simulations. Anyway, given that our Ψ estimates tend to be mostly close to 0, our data analysis should not be strongly affected by this assumption.

Due to our reliance on the expected SFS entries, that are averages over the tree space, our inference method (and also our goodness-of-fit assessment) should perform worse (given identical sample sizes and mutation counts) when used on species with small genomes and low recombination rates, a tendency we see for the goodness-of-fit for bacterial data sets.

Alternative processes leading to U-shaped SFS, further confounding factors On top of allele confusion and MMC genealogies as sources of U-shaped allele frequency spectra, we also made (cursory) checks for the potential influence of structure (e.g. gene flow or admixture) and biased gene conversion. We assessed population structure for 12 of the 17 data sets [*results for the other data sets will follow soon (analysis by F. Freund)*] that fitted well to a model with a meaningful MMC component (category A, B in Table 1, inferred MMC genealogy with $\alpha \leq 1.8$ or $\Psi \geq 0.04$, sample size ≥ 20). For this, we applied two clustering algorithms, PCA and k-means (see Section A.8 for details).

Overall, we assessed that population structure was visible in our checks for three of the data sets, see Table 8. However, a simulation study [KB19] showed that population structure causes a different imprint on the multi-locus SFS than MMC with exponential growth, so U-shaped SFS fitting well to our chosen models are likely at least not strongly affected by admixture or gene flow, which may indicate that the structure observed is relatively weak. Moreover, for data sets with small genomes and low recombination rate, the structure detected does not necessarily equate to population structure, but may also be reflecting selection and/or properties leading to MMC genealogies.

To check for the effect of biased gene conversion, we compared the SFS of unbiased mutations with the SFS of all mutations (details in A.6, the unbiased SFS are added in supplementary files 1,2). Many SFS were only slightly changed, others were changed but keep their U-shape, but 6 species (*A. cunicularia*, *F. albicollis*, *P. maior*, *E. garzetta*, *O. edulis*, *P. troglodytes e.*, all but one vertebrates), lost their U-shape. The U-shape in these few cases could be an effect of biased gene conversion. Since these data sets include some that fit very well to MMC (e.g. *P. maior*, *P. troglodytes e.*), this raises the follow-up question of whether the distortion of the SFS by biased gene conversion has a similar genetic signature as a multiple merger genealogy.

Even if we exclude all species with structured genetic diversity and species with patterns of biased gene conversion, there are still several species (how many) for which the U-shape is captured well by a MMC model, e.g. for *P. caeruleus*, *A. franciscana* or *C. japonica*. We believe that at least for these cases, sweepstake reproduction, positive selection and/or range expansion should indeed be considered as potential underlying drivers of genetic diversity.

MMC and biological properties Despite that we only analysed a small number of species sampled non-uniformly in the tree of life, we see multiple merger signals all across the tree of life. Reassuringly, our analysis supports multiple merger genealogies for *Mycobacterium tuberculosis*, which was recently proposed in [MASSJ20] and [MGF20] (the non-optimal goodness-of-fit likely stems from a small and essentially non-recombining genome). The strongest multiple merger effects estimated within the class of Beta coalescents ($\alpha \leq 1.1$) were found in samples of bacterial pathogens with low or intermediate recombination rates (*M. tuberculosis* and *P. aeruginosa*). Such estimates close to $\alpha = 1$, corresponding to the Bolthausen-Sznitman coalescent, point towards rapid selection as the driver of the MMC signal for these species as strong and

ongoing selection pressure are expected for both species.

Our sampling does not include fish, where reproduction sweepstakes can lead to MMC genealogies (see introduction). However, we included SFS of two cryptic subspecies of *C. intestinalis* which has been linked to both sweepstake reproduction and rapid selection [ZDB+12]. Our analysis supports assuming a MMC genealogy for one subspecies.

We stress that links between MMC model parameters and biological properties are often not obvious. For example, while reproduction sweepstakes can lead to both Beta- and Psi-coalescents, there is no actual translation of the parameters α and Ψ to realistic offspring distributions. For instance the Eldon-Wakeley model hypothesizes that occasional individual contribute to a fraction Ψ of the next generation, that is not biologically realistic. Still, the coalescent approximations do fit well to data. We would like to emphasize that different reproduction model can result in the same model on the *coalescent time scale*. The large families of the Eldon-Wakeley model could well be accumulated over multiple generations instead of in a single one. This both calls for working on linking mathematical models to actual biological properties (as it has been done in [DWF13] or in [Sch17] for rapid selection) and also to not discard multiple merger genealogies just because the underlying mathematical reproduction model does not perfectly fit to a species at hand.

Conclusion We collected genomic data for 45 species all across the tree of life and show that a vast majority exhibit a U-shaped SFS. We thus develop a statistical approach to distinguish the genetic signatures of two potential sources of this U-shape: allele confusion and MMC genealogies, together with exponential population growth. Our results show that some U-shaped SFS are well-described by only allele confusion, but that the majority of SFS are better described when adding an MMC component. While the U-shapes of some SFS fitting well to MMC genealogies may be explained by other sources (biased gene conversion for 6 species and population structure for 3 species), for half of the collected species, MMC genealogies are a good explanation for the observed pattern of genetic diversity. These do not only come from species already proposed as candidates for MMC (e.g. *P. caeruleus*). This study thus invites both closer inspection for the species at hand, but also suggests that MMC genealogies may appear in a wider range of species than previously reported (e.g. few marine species, pathogens, parasites or other organisms under rapid selection). More generally, our results show that a simple extension of the standard neutral model explains many more genome-wide SFS patterns. This echos the notion that the SNM, considered as *the* reference model in molecular evolution, routinely overlooks biological properties appearing in many species and thus that MMC could well be better candidate for being a biologically relevant reference model, especially when augmented with demography and population structure.

References

[ABAB+16] Carlos Alonso-Blanco, Jorge Andrade, Claude Becker, Felix Bemm, Joy Bergelson, Karsten M. Borgwardt, Jun Cao, Eunyoung Chae, Todd M. Dezwaan, Wei Ding, Joseph R. Ecker, Moises Exposito-Alonso, Ashley Farlow, Joffrey Fitz, Xiangchao Gan, Dominik G. Grimm, Angela M. Hancock, Stefan R. Henz, Svante Holm, Matthew Horton, Mike Jarsulic, Randall A. Kerstetter, Arthur Korte, Pamela Korte, Christa Lanz, Cheng-Ruei Lee, Dazhe Meng, Todd P. Michael, Richard Mott, Ni Wayan Muliyati, Thomas Nägele, Matthias Nagler, Viktoria Nizhynska, Magnus Nordborg, Polina Yu. Novikova, F. Xavier Picó, Alexander Platzer, Fernando A. Rabanal, Alex Rodriguez, Beth A. Rowan, Patrice A. Salomé, Karl J. Schmid, Robert J. Schmitz, Ümit Seren, Felice Gianluca Sperone, Mitchell Sudkamp, Hannes Svardal, Matt M. Tanzer, Donald Todd, Samuel L. Volchenboum, Congmao Wang, George Wang, Xi Wang, Wolfram Weckwerth,

Detlef Weigel, and Xuefeng Zhou. 1,135 genomes reveal the global pattern of polymorphism in arabidopsis thaliana. *Cell*, 166(2):481–491, 2016.

[Ach09] Guillaume Achaz. Frequency spectrum neutrality tests: one for all and all for one. *Genetics*, 183(1):249–58, Sep 2009.

[ÁH14] Einar Árnason and Katrín Halldórsdóttir. Nucleotide variation and balancing selection at the ckma gene in atlantic cod: Analysis with multiple merger coalescent models. *PeerJ PrePrints*, 2, 2014.

[BBE13] Matthias Birkner, Jochen Blath, and Bjarki Eldon. An ancestral recombination graph for diploid populations with skewed offspring distribution. *Genetics*, 193(1):255–290, 2013.

[BBM+09] Matthias Birkner, Jochen Blath, Martin Möhle, Matthias Steinrücken, and Johanna Tams. A modified lookdown construction for the Ξ-fleming-viot process with mutation and populations with recurrent bottlenecks. *Alea*, 6:25–61, 2009.

[BBS13] Julien Berestycki, Nathanaël Berestycki, and Jason Schweinsberg. The genealogy of branching brownian motion with absorption. *The Annals of Probability*, 41(2):527–618, 2013.

[BCEH16] Jochen Blath, Mathias Christensen Cronjäger, Bjarki Eldon, and Matthias Hammer. The site-frequency spectrum associated with Ξ-coalescents. *Theoretical Population Biology*, 110:36–50, 2016.

[BD03] Emmanuelle Baudry and Frantz Depaulis. Effect of misoriented sites on neutrality tests with outgroup. *Genetics*, 165(3):1619–1622, 2003.

[BD13] Éric Brunet and Bernard Derrida. Genealogies in simple models of evolution. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(01):P01006, 2013.

[BHK21] Gabriel Birzu, Oskar Hallatschek, and Kirill S Korolev. Genealogical structure changes as range expansions transition from pushed to pulled. *Proceedings of the National Academy of Sciences*, 118(34), 2021.

[BLS18] Matthias Birkner, Huili Liu, and Anja Sturm. Coalescent results for diploid exchangeable population models. *Electronic Journal of Probability*, 23, 2018.

[BMHR+17] Jean-Tristan Brandenburg, Tristan Mary-Huard, Guillem Rigaill, Sarah J. Hearne, Hélène Corti, Johann Joets, Clémentine Vitte, Alain Charcosset, Stéphane D. Nicolas, and Maud I. Tenaillon. Independent introductions and admixtures have contributed to adaptation of european maize and its american counterparts. *PLOS Genetics*, 13(3):e1006666, Mar 2017.

[BWSH01] Carlos D Bustamante, John Wakeley, Stanley Sawyer, and Daniel L Hartl. Directional selection and the site-frequency spectrum. *Genetics*, 159(4):1779–1788, 2001.

[CCSWB20] Fernando Cordero, Adrián González Casanova, Jason Schweinsberg, and Maite Wilke-Berenguer. Λ-coalescents arising in populations with dormancy. *arXiv preprint arXiv:2009.09418*, 2020.

[CGD18] Ivana Cvijović, Benjamin H Good, and Michael M Desai. The effect of strong purifying selection on genetic diversity. *Genetics*, 209(4):1235–1278, 2018.

[Con15] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[DK99] Peter Donnelly and Thomas G Kurtz. Particle representations for measure-valued population models. *The Annals of Probability*, 27(1):166–205, 1999.

[DS05] Rick Durrett and Jason Schweinsberg. A coalescent model for the effect of advantageous mutations on the genealogy of a population. *Stochastic Processes and their Applications*, 115(10):1628 – 1657, 2005.

[DWF13] Michael M. Desai, Aleksandra M. Walczak, and Daniel S. Fisher. Genetic diversity and the structure of genealogies in rapidly adapting populations. *Genetics*, 193(2):565–585, 2013.

[EBBF15] Bjarki Eldon, Matthias Birkner, Jochen Blath, and Fabian Freund. Can the site-frequency spectrum distinguish exponential population growth from multiple-merger coalescents? *Genetics*, 199(3):841–856, 2015.

[Eld20] Bjarki Eldon. Evolutionary genomics of high fecundity. *Annual Review of Genetics*, 54, 2020.

[EW06] Bjarki Eldon and John Wakeley. Coalescent processes when the distribution of offspring number among individuals is highly skewed. *Genetics*, 172(4):2621–2633, 2006.

[Ewe72] W J Ewens. The sampling theory of selectively neutral alleles. *Theor Popul Biol*, 3(1):87–112, Mar 1972.

[Fre20] Fabian Freund. Cannings models, population size changes and multiple-merger coalescents. *Journal of mathematical biology*, 80(5):1497–1521, 2020.

[FSJ21] Fabian Freund and Arno Siri-Jégousse. The impact of genetic diversity statistics on model selection between coalescents. *Computational Statistics & Data Analysis*, 156:107055, 2021.

[Fu95] Yun-Xin Fu. Statistical properties of segregating sites. *Theoretical population biology*, 48(2):172–197, 1995.

[Gil00] John H Gillespie. Genetic Drift in an Infinite Population: The Pseudohitchhiking Model. *Genetics*, 155(2):909–919, 06 2000.

[GJQP+19] Hugh G Gauch Jr, Sheng Qian, Hans-Peter Piepho, Linda Zhou, and Rui Chen. Consequences of pca graphs, snp codings, and pca variants for elucidating population structure. *PloS one*, 14(6):e0218306, 2019.

[GT94] Robert C Griffiths and Simon Tavare. Sampling theory for neutral alleles in a varying environment. *Philosophical transactions: biological sciences*, pages 403–410, 1994.

[HJ20] Rebecca B. Harris and Jeffrey D. Jensen. Considering genomic scans for selection as coalescent model choice. *Genome biology and evolution*, 12(6):871–877, 2020.

[HM13] Thierry Huillet and Martin Möhle. On the extended moran model and its relation to coalescents with multiple collisions. *Theoretical population biology*, 87:5–14, 2013.

[HSDB08] Emilia Huerta-Sanchez, Rick Durrett, and Carlos D Bustamante. Population genetics of polymorphism and divergence under fluctuating selection. *Genetics*, 178(1):325–337, 2008.

[Hud90] Richard R Hudson. Gene genealogies and the coalescent process. *Oxford surveys in evolutionary biology*, 7(1):44, 1990.

[IM02] Alex Iksanov and Martin Möhle. On the number of jumps of random walks with a barrier. *Advances in Applied Probability*, 40(01):206–228, 2002.

[JA11] Thibaut Jombart and Ismaïl Ahmed. adegenet 1.3-1: new tools for the analysis of genome-wide snp data. *Bioinformatics*, 27(21):3070–3071, 2011.

[JC+69] Thomas H Jukes, Charles R Cantor, et al. Evolution of protein molecules. *Mammalian protein metabolism*, 3:21–132, 1969.

[JPS+19] Jeffrey D. Jensen, Bret A. Payseur, Wolfgang Stephan, Charles F. Aquadro, Michael Lynch, Deborah Charlesworth, and Brian Charlesworth. The importance of the neutral theory in 1968 and 50 years on: A response to kern and hahn 2018. *Evolution*, 73(1):111–114, 2019.

[KB19] Jere Koskela and Maite Wilke Berenguer. Robust model selection between population growth and multiple merger coalescents. *Mathematical biosciences*, 311:1–12, 2019.

[KH18] Andrew D Kern and Matthew W Hahn. The Neutral Theory in Light of Natural Selection. *Molecular Biology and Evolution*, 35(6):1366–1371, 05 2018.

[Kim68] Motoo Kimura. Evolutionary rate at the molecular level. *Nature*, 217(5129):624–626, 1968.

[Kim83] Motoo Kimura. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, 1983.

[Kin82] J.F.C. Kingman. The coalescent. *Stochastic Processes and their Applications*, 13(3):235–248, Sep 1982.

[Kos18] Jere Koskela. Multi-locus data distinguishes between population growth and multiple merger coalescents. *Statistical applications in genetics and molecular biology*, 17(3), 2018.

[KVS+17] Mamoru Kato, Daniel A. Vasco, Ryuichi Sugino, Daichi Narushima, and Alexander Krasnitz. Sweepstake evolution revealed by population-genetic analysis of copy-number alterations in single genomes of breast cancer. *Royal Society Open Science*, 4(9), 2017.

[Lap17] Marguerite Lapierre. *Extensions du modèle standard neutre pertinentes pour l'analyse de la diversité génétique*. PhD thesis, Université Pierre et Marie Curie-Paris VI, 2017.

[LBL+16] Marguerite Lapierre, Camille Blin, Amaury Lambert, Guillaume Achaz, and Eduardo PC Rocha. The impact of selection, gene conversion, and biased sampling on the assessment of microbial demography. *Molecular biology and evolution*, 33(7):1711–1725, 2016.

[LCC+15] Justin B Lack, Charis M Cardeno, Marc W Crepeau, William Taylor, Russell B Corbett-Detig, Kristian A Stevens, Charles H Langley, and John E Pool. 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, 01 2015.

[LLA17] Marguerite Lapierre, Amaury Lambert, and Guillaume Achaz. Accuracy of demographic inferences from the site frequency spectrum: The case of the yoruba population. *Genetics*, 206(1):439–449, 05 2017.

[MASSJ20] Ana Y Morales-Arce, Susanna J Sabin, Anne C Stone, and Jeffrey D Jensen. The population genomics of within-host mycobacterium tuberculosis. *Heredity*, pages 1–9, 2020.

[MC05] Gilean AT McVean and Niall J Cardin. Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 360(1459):1387–1393, 2005.

[ME20] Nina Marchi and Laurent Excoffier. Gene flow as a simple cause for an excess of high-frequency-derived alleles. *Evolutionary applications*, 13(9):2254–2263, 2020.

[MGF20] Fabrizio Menardo, Sébastien Gagneux, and Fabian Freund. Multiple Merger Genealogies in Outbreaks of Mycobacterium tuberculosis. *Molecular Biology and Evolution*, 38(1):290–306, 07 2020.

[MHAJ18] Sebastian Matuszewski, Marcel E. Hildebrandt, Guillaume Achaz, and Jeffrey D. Jensen. Coalescent processes with skewed offspring distributions and non-equilibrium demography. *Genetics*, 208(1):323–338, 2018.

[Mon16] Valeria Montano. Coalescent inferences in conservation genetics: should the exception become the rule? *Biology letters*, 12(6):20160211, 2016.

[MS01] Martin Möhle and Serik Sagitov. A classification of coalescent processes for haploid exchangeable population models. *The Annals of Probability*, 29(4):1547–1562, 2001.

[MW06] Paul Marjoram and Jeff D Wall. Fast "coalescent" simulation. *BMC genetics*, 7(1):16, 2006.

[NH13] Richard A. Neher and Oskar Hallatschek. Genealogies of rapidly adapting populations. *Proc. Natl. Acad. Sci. USA*, 110(2):437–442, 2013.

[Nie00] R Nielsen. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics*, 154(2):931–942, 02 2000.

[NNY16] Hiro-Sato Niwa, Kazuya Nashida, and Takashi Yanagimoto. Reproductive skew in japanese sardine inferred from dna sequences. *ICES Journal of Marine Science*, 73(9):2181–2189, 2016.

[NT08] Nobukazu Nawa and Fumio Tajima. Simple method for analyzing the pattern of dna polymorphism and its application to snp data of human. *Genes & Genetic Systems*, 83(4):353–360, 2008.

[PATE18] Fanny Pouyet, Simon Aeschbacher, Alexandre Thiéry, and Laurent Excoffier. Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences. *Elife*, 7:e36317, 2018.

[Pit99] Jim Pitman. Coalescents with multiple collisions. *Annals of Probability*, 27(4):1870–1902, 1999.

[PMSK+13] Javier Prado-Martinez, Peter H. Sudmant, Jeffrey M. Kidd, Heng Li, Joanna L. Kelley, Belen Lorente-Galdos, Krishna R. Veeramah, August E. Woerner, Timothy D. O'Connor, Gabriel Santpere, Alexander Cagan, Christoph Theunert, Ferran Casals, Hafid Laayouni, Kasper Munch, Asger Hobolth, Anders E. Halager, Maika Malig, Jessica Hernandez-Rodriguez, Irene Hernando-Herraez, Kay Prüfer, Marc Pybus, Laurel Johnstone, Michael Lachmann, Can Alkan, Dorina Twigg, Natalia Petit, Carl Baker, Fereydoun Hormozdiari, Marcos Fernandez-Callejo,

Marc Dabad, Michael L. Wilson, Laurie Stevison, Cristina Camprubí, Tiago Carvalho, Aurora Ruiz-Herrera, Laura Vives, Marta Mele, Teresa Abello, Ivanela Kondova, Ronald E. Bontrop, Anne Pusey, Felix Lankester, John A. Kiyang, Richard A. Bergl, Elizabeth Lonsdorf, Simon Myers, Mario Ventura, Pascal Gagneux, David Comas, Hans Siegismund, Julie Blanc, Lidia Agueda-Calpena, Marta Gut, Lucinda Fulton, Sarah A. Tishkoff, James C. Mullikin, Richard K. Wilson, Ivo G. Gut, Mary Katherine Gonder, Oliver A. Ryder, Beatrice H. Hahn, Arcadi Navarro, Joshua M. Akey, Jaume Bertranpetit, David Reich, Thomas Mailund, Mikkel H. Schierup, Christina Hvilsom, Aida M. Andrés, Jeffrey D. Wall, Carlos D. Bustamante, Michael F. Hammer, Evan E. Eichler, and Tomas Marques-Bonet. Great ape genetic diversity and population history. *Nature*, 499(7459):471–475, 2013.

[RdSB+18] Olaya Rendueles, Jorge A. Moura de Sousa, Aude Bernheim, Marie Touchon, and Eduardo P. C. Rocha. Genetic exchanges are more frequent in bacteria encoding capsules. *PLOS Genetics*, 14(12):1–25, 12 2018.

[RGB+14] J. Romiguier, P. Gayral, M. Ballenghien, A. Bernard, V. Cahais, A. Chenuil, Y. Chiari, R. Dernat, L. Duret, N. Faivre, and et al. Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature*, 515(7526):261–263, Aug 2014.

[RNW+14] Christian Rödelsperger, Richard A Neher, Andreas M Weller, Gabi Eberhardt, Hanh Witte, Werner E Mayer, Christoph Dieterich, and Ralf J Sommer. Characterization of genetic diversity in the nematode pristionchus pacificus from population-scale resequencing data. *Genetics*, 196(4):1153–1165, 2014.

[RZL+18] Aurélien Richaud, Gaotian Zhang, Daehan Lee, Junho Lee, and Marie-Anne Félix. The local coexistence pattern of selfing genotypes in caenorhabditis elegans natural metapopulations. *Genetics*, 208(2):807–821, 2018.

[Sag99] Serik Sagitov. The general coalescent with asynchronous mergers of ancestral lines. *Journal of Applied Probability*, 36(4):1116–1125, 1999.

[Sch00] Jason Schweinsberg. Coalescents with simultaneous multiple collisions. *Electronic Journal of Probability*, 5:1–50, 2000.

[Sch03] Jason Schweinsberg. Coalescent processes obtained from supercritical Galton–Watson processes. *Stochastic Proc. Appl.*, 106(1):107–139, 2003.

[Sch17] Jason Schweinsberg. Rigorous results for a population model with selection ii: genealogy of the population. *Electronic Journal of Probability*, 22, 2017.

[SD05] Jason Schweinsberg and Rick Durrett. Random partitions approximating the coalescence of lineages during a selective sweep. *The Annals of Applied Probability*, 15(3):1591 – 1651, 2005.

[SH74] John Maynard Smith and John Haigh. The hitch-hiking effect of a favourable gene. *Genetical Research*, 23(1):23–35, 1974.

[SKS16] Jeffrey P. Spence, John A. Kamm, and Yun S. Song. The site frequency spectrum for general coalescents. *Genetics*, 202(4):1549–1561, 2016.

[SW08] Ori Sargsyan and John Wakeley. A coalescent process with simultaneous multiple mergers for approximating the gene genealogies of many marine organisms. *Theoretical Population Biology*, 74(1):104–114, Aug 2008.

[TL14] Aurelien Tellier and Christophe Lemaire. Coalescence 2.0: a multiple branching of recent theoretical developments and their applications. *Molecular ecology*, 23(11):2637–2652, 2014.

[TV09] Jesse E Taylor and Amandine Véber. Coalescent processes in subdivided populations subject to recurrent mass extinctions. *Electron. J. Probab*, 14:242–288, 2009.

[Wal00] Jeffrey D. Wall. A Comparison of Estimators of the Population Recombination Rate. *Molecular Biology and Evolution*, 17(1):156–163, 01 2000.

[ZDB+12] Aibin Zhan, John A Darling, Dan G Bock, Anaïs Lacoursière-Roussel, Hugh J MacIsaac, and Melania E Cristescu. Complex genetic patterns in closely related colonizing invasive species. *Ecology and Evolution*, 2(7):1331–1346, 2012.

Appendix A Appendix

A.1 Reproduction models linked to MMC and time scalings

The coalescent approximations from the main text are indeed the coalescent limits for population size $N \to \infty$ (with changed time-scale) of genealogy trees in some reproduction model. We focus on Cannings models as reproduction models, which are discrete-generation models, usually with fixed population size, and exchangeable offspring numbers between individuals. This is a standard model choice, see e.g. [Sag99], also the modified Moran models present in the Methods are Cannings models. Different reproduction models can lead to the same coalescent limit, e.g. the Wright-Fisher and Moran model both lead to Kingman's coalescent. If the coalescent limit is identical for constant population size reproduction models (and the number of generations to form one coalescent time unit is of order N^η), we can describe the limit as in Eq. (1). Thus, adding population size changes can still lead to a difference in coalescent limit via changing the power η of the population size ratio ν. For instance, $\eta = 2$ for the standard Moran model but $\eta = 1$ for the Wright-Fisher model (Λ the point mass in 0 in both cases). In the case of exponential growth (on the coalescent time scale), we simply have that the factor influenced by η in Eq. (1) equals $\nu(t)^{-\eta} = \exp(\eta\rho t)$. This means that we can still interpret parameters assuming one reproduction model (model 1) leading to the coalescent (with scale parameter $\eta = x_1$) under the assumption of an alternative reproduction model (model 2 with scale parameter $\eta = x_2$) by simply rescaling the exponential growth parameter g from model 1 as $g' = g\frac{\eta_1}{\eta_2}$. For instance, a growth rate of $2g$ in the Wright-Fisher model corresponds to a growth rate of g in the Moran model.

For our two MMC models, this also means that we could analyse the models based on alternative reproduction models. For instance, we set $\gamma = 1.5$ for the discrete reproduction model leading to the Psi-coalescents, but we could also choose any other $1 < \gamma < 2$. For the Beta-coalescents, there is indeed a very appealing alternative reproduction model, Schweinsberg's model from [Sch03]. This alternative model assumes, for $1 \leq \alpha < 2$, that each individual at some generation independently produces a number offspring following a power law distribution with tail parameter α (infinite variance), and that then the next generation (the individuals surviving long enough to reproduce) is sampled from these offspring. In this model, one unit of coalescent time corresponds to an order of $N^{\alpha-1}$ generations. While this model is appealing since it really captures a high fecundity, high early life mortality scenario, for $\alpha = 1$, it has the surprising property that any population size change cannot be seen in the resulting Beta-n-coalescent (Bolthausen-Sznitman-n-coalescent). However, the Bolthausen-Sznitman-n-coalescent is also appearing as limit genealogy for many other discrete-time reproduction models. This is why we decided to see the Beta-n-coalescents as coming

from the modified Moran model presented in the Methods, since it allows measurable population growth for the Bolthausen-Sznitman coalescent. Moreover, as discussed above, if $\alpha > 1$, we can still interpret any growth rate g when seeing the Beta-coalescent as the genealogy model based on the modified Moran model as growth rate $g' = \left(1 + \frac{1}{\alpha-1}\right) g$ under Schweinsberg's model.

A.2 Properties of the reproduction model underlying the Beta-coalescent

The modified Moran model with distribution given on p.4 leading to the Beta$(2-\alpha,\alpha)$-coalescent was introduced in [HM13] and the properties of U have been additionally analysed in [IM02]. U, or more precisely U_N since it depends on N, is distributed as the number of lineages merged at the first merger in a Beta$(2-\alpha,\alpha)$-coalescent starting with sample size N. Since when increasing the sample size, the first merger can only include more lineages, $U_N \leq U_M$ holds for $M \geq N$ (we can assume coalescents with increasing sample sizes just add branches to the tree from smaller sample sizes, see [Pit99]). This then also holds for the expected values, so $E(U_N) \leq E(U_M) \leq E(U_\infty) = \frac{\alpha}{\alpha-1}$, where U_∞ is the limit of U for $N \to \infty$. See [HM13, p.9] for the existence of the limit, whose properties including mean are described on the cited page combined with [IM02, p.226], including its infinite variance. See Table 4 for some properties of U_N for different N and α, computed from the definition of U_N and the listed properties of its limit.

N	α	$E(U_N)$	$E(U_\infty)$	$\sqrt{Var(U_N)}$	$P(U \leq x_{min}) \geq 0.99$
5000	1.1	6.54	11.00	46.87	62
10000	1.1	6.83	11.00	64.24	62
25000	1.1	7.20	11.00	97.26	62
5000	1.5	2.96	3.00	9.39	15
10000	1.5	2.97	3.00	11.27	15
25000	1.5	2.98	3.00	14.29	15
5000	1.9	2.11	2.11	1.39	4
10000	1.9	2.11	2.11	1.50	4
25000	1.9	2.11	2.11	1.64	4

Table 4: Properties of U_N for the modified Moran model underlying the Beta-coalescents. x_{min}: Minimal integer x_min so that $P(U \leq x_{min}) \geq 0.99$ is satisfied.

A.3 Mathematical derivation of the pseudolikelihood function Eq. (2)

We follow the derivation from [EBBF15, Eq. 11]. We want to compute the likelihood of seeing the observed site frequency spectrum $s_1, \ldots, s_{n-1}$ under a given coalescent (here a Beta-n-coalescent or a Psi-n-coalescent with exponential growth, but the derivation works for any coalescent model). Let $s = \sum_{i=1}^{n-1} s_i$ be the number of observed segregating sites. We assume the fixed-s approach, e.g. we assume that the distribution of the SFS is given by placing s mutations at random on the genealogical tree. Under the fixed-s assumption, the probability of observing the SFS is given by the multinomial distribution

$$P(SFS = (s_1, \ldots, s_{n-1})) = \mathbb{E}\left[\frac{s!}{s_1! \cdots s_{n-1}!} \prod_{i=1}^{n-1} \left(\frac{T_i}{T_{tot}}\right)^{s_i}\right], \tag{4}$$

since a segregating site has mutant allele frequency i if it lands on a branch that supports i leaves (T_i is the sum of lengths of branches supporting i leaves, $T_{tot} = \sum_{i=1}^{n-1} T_i$ is the total length of the genealogy).

Under further assumptions of independence of the different fractions $\frac{T_i}{T_{tot}}$ of the total branch length and approximating $E\left(\frac{T_i}{T_{tot}}\right) \approx \frac{E(T_i)}{E(T_{tot})}$, we have a further approximation

$$P(SFS = (s_1, \ldots, s_{n-1})) = \frac{s!}{s_1! \cdots s_{n-1}!} \prod_{i=1}^{n-1} \left(\frac{E(T_i)}{E(T_{tot})}\right)^{s_i}, \tag{5}$$

Now, what happens if we add a chance, the confusion probability, of e that each allele's state as ancestral or derived may be switched? Eq. (5) constitutes a multinomial distribution, which can be interpreted as throwing s balls into compartments 1,...,$n-1$, where compartment i is hit with probability $\frac{E(T_i)}{E(T_{tot})}$. Confusing the allele in this interpretation means that a ball that originally lands in compartment i is placed in in compartment $n-i$ instead. If this happens with probability e, a ball consequently lands in compartment i with probability $(1-e)\frac{E(T_i)}{E(T_{tot})} + e\frac{E(T_{n-i})}{E(T_{tot})}$. So the probability to observe a specific SFS when ancestral and derived types can be confused is

$$P(SFS{=}(s_i)_{i=1}^{n-1}) = \frac{s!}{s_1! \cdots s_{n-1}!} \prod_{i=1}^{n-1} \left(\frac{(1-e)E(T_i) + eE(T_{n-i})}{E(T_{tot})}\right)^{s_i}, \tag{6}$$

where $(s_i)_{i=1}^{n-1} = (s_1, \ldots, s_{n-1})$ is the observed SFS. This is again a multinomial distribution.
Simulations showed that inferring parameters via a pseudo-likelihood approach based on Eq. 6 tends to over-estimate e. To counteract this, we couple this equation with an alternative estimation of e by using polymorphic sites discarded in the process of polarizing the SFS due to having a third allele in the outgroup. As described in [Lap17, Section 4.2] or [BD03, p. 1620], these sites carry information about e. Let $S_0 = S + S_{\neq}$ be the total number of biallelic SNPs in the sample, where S is the (random) number of sites where the outgroup does not show a third allele not observed in the sample and $S_{\neq}$ the number of sites where it does. Observe that s is the observed outcome of S, the total sum of the observed site frequency spectrum.
Consider a biallelic site in the sample. Assume you know the ancestral state of the common ancestor of sample and outgroup (and that the state in the outgroup is monomorphic). Further assume that the SNP is caused by a single mutation (we assume an infinite-sites model on the sample's genealogy, so within the sample this is met already). Assume thirdly a Jukes-Cantor setting: Any change from one base to another across a phylogenetic path has the same probability. We denote the probability of such a mutation on the path τ between sample and outgroup, i.e. between the MRCA of the sample and the outgroup, by p.
Then, a three-allelic SNP happens if there is a mutation on this path τ which causes a third allele. The probability of this equals $2p$, since there are two nucleotides not covered by the sample at this site.
If there are two alleles at the site found in sample and outgroup, we identify the ancestral allele as the allele of the outgroup. Now, confusing the ancestral with the derived allele happens if there is a mutation on τ with the same effect as the mutation causing the SNP in the sample, which means its probability is p, since the site in the outgroup has to mutate to a specific nucleotide from the ancestral state. While we assume an infinite-sites model for the site frequency spectrum, this would clearly be violated for this mutation affecting the outgroup. However, since the branch connecting to the outgroup should be considerably longer than the genealogy within the species, we think it is still a reasonable approximation that while the infinite sites model is modelling the genealogy well, it breaks down on the longer phylogenetic branches and allows sites to be hit twice there.
Following this, we can compute the probability $P(S_{\neq} = s_{\neq} | S_0 = s + s_{\neq})$ that we observe exactly $s_{\neq}$ sites which are biallelic within the sample but have a third allele for the outgroup. This is just binomial sampling from S_0 biallelic sites with success probability $2p$. We can also express this probability in terms of the misidentification probability e. Let $v \in SFS$ be the event that a biallelic site (variant) can be polarized

via outgroup (which means it has one of the two alleles of the sample also for the outgroup) and $mis(v)$ the event that the ancestral state of v is misidentified. The probability that a site of the SFS can be polarized is $1-p$. We have

$$e = P(mis(v)|v \in SFS) = \frac{P(mis(v), v \in SFS)}{P(v \in SFS)} = \frac{p}{1-2p},$$

and thus equivalently $p = \frac{e}{1+2e}$. This leads to

$$P(S_{\neq} = s_{\neq}|S_0 = s + s_{\neq}) = \binom{s+s_{\neq}}{s_{\neq}} \left(\frac{2e}{1+2e}\right)^{s_{\neq}} \left(\frac{1}{1+2e}\right)^{s}. \tag{7}$$

We now simply assume a composite likelihood, multiplying Eqs. (6) and (7). Conditional on observing $s + s_{\neq}$ segregating sites from which s can be polarized via outgroup and form the SFS, the pseudo-likelihood of observing a specific SFS is given by Eq. (2).

Remark A.1. *Eq.* (7) *shows that S given S_0 is binomially distributed with S_0 draws with success probability $1-2p = (1+2e)^{-1}$. Let $X(S_0)$ be a r.v. with this distribution. We will use this to simulate S_0 based on S and the misclassification probability e: The maximum likelihood estimate for the number of trials S_0 of the binomial r.v. $X(S_0) \sim Bin(S_0, (1+2e)^{-1})$, in the sense of maximising $P(X(S_0) = s)$, is $\hat{S_0} = \lfloor (1+2e)s \rfloor$, since*

$$\frac{P(X(S_0+1)=s)}{P(X(S_0)=s)} = \frac{S_0+1}{S_0+1-s}\frac{2e}{1+2e} \geq 1 \Leftrightarrow S_0 \leq s(1+2e)-1.$$

We will use this estimate to simulate a reasonable $s_{\neq}$. If we simulate a SFS with s mutations, and we flip each mutation in it with frequency e from class i to $n-i$, we then simulate $s_{\neq}$ as a binomial draw from $\hat{S_0} = \lfloor (1+2e)s \rfloor$ Bernoulli r.v.'s with success probability $\frac{2e}{1+2e}$. This is denoted as the $\hat{S_0}$ approach.

A.4 Cramér's V as a goodness-of-fit measure

Our assumptions leading to Equations (5), (6) can be interpreted that each variant observed for the SFS is sampled from a multinomial distribution from the 'true' allele frequency spectrum. In the following, we denote the multinomial approximation of the SFS entry frequencies, the 'true' spectrum, by $(p_1, \ldots, p_{n-1})$. Since assuming sampling from a multinomial distribution is also the statistical model behind the χ^2 goodness-of-fit test, we chose the effect size measure Cramér's V of this test, defined as

$$V = \sqrt{\sum_{i=1}^{n-1} \frac{(o_i - p_i)^2}{p_i(n-2)}},$$

to quantify the lack of goodness of fit (o_i is the observed frequency of mutations with frequency i/n among all mutations). This measure can be interpreted as a dimensionless version of the χ^2 test statistic, since the mutation counts do not enter, just the mutation frequencies and the additional factor $n-2$ corrects for unequal sample sizes.

A.5 Assessing estimation errors

A.5.1 Simulation and inference setup

As a rough approximation of a genome, we simulated 100 independent loci (ignoring the fine structure of weakly physically linked loci and long range LD, see Section A.7). This means that the genealogical trees of the loci are independent and follow the same tree distribution, e.g. realisations of a Beta coalescent with exponential growth with rate g and coalescent parameter α. The mutations on each tree are independent

of all trees (and mutations on other trees) and given by a Poisson process with rate $\frac{\theta}{2}$. We assumed two different sample sizes $n = 20$ and $n = 100$. For each locus, we set the mutation rate so that on average 50 mutations appear, i.e. we set $\theta = 100/E/(T_{tot})$ (generalized Watterson estimate), where T_{tot} is the sum of all branch lengths of the locus' genealogy. Mutations are interpreted under the infinite-sites model, resulting in simulated SNP sequences (ancestral vs. derived type). For each SNP, we then flip ancestral and derived allele with probability e. We simulate 500 SNP sequences as described above for each combination of coalescent parameter α or Ψ, growth rate g and confusion probability e from the following two sets (the first set has α, Ψ and g on the inference grid, the second uses off-grid values).

- Set 1: equidistant $\alpha \in \{1, 1.05, 1.1, \ldots, 2\}$ and $\Psi \in \{0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$
  Set 2: $\alpha \in \{1.025, 1.325\}$, $\Psi \in \{0.025, 0.075\}$
- Set1: $g \in \{0, 0.5, 1, 10\}$, Set 2: $g \in \{0.25, 2.25, 11.25\}$
- Set 1: $e \in \{0, 0.01, 0.05, 0.1\}$ (essentially on grid), Set 2: $e \in \{0, 0.015, 0.045, 0.095\}$

To infer via Eq. (2), we also need the total number of segregating sites $s + s_{\neq}$, adding the number of segregating positions not included in the SFS due to not being able to polarize them. For this, we use the $\hat{S}_0$ approach described in Remark A.1.
First, we estimate parameters using Eq. 2 using the same coalescent model (Beta or Psi) on equidistant grids with $\alpha \in \{1, 1.05, 1.1, \ldots, 2\}$ or $\Psi \in \{0, 0.5, 0.1, \ldots, 1\}$, $g \in \{0, 0.05, \ldots, 25\}$, $e \in \{0.001, 0.011, \ldots, 0.201\}$. For this, we only used the on-grid values (Set 1). Results are shown in Figures 1, 5 – 7,9 – 20.
Second, we assess the error of our model selection approach based on approximated Bayes factors. For this, we fixed different values of Ψ and α from Set 1 and Set 2 including $\alpha = 2$. We then picked 2,000 simulations at random from all parameter combinations with this fixed coalescent parameter (as described above) and performed model selection via Bayes factors as described in Section 2.2. The maximum was taken on the same equidistant grid as for the parameter estimation, for Ψ additionally combined with a further set with $\Psi \in \{0.01, 0.02\}$, ρ and e as above. The results are summarised in Tables 2 and 5.

A.5.2 Results

For inferring parameters under the Beta-coalescent or the Psi-coalescent, Figures 1, 5–7 show the error distribution of all three parameters for $n \in \{20, 100\}$ across all simulation parameter choices. While g cannot be estimated precisely in some cases, e, Ψ and, to a lesser degree, α, can generally be estimated rather well, especially if sample size $n = 100$.

These errors distribute over the different parameter settings as shown in Figures 9 – 20. Most notably, large errors when estimating growth rates only happen if the growth rate is also large. For Psi-coalescents, we see that choosing Ψ between grid points are still mostly captured by the adjacent Ψ grid points.

A.6 Correction for GC-bias

We use the approach from [PATE18] and consider the subset of SNPs corresponding to $A \leftrightarrow T$ and $G \leftrightarrow C$ conversions that are not affected by biased gene conversion. We added these neutralized SFS to the observed SFS and the predictions of the fitted models in supplementary files 1,2.

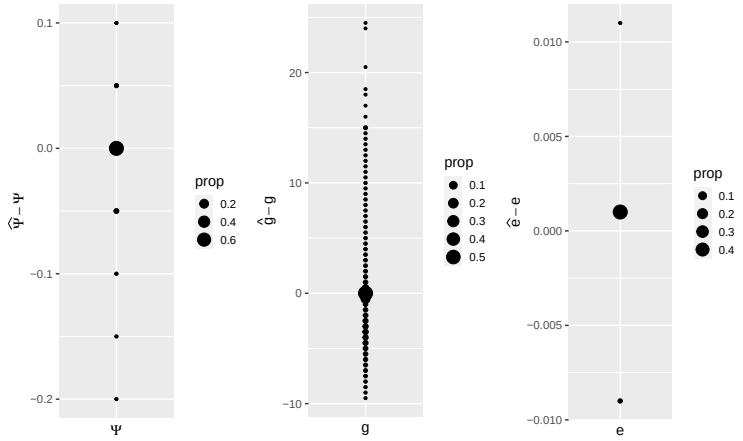


Figure 5: Error for estimating parameters for Psi-coalescents ($n = 100$, with growth and misclassification) across all simulation scenarios

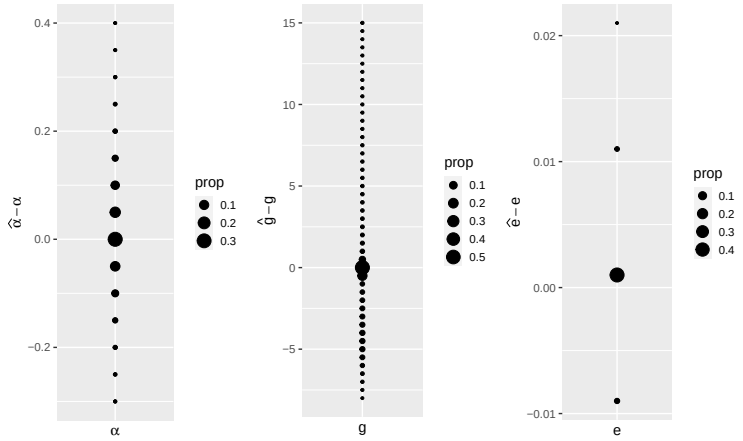


Figure 6: Error for estimating parameters for Beta-coalescents ($n = 20$, with growth and misclassification) across all simulation scenarios

A.7 Non-independence of unlinked loci under multiple merger genealogies

Here, we address the issue that physically unlinked loci in multiple merger genealogies still have dependent genetic diversity. For Λ-coalescents, which all our coalescent models are, the issue can be easily understood within the approximate multi-unlinked-locus model from the appendix of [Kos18]: Multiple mergers are resulting from large families appearing in a short amount of evolutionary time (see also a more thorough explanation in [MGF20]), so these families affect not only one, but all loci. Due to the model definition of MMC, each ancestral lineage can join one of such events with the same probability. Thus, if this probability is high, there will be a merger of similar size at each or nearly each locus in the genome, introducing a dependency between loci. The strength of this dependency should be proportional to the probability with which an ancestral lineage merges. This probability x is generated by a Poisson process whose rate is proportional to $x^{-2}\Lambda(dx)$, where Λ is the associated measure of the coalescent (a Beta distribution or the point mass in Ψ for our model classes). This probability is rather small for Beta coalescents, but high for high Ψ values, see also Figure 21.

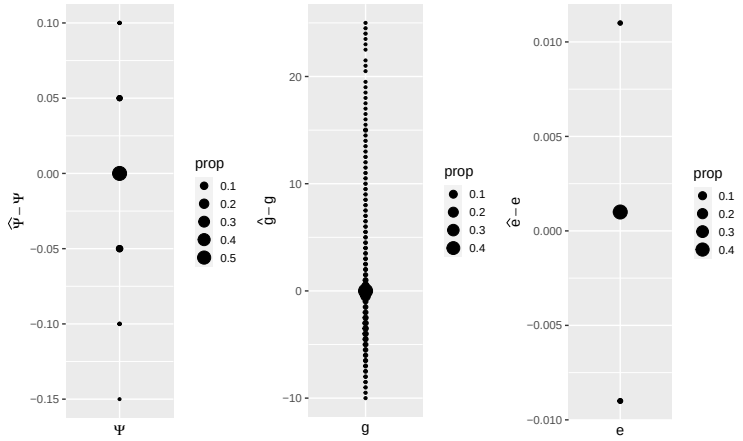


Figure 7: Error for estimating parameters for Psi-coalescents ($n = 20$, with growth and misclassification) across all simulation scenarios

sample size	$n = 20$	$n = 100$
$\alpha = 1.75$	0.01	
$\alpha = 1.8$	0.04	
$\alpha = 1.85$	0.22	0.04
$\alpha = 1.9$	0.37	0.53
$\alpha = 1.95$	0.34	0.43
$\Psi = 0.05$	0.02	

Table 5: Fractions of estimated parameters of model-misidentified coalescent simulations with $\alpha = 2$. If the two-step Bayes factor model inference recorded "MMC", the Beta parameter is reported.

A.8 Population structure scans

We performed two simple checks for population structure: PCA and `find.clusters` from the R package `adegenet` [JA11]. For PCA, we coded alleles as 0 and 1, imputed missing data as the mean allele at the site, to then perform a double-centered PCA: PCA, as implemented in `adegenet`, was performed on the SNP matrix after substracting row means and column means (and adding the overall mean), see [GJQP$^+$19, p.20]. The approach behind `find.clusters` is to first perform a standard PCA and then group individuals by running the k-means clustering algorithm on the principal component coordinates for different numbers of clusters. Based on the goodness-of-fit criterion BIC, we chose the 'optimal' k as the smallest value of k that is a local minimum (smaller and larger values around that k are higher, the elbow criterion). For large data sets of more than 1 million SNPs, we performed the analysis with a reduced data set by filtering down the number of SNPs by only retaining each xth SNP where $x = \frac{\#\ SNPs}{1000000}$, rounded to the lower integer. Results are shown in Supplementary file 3 and Table 8.

A.9 Psi-coalescent graphs

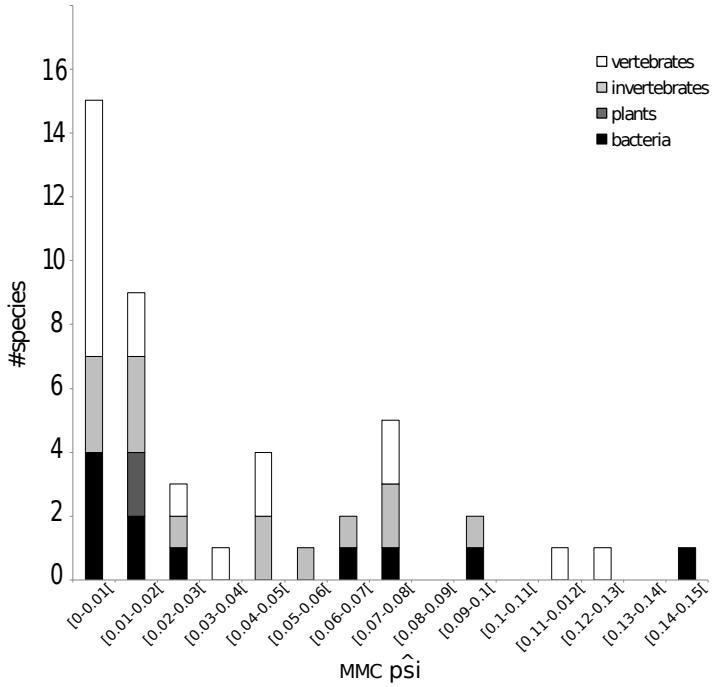


Figure 8: Distribution of psi parameter in function of the order of the species (white: vertebrates, light grey: invertebrates, dark grey: plants, black: bacteria).

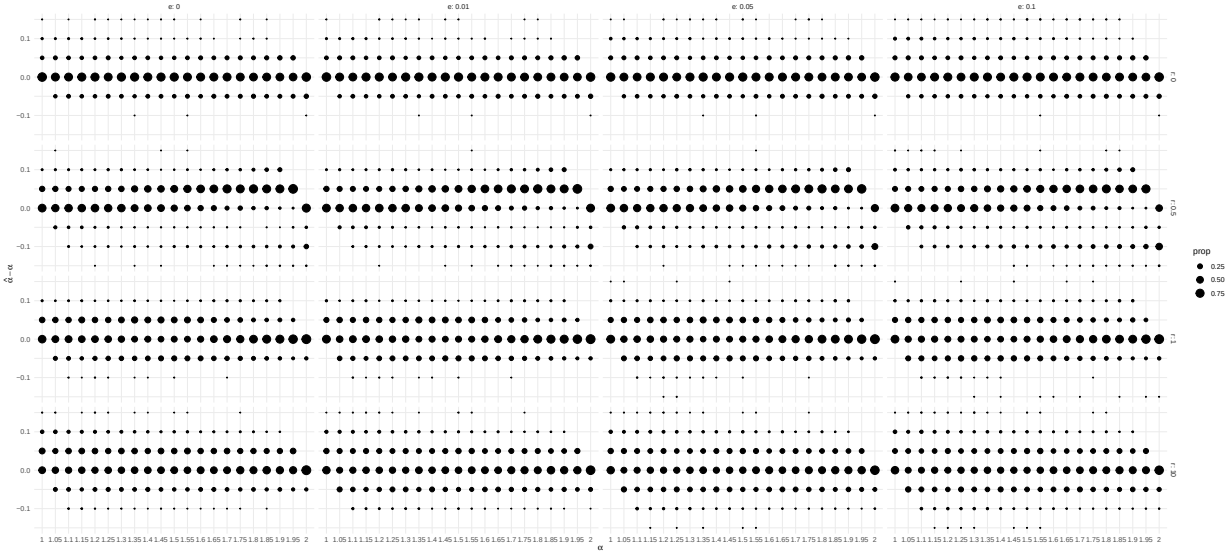


Figure 9: Error for estimating coalescent parameter α for Beta-coalescents with growth and misclassification ($n = 100$). Growth rate is denoted by g.

Table 6: Parameters Estimations

Species	g_{KM}	e_{KM}	V_{KM}	α	g_{Beta}	e_{Beta}	V_{Beta}	Ψ	g_{Ψ}	e_{Ψ}	V_{Ψ}	Model
Acinetobacter baumannii	0	0.1	0.064	1.8	0	0.1	0.059	0.02	0	0.1	0.061	Beta
Aptenodytes patagonicus	24	0.01	0.071	1.25	1.5	0	0.047	0.04	15.5	0.01	0.063	Beta
Arabidopsis thaliana	3.5	0.08	0.019	1.6	0	0.07	0.010	0.01	1	0.08	0.017	Beta
Armadillidium vulgare	0	0.03	0.089	1.7	0	0.03	0.069	0.06	0	0.02	0.083	Beta
Artemia franciscana	0.5	0.03	0.067	1.65	0	0.03	0.055	0.02	0.5	0.03	0.065	Beta
Athene cunicularia	2.5	0.03	0.040	1.8	1	0.03	0.037	0	2.5	0.03	0.040	Beta
Bacillus subtilis	5.5	0.15	0.085	1.25	0	0.15	0.079	0.14	0	0.2	0.062	Ψ
Caenorhabditis brenneri	1.5	0.09	0.094	1.5	0	0.06	0.086	0.09	0	0.06	0.105	Beta
Caenorhabditis elegans	0	0.06	0.142	2	0	0.06	0.1422	0	0	0.06	0.1422	KM
Chlamydia trachomatis	0	0.11	0.105	2	0	0.11	0.105	0	0	0.11	0.105	KM
Ciona intestinalis A	0	0.11	0.053	1.95	0	0.11	0.052	0	0	0.11	0.053	KM
Ciona intestinalis B	0.5	0.03	0.085	1.65	0	0.02	0.061	0.07	0	0.02	0.068	Beta
Clostridium difficile	15	0.15	0.214	1	0	0.15	0.221	0	17.5	0.2	0.214	KM
Corvus cornix	1.5	0	0.023	1.95	1	0	0.020	0	1.5	0	0.023	Beta
Coturnix japonica	4	0.02	0.048	1.45	0.5	0.01	0.020	0.07	1.5	0.01	0.044	Beta
Culex pipiens	2.5	0.02	0.069	1.55	0.5	0.01	0.057	0.07	1	0.01	0.063	Beta
Drosophila melanogaster	5.5	0.02	0.019	1.65	0.5	0.02	0.005	0.01	3	0.02	0.017	Beta
Egretta garzetta	0	0.02	0.055	1.75	0	0.02	0.037	0.07	0	0.02	0.039	Beta
Emys orbicularis	0.5	0	0.068	1.85	0	0	0.060	0.04	0	0	0.059	KM
Escherichia coli	0	0.06	0.054	2	0	0.06	0.054	0	0	0.06	0.054	KM
Ficedula albicollis	0.5	0.01	0.029	2	0.5	0.01	0.029	0.01	0.5	0.01	0.028	Ψ
Gorilla gorilla gorilla	0	0	0.042	1.9	0	0	0.040	0	0	0	0.042	Beta
Halictus scabiosae	0	0.01	0.069	1.85	0	0.01	0.064	0.04	0	0.01	0.062	MMC
Helicobacter pilori	1	0.15	0.052	1.65	0	0.15	0.060	0.01	1	0.2	0.050	Ψ
Homo sapiens	0.5	0.01	0.010	1.85	0	0	0.011	0	0.5	0.01	0.010	Beta
Klebsiella pneumoniae	18.5	0.15	0.122	2	18.5	0.15	0.126	0	18.5	0.16	0.122	KM
Lepus granatensis	0.5	0.04	0.102	1.5	0	0.03	0.069	0.12	0	0.03	0.066	MMC
Melitaea cinxia	1.5	0.04	0.061	1.7	0.5	0.03	0.059	0.01	2	0.04	0.061	Beta
Messor barbarus	0.5	0	0.069	2	0.5	0	0.069	0	0.5	0	0.069	KM
Mycobacterium tubercolosis	25	0.01	0.118	1.05	2.5	0	0.090	0.07	29	0	0.126	Beta
Nipponia nippon	0	0.03	0.160	2	0	0.03	0.160	0	0	0.03	0.160	KM
Ostrea edulis	0	0.02	0.052	1.8	0	0.02	0.044	0.04	0	0.02	0.042	MMC
Pan paniscus	2	0	0.068	1.85	1	0	0.056	0	2	0	0.068	Beta
Pan troglodytes ellioti	0.5	0	0.052	1.7	0	0	0.028	0.02	0.5	0	0.045	Beta
Parus major	0.5	0.01	0.031	1.75	0	0.01	0.010	0.03	0	0.01	0.022	Beta
Parus caeruleus	6	0.04	0.062	1.2	0	0	0.037	0.11	1.5	0.03	0.031	MMC
Passer domesticus	0	0	0.022	2	0	0	0.022	0	0	0	0.022	KM
Phylloscopus trochilus	12.5	0	0.022	2	12.5	0	0.022	0	12.5	0	0.022	KM
Physa acuta	1	0.03	0.068	1.5	0	0.02	0.035	0.05	0.5	0.03	0.055	Beta
Pseudomonas aeruginosa	25	0.15	0.073	1.1	0	0.15	0.063	0.06	3	0.2	0.050	Ψ
Sepia officinalis	0	0.02	0.091	1.95	0	0.02	0.090	0.01	0	0.02	0.090	KM
Staphylococcus aureus	1	0.15	0.054	1.7	0	0.15	0.059	0.01	1	0.2	0.055	Ψ
Streptococcus pneumoniae	1	0.12	0.103	1.5	0	0.08	0.099	0.09	0	0.08	0.102	Beta
Taeniopygia guttata	9.5	0	0.034	1.75	4	0	0.019	0.01	10	0	0.030	Beta
Zea mays	0	0	0.033	1.95	0	0	0.031	0.01	0	0	0.030	Ψ

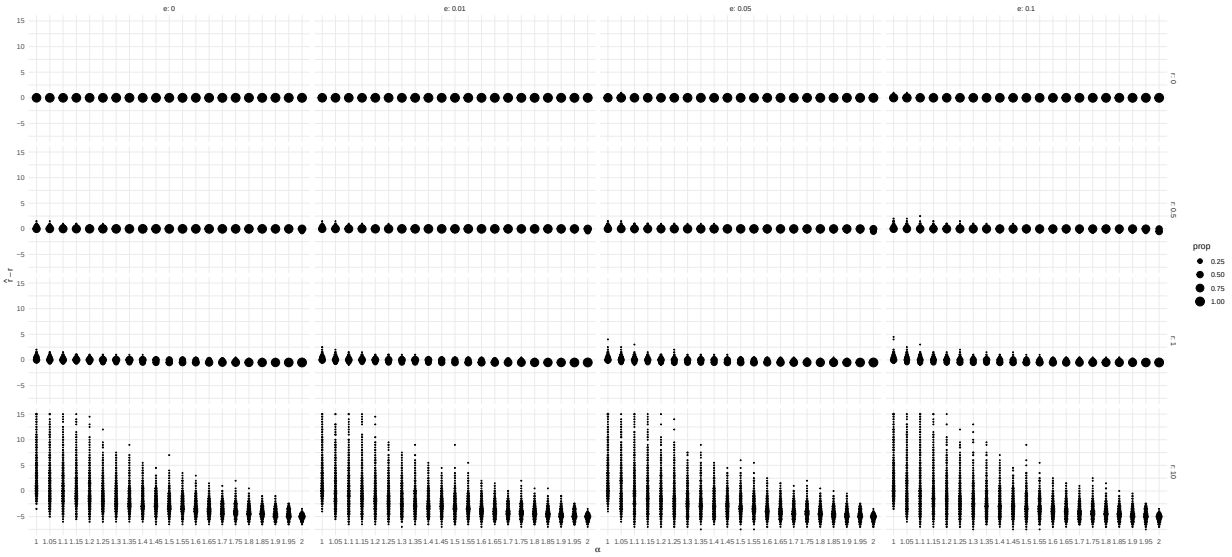


Figure 10: Error for estimating growth rate g for Beta-coalescents with growth and misclassification ($n = 100$)

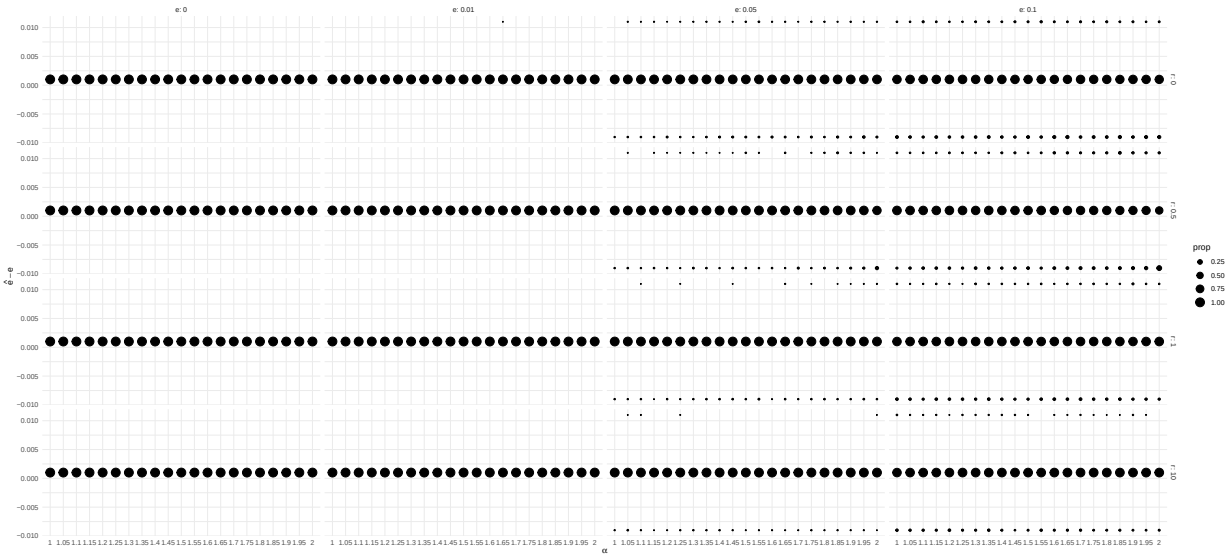


Figure 11: Error for estimating confusion rate e for Beta-coalescents with growth and misclassification ($n = 100$). Growth rate is denoted by g.

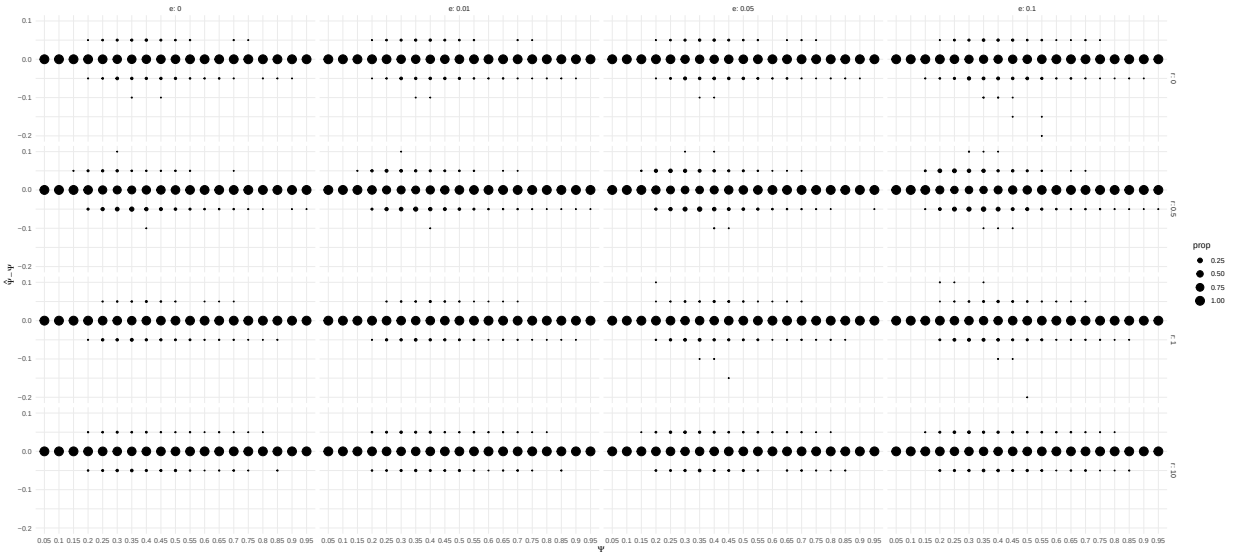


Figure 12: Error for estimating coalescent parameter Ψ for Psi-coalescents with growth and misclassification ($n = 100$). Growth rate is denoted by g.

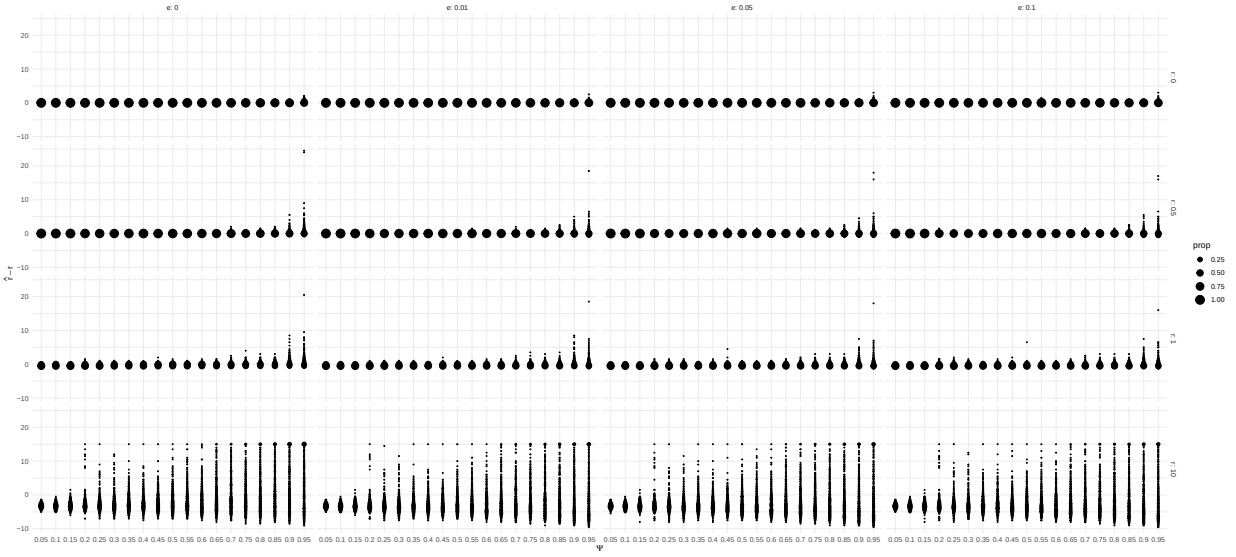


Figure 13: Error for estimating growth rate g for Psi-coalescents with growth and misclassification ($n = 100$).

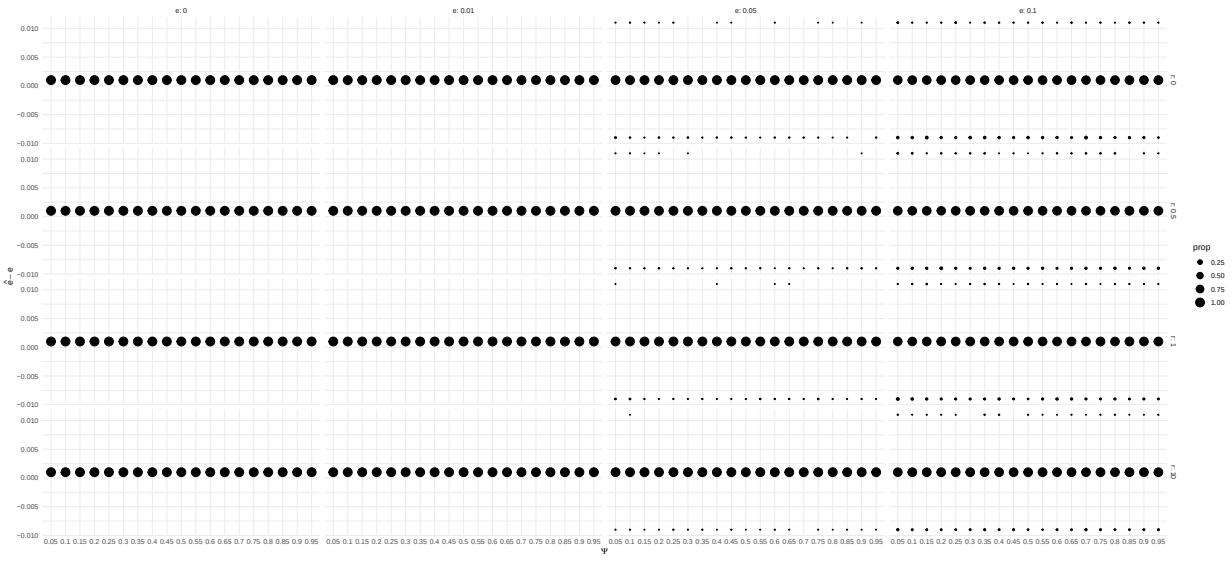


Figure 14: Error for estimating confusion rate e for Psi-coalescents with growth and misclassification ($n = 100$). Growth rate is denoted by g.

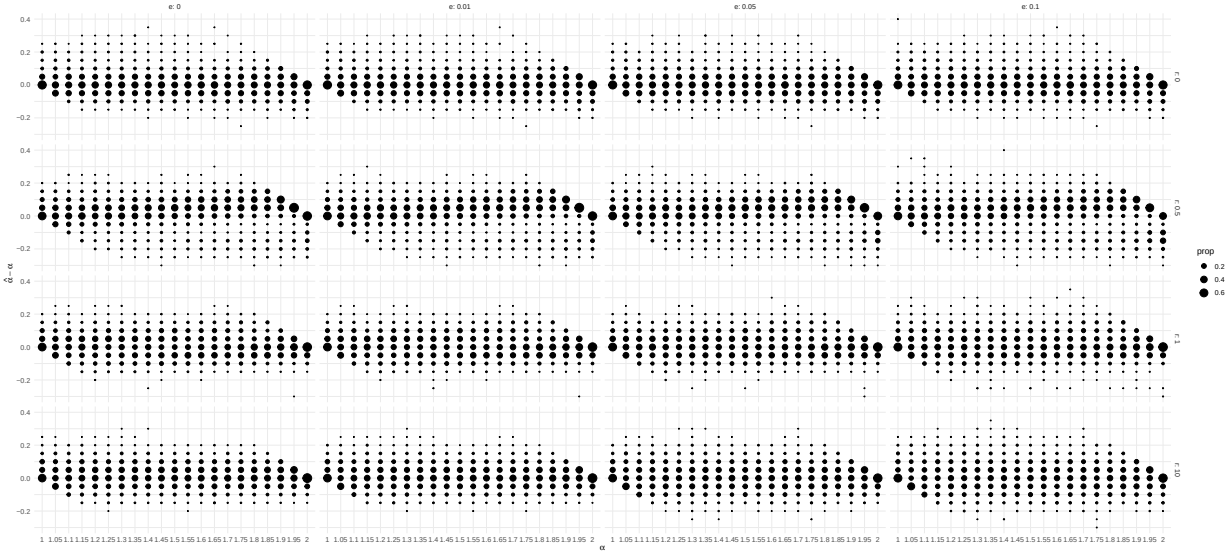


Figure 15: Error for estimating coalescent parameter α for Beta coalescents with growth and misclassification ($n = 20$). Growth rate is denoted by g.

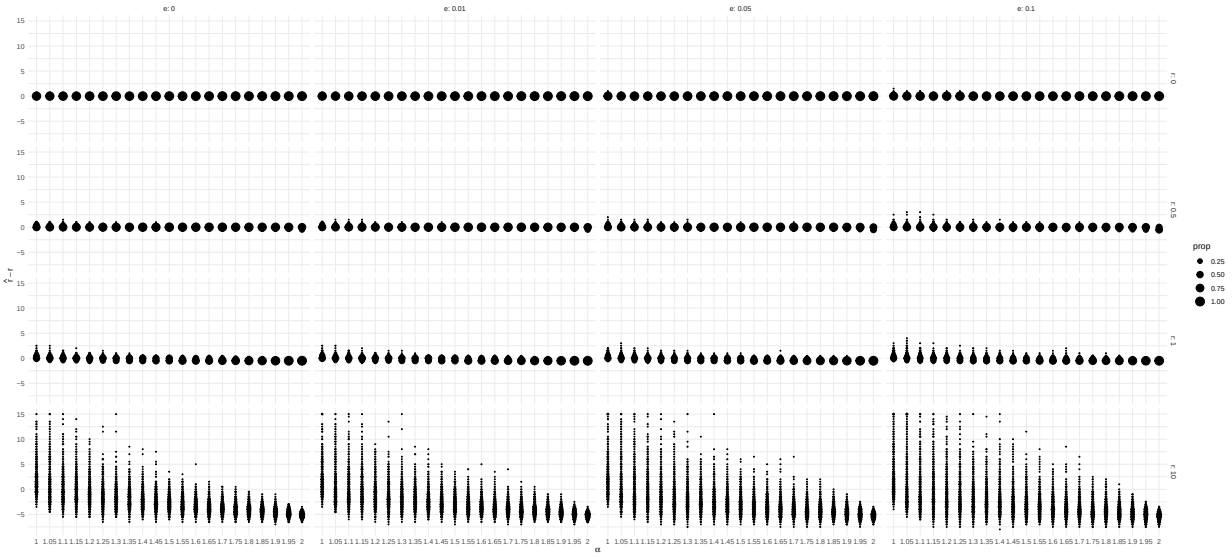


Figure 16: Error for estimating growth rate g for Beta-coalescents with growth and misclassification ($n = 20$)

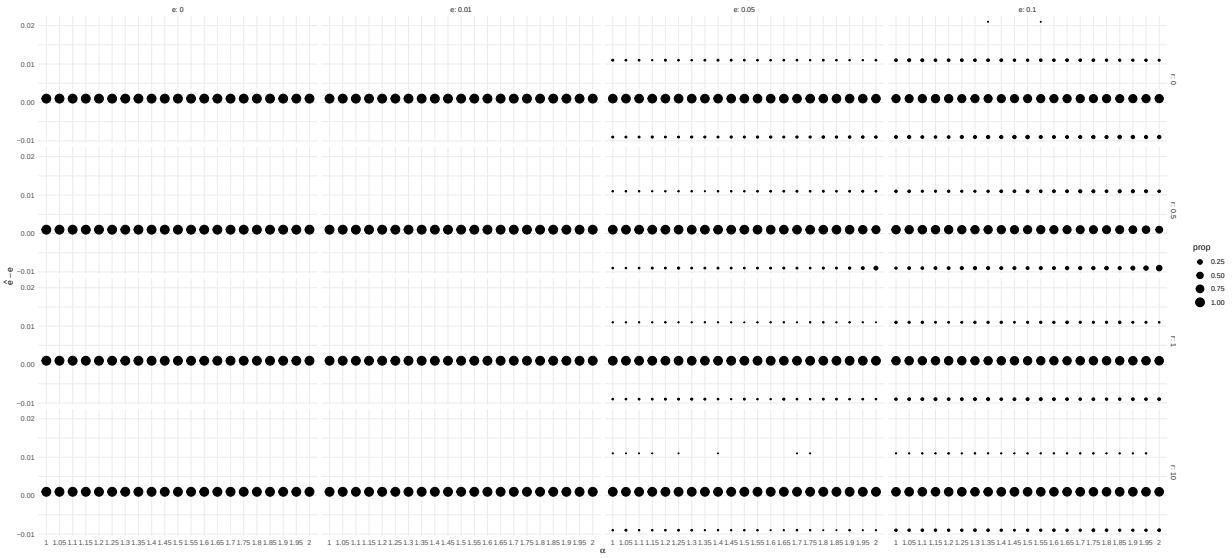


Figure 17: Error for estimating confusion rate e for Beta-coalescents with growth and misclassification ($n = 20$). Growth rate is denoted by g.

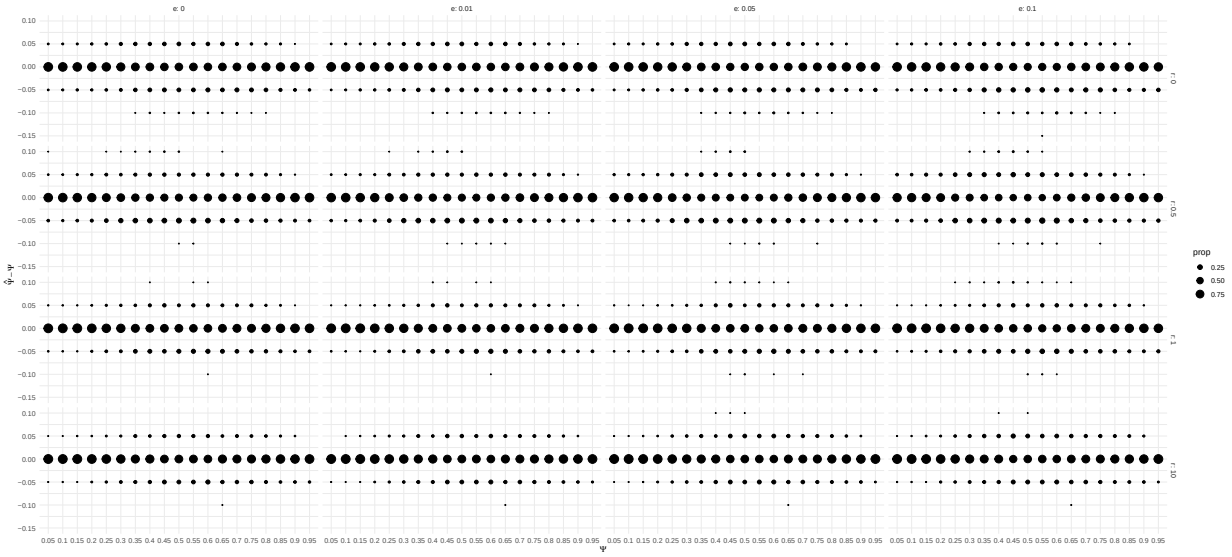


Figure 18: Error for estimating coalescent parameter Ψ for Psi-coalescents with growth and misclassification ($n = 20$). Growth rate is denoted by g.

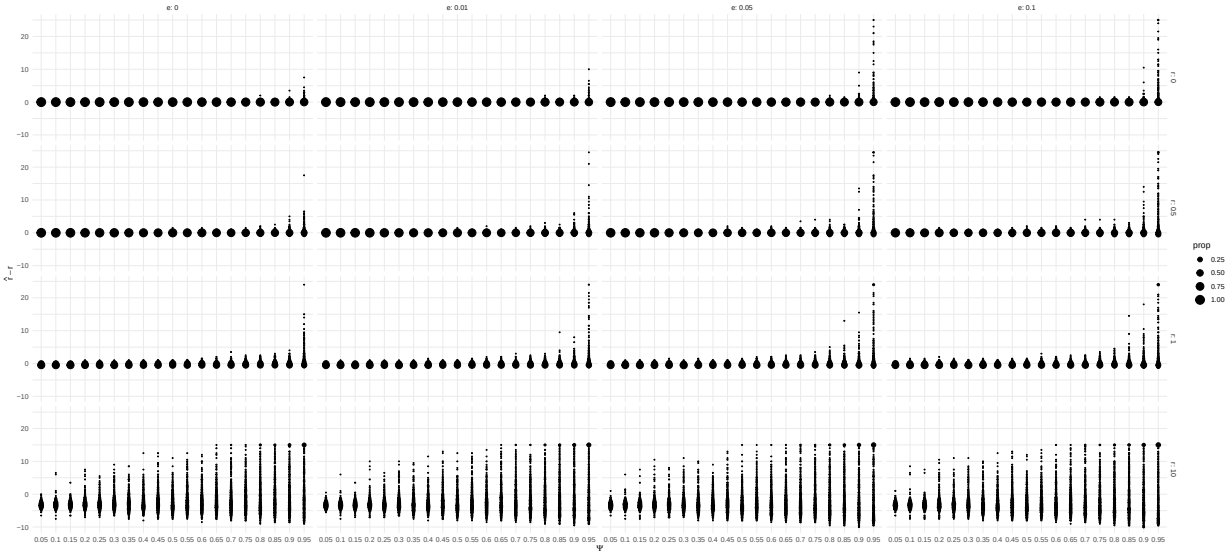


Figure 19: Error for estimating growth rate g for Psi-coalescents with growth and misclassification ($n = 20$)

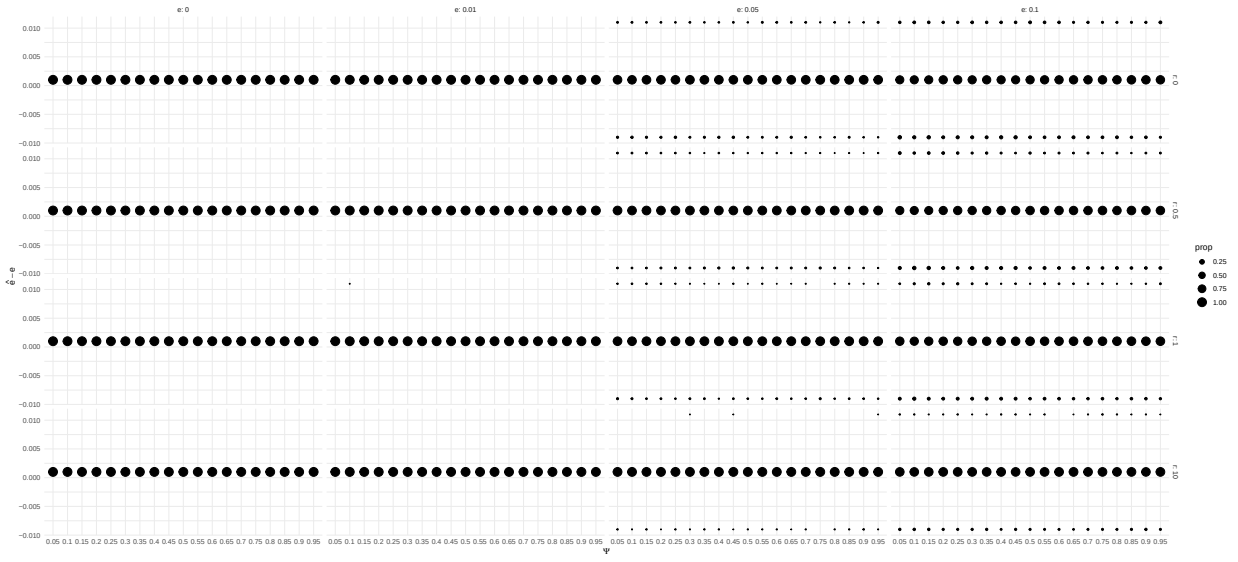
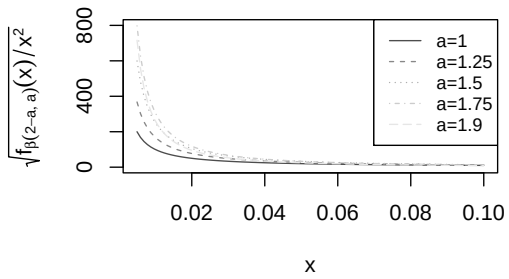
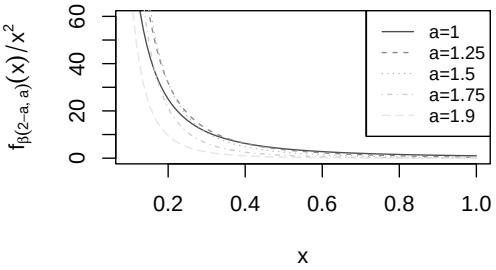


Figure 20: Error for estimating confusion rate e for Psi-coalescents with growth and misclassification ($n = 20$). Growth rate is denoted by g.



(a) (Improper) distribution of mergP x close to 0



(b) (Improper) distribution of bigger mergP x

Figure 21: Distribution of merger rates of Beta-coalescents: Each lineage merges with merger probability x (abbeviated as mergP), where x is chosen with rate $x^{-}2 * \Lambda(dx)$, where Λ is a Beta distribution with parameters $2 - a$ and a. Mergers only are realized if at least two lineages merge. The figures depict the corresponding (improper) density $x^{-}2 * f_\beta(2 - a, a)$, where f_β is the density of the Beta distribution used. The detailed (Poisson) construction can be found in [Pit99].

Table 7: Data set information

Species	Outgroup	n	Polarized SNP	$\neq$ outgroup	Diallelic outgroup	Source
Acinetobacter baumannii	*A. nosocomialis*	79	78175	6006		[RdSB$^+$18]*
Aptenodytes patagonicus	*A. forsteri*	20	1278	12	32	[RGB$^+$14]
Arabidopsis thaliana	*A. lyrata*	345	10322757	1023148	398365	[ABAB$^+$16]
Armadillidium vulgare	*A. nasatum*	20	23323	745		[RGB$^+$14]
Artemia franciscana	*A. sinica*	20	5548	247		[RGB$^+$14]
Athene cunicularia	*Strix occidentalis*	40	11268203	383702	68196	⋆
Bacillus subtilis	*B. atrophaeus*	38	105523	29934		[RdSB$^+$18]*
Caenorhabditis brenneri	*Caenorhabditis sp. 10*	20	1339	106		[RGB$^+$14]
Caenorhabditis elegans (Orsay population)	*C. elegans* ECA396 ECA723 ECA744	573	165	5	22	[RZL$^+$18]
Chlamydia trachomatis	*C. muridarum*	59	9924	1694		[RdSB$^+$18]*
Ciona intestinalis A	*C. intestinalis B*	20	480	94	1641	[RGB$^+$14]
Ciona intestinalis B	*C. intestinalis A*	20	1883	59	490	[RGB$^+$14]
Clostridium difficile	*Anaerococcus prevotii*	11	192	49		[RdSB$^+$18]*
Corvus cornix	*C. monedula*	38	7167395	25949	664205	⋆
Coturnix japonica	*Gallus varius*	20	5061864	87069	220450	⋆
Culex pipiens	*C. torrentium*	20	5442	106		[RGB$^+$14]
Drosophila melanogaster	*D. simulans*	196	4662706	151138		[LCC$^+$15]
Egretta garzetta	*Pelecanus crispus*	10	9318499	361539	10242	⋆
Emys orbicularis	*Trachemys scripta*	20	515	14		[RGB$^+$14]
Escherichia coli	*E. fergusonii*	62	84222	6903		RefSeq used in [LBL$^+$16]
Ficedula albicollis	*F. hypoleuca*	24	14697230	269430	229260	⋆
Gorilla gorilla	ancestral allele call from [PMSK$^+$13]	54	9878547	42	569321	[PMSK$^+$13]
Halictus scabiosae	*H simplex*	22	712	10		[RGB$^+$14]
Helicobacter pilori	*H. felis*	70	27498	8235		[RdSB$^+$18]*
Homo sapiens (Yoruba population)	ancestral allele call from [Con15]	216	19441528	105146		[Con15]
Klebsiella pneumoniae	*K. varicola*	156	203601	375		[RdSB$^+$18]*
Lepus granatensis	*L americanus*	20	769	31		[RGB$^+$14]
Melitaea cinxia	*M didyma*	18	1695	101		[RGB$^+$14]
Messor barbarus	*M structor*	20	9651	50		[RGB$^+$14]
Mycobacterium tubercolosis	MYCN001 - MYCN005	33	7142	13	78	RefSeq
Nipponia nippon	*Pelecanus crispus*	16	1140694	44153	2034	⋆
Ostrea edulis	*O. chilensis*	20	939	28		[RGB$^+$14]
Pan paniscus	ancestral allele call from [PMSK$^+$13]	26	6293657	63	284527	[PMSK$^+$13]
Pan troglodytes ellioti	ancestral allele call from [PMSK$^+$13]	20	10009190	44	459884	[PMSK$^+$13]
Parus major	*Cyanistes caeruleus*	54	14174305	143760	126876	⋆
Parus caeruleus	*P. major*	20	866	51	19	[RGB$^+$14]
Passer domesticus	*P. montanus*	16	18501992	90623	633399	⋆
Phylloscopus trochilus	*P. tristis*	24	33401127	8605	6092936	⋆
Physa acuta	*P. gyrina*	18	4286	176		[RGB$^+$14]
Sepia officinalis	*Sepiella japonica*	18	1740	52		[RGB$^+$14]
Pseudomonas aeruginosa	*P. knackmussii*	86	90258	17208		[RdSB$^+$18]*
Staphylococcus aureus	*S. epidermis*	152	30052	8694		[RdSB$^+$18]*
Streptococcus pneumoniae	*S. mitis*	32	49917	2468		[RdSB$^+$18]*
Taeniopygia guttata	*Poephila acuticauda*	38	53263038	118506	4346767	⋆
Zea mays	*Tripsacum dactyloides*	66	520310	214398		[BMHR$^+$17]

*Core genomes were computed and aligned as is [RdSB$^+$18], with the difference that for every species we added to the set of a species genome one further genome from the closest species (to be able to orient the changes).

⋆ `https://github.com/harvardinformatics/shortRead_mapping_variantCalling`

Table 8: Samples checked for population structure via optimal k-means and PCA. Number of clusters: k inferred by BIC criterion for subsequent k-means clustering. PCA plots with coloured inferred clusters are available in supplementary file 3.

Species	population structure	Number of clusters
Acinetobacter baumannii	yes	14
Aptenodytes patagonicus	potentially	5
Artemia franciscana	no	1
Athene cunicularia	no	1
Bacillus subtilis	potentially	5
Coturnix japonica	no	1
Culex pipiens	yes	2
Halictus scabiosae	potentially	2
Ostrea edulis	yes	2
Parus caeruleus	no	1
Parus maior	no	1
Pseudomonas aeruginosa	potentially	9

4.3 Fichiers supplémentaires

Les fichiers supplémentaires cités dans l'article se trouve en Annexes 6.2 de cette thèse.

— Supplementary file 1 : Annexes 6.2.1.

— Supplementary file 2 : Annexes 6.2.2.

— Supplementary file 3 : Annexes 6.2.3.

Chapitre 5

Discussion

Contents

5.1 Conclusion générale

5.1.1 Résumé

Dans cette thèse, je me suis intéressée à différents facteurs mesurables sur un alignement de séquences et j'ai étudié ce qu'ils pouvaient nous enseigner sur la démographie passée d'une population.

Dans le cadre de ma thèse je me suis intéressée à l'information portée par les évènements de mutation, observables par les SNPs présents dans le génome d'individus d'une même population. Pour cela, j'ai utilisé le SFS et montré par quelle gamme de changements de taille de population il était affecté. Par exemple, les logiciels utilisant le SFS détectent plus facilement une croissance qu'un déclin (Chapitre Inférences de changement de taille de population à partir de SFS). J'ai également utilisé le SFS d'espèces disséminées sur l'ensemble de l'arbre du vivant pour remettre en question l'utilisation du modèle standard neutre au profit de modèles plus généraux, permettant les multifurcations au sein des arbres de coalescence (Chapitre U-shaped genome site frequency spectra : challenging the reference model of molecular evolution ?).

Je me suis également intéressée à l'information contenue dans les évènements de recombinaison le long des génomes. Ces évènements sont de natures différentes et peuvent être plus ou moins difficiles à mettre en évidence suivant l'histoire évolutive de l'arbre de coalescence (Chapitre Impact de la démographie sur les évènements de recombinaison). La distribution des évènements de recombinaison le long du génome, ainsi que celle des évènements de recombinaison détectables, sont étudiables en considérant la distribution des distances séparant deux évènements. Ces distributions sont très informatives et peuvent, par exemple, servir à l'inférence démographique (Chapitre Testing for population decline using maximal linkage disequilibrium blocks). Il est cependant important de noter qu'elles sont très difficiles à échantillonner. De nombreux facteurs influencent les distributions mesurables sur les alignements de séquences ce qui peut rendre leur étude complexe (Chapitre Comment expliquer les distributions de longueurs de blocs MLD observées ?).

Pour finir, j'ai souhaité combiner l'information des évènements de mutation et celle des évènements de recombinaison. Dans un premier temps j'ai développé un cadre probabiliste permettant l'inférence démographique en combinant l'information du SFS et de la distribution des longueurs de blocs MLD (Chapitre Inférence

combinant SFS et blocs MLD). Puis dans un deuxième temps, j'ai utilisé une mesure affectée à la fois par les évènements de mutation et par les évènements de recombinaison : le déséquilibre de liaison (ou LD). Pour utiliser au mieux les deux informations, il est possible de compartimenter les mesures de LD en fonction du nombre d'évènements de recombinaison les séparant. Il est également possible de considérer, par exemple, le LD entre des sites appartenant au même arbre de coalescence ou le LD entre des sites séparés par seulement un évènement de recombinaison, etc. Les distributions de LD varient fortement en fonction de la démographie et du nombre d'évènements de recombinaison (Chapitre Étude de la distribution du déséquilibre de liaison). Les évènements de recombinaison étant difficilement détectables, il a fallu modifier légèrement la statistique pour pouvoir l'appliquer à des données réelles. Cependant, les mesures basées sur le LD compartimenté permettent de détecter des déclins connus non identifiables par d'autres méthodes (Chapitre Inférences utilisant D_0 et D_1).

5.1.2 Définition des blocs *Maximal Recombination Free* et des blocs *Maximal Linkage Disequilibrium*

5.1.2.1 Des objets pour étudier l'histoire évolutive

Les segments IBD, bien que définis par les théoriciens comme pouvant être portés par n individus, sont peu étudiés pour $n > 2$ individus. L'ensemble des techniques servant à les détecter ou bien à les utiliser pour inférer l'histoire démographique d'une population, s'intéressent aux segments IBD entre deux génomes. Au cours de ma thèse, j'ai voulu exploiter l'information des segments IBD à l'échelle d'un échantillon. C'est de cette manière que se définissent les blocs MRF, il s'agit de segments portés par tous les n individus (avec $n \geq 2$), dont les extrémités sont définis par l'emplacement d'évènements de recombinaison ayant lieu dans l'histoire d'un ou plusieurs individus échantillonnés. La longueur de ces segments dépend de l'histoire évolutive de la population, il peuvent donc être d'une grande utilité pour l'inférence démographique.

Ces blocs ne sont pas détectables sur un génome. C'est pour cela que nous avons défini des blocs aux caractéristiques similaires : les blocs MLD. Ces segments contenant des sites compatibles avec une seule topologie d'arbre, sont délimités par les évènements de recombinaison détectables par le test des 4-gamètes. Comme il a été démontré dans le chapitre Testing for population decline using maximal linkage disequilibrium blocks, la distribution de longueurs de ces blocs peut aussi être utilisée pour étudier l'histoire évolutive d'une population.

Dans le cadre de ma thèse, je me suis seulement intéressée à l'inférence de la démographie passée. Mais ces blocs, dont la distribution dépend de la topologie et des longueurs de branches de l'arbre de coalescence, pourraient également être utilisés pour mettre en évidence d'autres phénomènes influençant les arbres de coalescence d'une population, comme c'est le cas de la structure de population.

5.1.2.2 Des supports pour étudier l'histoire évolutive

Ces blocs peuvent non seulement servir en tant que tels pour étudier l'histoire évolutive d'une population, mais également servir de support à d'autres mesures statistiques. C'est ce qui a été fait des les chapitres Étude de la distribution du déséquilibre de liaison et Inférences utilisant D_0 et D_1, les alignements de génomes ont été découpés grâce aux blocs MRF et MLD pour en étudier les distributions de D les contenant. Cela a permis de sous-échantillonner la distribution globale pour séparer les effets des différents facteurs, afin d'affiner les inférences possibles. Ce découpage a été fait pour mesurer les valeurs de D, mais pourrait servir pour d'autres mesures statistiques. De plus, la longueur des blocs apporte une information sur l'histoire passée, p. ex. un bloc long n'a pas le même âge qu'un bloc court. Il pourrait être intéressant de regrouper les blocs par catégorie et ainsi de mesurer les statistiques en fonction de ces catégories. Cela permettrait de séparer différentes temporalités et également différentes histoires évolutives disjointes.

En effet, chaque bloc s'appuie sur une topologie. Ces dernières peuvent être compatibles entre elles ou non. Des incompatibilités très prononcées peuvent provenir d'évènements passés, comme un mélange de populations ancien, ou de la structure de population. Le regroupement de blocs en fonction de leur topologie sous-jacente pourrait également nous informer sur l'histoire évolutive de la population. Dans le cadre d'un mélange ancien de deux populations, certaines topologies seraient plus influencées par l'histoire évolutive d'une population ou de l'autre. Inférer les topologies de chaque bloc peut être complexe, mais il est envisageable de regrouper les blocs dont les topologies sont compatibles avec la même histoire. Deux blocs MLD adjacents ne sont pas compatibles, mais des blocs à une distance plus importante peuvent l'être.

5.1.3 La distribution de D

5.1.3.1 Une autre façon d'accéder au 2-SFS

Une mesure du déséquilibre de liaison D, se calculant $D_{AB} = f_{AB} - f_A f_B$ avec f_A et f_B respectivement les fréquences de A et de B et f_{AB} la fréquence de la co-occurrence AB, est peu utilisée dans sa forme non-normalisée. La covariance D peut cependant s'apparenter à l'écart du 2-SFS par rapport à son attendu théorique si les sites ne sont pas liés, quand la recombinaison est maximale. Le 2-SFS (Ferretti et al. 2018) ou *two-locus frequency spectrum* (Hudson 2001) est le SFS joint de deux locus, son attendu théorique pour deux sites conditionnés à avoir le même arbre est l'association aléatoire de deux SFS simples. Le 2-SFS est très difficile à étudier analytiquement car il dépend à la fois de l'histoire évolutive (comme la démographie) affectant le SFS simple et de la recombinaison créant des associations (ou linkage) plus ou moins fortes entre les locus.

5.1.3.2 Dépendant de la démographie et de la recombinaison

La distribution de D dépend donc logiquement à la fois de la démographie et de l'effet de la recombinaison. La normalisation de D a pour conséquence d'effacer l'effet de la fréquence allélique des sites (pour $|D'|$ et r), grandement affectée par la démographie. Cependant, comme nous l'avons montré dans le chapitre Inférences utilisant D_0 et D_1 la distribution de D peut être très informative et aider à l'inférence de la démographie. Elle pourrait servir à mettre en avant d'autres processus évolutifs comme la sélection ou la structure de population.

5.1.3.3 Permettant d'étudier la démographie et la recombinaison séparément

De plus, nous avons montré dans le chapitre Étude de la distribution du déséquilibre de liaison qu'il est possible de séparer l'effet de la topologie et des longueurs de branches (dépendant de la démographie), de l'effet de la recombinaison, en étudiant la distribution du déséquilibre de liaison par bloc MRF ainsi que par bloc MLD. L'étude de la distribution du déséquilibre de liaison par bloc MLD ou par paire de blocs adjacents, ne prend pas en compte l'effet combiné de tous les évènements de recombinaison, seulement de ceux qui sont détectables mais permet tout de même de gagner en information : la distribution à l'intérieur d'un même MLD ou entre deux MLD adjacents pouvant ne pas porter la même information (Chapitre Inférences utilisant D_0 et D_1).

Il serait donc intéressant de procéder de la même manière avec le 2-SFS, pour lequel des résultats analytiques sont disponible pour deux sites conditionnés à avoir le même arbre de coalescence.

5.1.4 Combinaison de l'information des mutations et des recombinaisons

Dans le cadre de ma thèse, j'ai combiné l'information portée par les évènements de mutation et celle portée par les évènements de recombinaison. La combinaison de différents types d'information permet d'améliorer les inférences d'histoires évolutives grâce aux différents paramètres et aux différentes temporalités auxquels ces statistiques sont sensibles.

5.1.4.1 Différents paramètres

En effet, comme montré dans le chapitre Inférence combinant SFS et blocs MLD, dans le cadre d'un scénario de déclin récent, le SFS résumant la distribution des fréquences de mutations, est plus sensible à la date du déclin qu'à l'intensité du déclin. La distribution des blocs MLD correspondant aux distances entre deux évènements de recombinaison détectables, est plus sensible à l'intensité du déclin qu'à sa date. C'est bien la combinaison des deux, la différence de sensibilité entre

ces deux statistiques, qui permet une meilleure inférence du couple date/intensité du déclin.

5.1.4.2 Différentes temporalités

Il est également possible que des statistiques différentes soient sensibles pour des temps différents de l'histoire évolutive d'une population. Par exemple, D_0 et D_1 permettent de détecter des changements de taille de population plus anciens que le SFS (voir Chapitre Inférences utilisant D_0 et D_1). Dans le cadre de ce travail, cette différence à permis de mettre en avant une erreur dans le choix du modèle à inférer. Le SFS et les distributions de D_0 et D_1 indiquant des changements démographiques différents, nous avons considéré un scénario comprenant deux changements de taille de population en accord avec les trois distributions.

5.1.4.3 Combinaison de différentes méthodes

Des méthodes s'appuyant sur le même type d'information peuvent être affectées différemment et ne pas inférer le même scénario avec les mêmes informations. Par exemple, deux logiciels utilisant l'information du SFS n'ont pas exactement les mêmes limites pour détecter un changement de taille de population (voir chapitre Inférences de changement de taille de population à partir de SFS).

Méthodes et statistiques peuvent être affectées différemment par des intrications complexes de l'histoire démographique d'une population. Il est donc recommandé de comparer les résultats obtenus par des méthodes différentes (Beichman et al. 2018). Certaines méthodes sont difficiles à appliquer à des espèces non humaines, mais de plus en plus d'études d'inférences démographiques, comme MacLeod et al. 2013; Prado-Martinez et al. 2013; Prates et al. 2016; Zhan et al. 2014, utilisent plusieurs méthodes.

Les nouvelles méthodes combinant plusieurs statistiques mesurées sur les données sont donc celles à privilégier. Comme Pop-SizeABC (Boitard et al. 2016) qui combine le SFS et le LD (r^2) moyen entre des sites à différentes distances ou SMC++ (Terhorst et al. 2017) qui combine l'information des T_{MRCA} et du SFS. Jay et al. (2019) ont démontré que l'utilisation jointe de statistiques basées sur des informations différentes, comme la longueur des segments IBD, ou la diminution du LD avec la distance physique, avec des statistiques plus classiques comme le D de Tajima, était possible pour inférer des scénarios démographiques.

5.1.5 La remise en cause du modèle neutre

5.1.5.1 Une question d'adéquation aux données

Le modèle standard neutre actuel n'est que rarement en adéquation avec les données, indifféremment de l'emplacement de l'espèce dans l'arbre de la vie. Des modèles, plus inclusifs, comme les Multiple-Merger Coalescents, s'ajustent beaucoup mieux aux données (et comprennent le coalescent de Kingman dans leur

gamme de paramètres). En effet, l'utilisation de modèles comprenant des coalescences multiples a permis, pour 32 des 45 SFS, de produire des scénarios en adéquation avec les SFS des espèces échantillonnées. (chapitre U-shaped genome site frequency spectra : challenging the reference model of molecular evolution ?). Les MMC semblent donc être des candidats potentiels pour devenir les nouveaux modèles standards neutres, même s'ils ne permettent pas d'expliquer l'ensemble des données observées.

5.1.5.2 Implications d'un mauvais choix de modèle

Les estimations de paramètres dépendent du modèle utilisé, un mauvais choix de modèle entraîne de mauvaises estimations.

Les estimations de taille de populations faites à partir des génomes d'individus de la population servent à la mise en place de politique de conservation. La conservation d'espèces en danger peut se faire grâce à la protection des liens entre des habitats fragmentés, liens mis en évidence par les inférences de flux génétiques ou de migrations entre populations. La plupart des outils statistiques utilisés pour ces inférences utilisent seulement le coalescent de Kingman. Cependant, si le modèle correspondant à la population n'est pas le coalescent de Kingman, les méthodes inférant le mauvais modèle ne rendront pas compte de la réalité biologique de l'espèce.

Le développement d'outils statistiques utilisant les MMC, offrant une plus grande gamme de modèles, devraient améliorer les estimations de paramètres démographiques ayant une importance pour la conservation (Montano 2016).

5.2 Limites

Mes études s'intéressent aux inférences démographiques à partir de données d'alignement de génomes. Il existe de nombreux phénomènes pouvant imiter des signaux identiques à ceux mesurés pour l'inférence démographique ou encore atténuer, voire biaiser, les signaux laissés par l'histoire démographique d'une population. Il est important de les identifier et, si possible, de les prendre en compte.

Je liste ici, les phénomènes pouvant biaiser les méthodes que j'ai utilisées ou développées pendant ma thèse.

5.2.1 Structure

Une des hypothèses du modèle standard neutre est l'absence de structure de population. Les individus se reproduisent aléatoirement les uns avec les autres. Bien que cela soit rarement le cas dans les populations naturelles, les scénarios de structure de population sont peu souvent considérés. En effet, certains logiciels, comme PSMC (Li and Durbin 2011 ou Stairway Plot (Liu and Fu 2015), proposent

uniquement l'inférence de scénarios démographiques, sans possibilité de prendre en compte la structure de la population.

Néanmoins, des analyses de structure de la population cherchant à catégoriser de potentielles sous-populations sont régulièrement effectuées. Elles utilisent, par exemple, des ACP (comme pour les données du chapitre U-shaped genome site frequency spectra : challenging the reference model of molecular evolution ?) ou se basent sur des logiciels spécialisés comme Structure (Hubisz et al. 2009.

Ces analyses ne sont pourtant pas suffisantes pour éliminer la présence de structure de population. En effet, il existe des scénarios de structure de population complexes, ne produisant pas ou peu de signal équivalant à des sous-populations, des ensembles d'individus possédant des marqueurs génétiques particuliers. Pour certaines statistiques, comme l'IICR, il n'est pas possible de faire la distinction entre de la structure de population ou de la démographie. En effet, pour chaque scénario démographie, correspond à un scénario de structure produisant la même distribution (Mazet et al. 2016; Chikhi et al. 2018).

Les mesures, ainsi que les méthodes en découlant, utilisées et développées au cours de ma thèse souffrent des mêmes maux. Par exemple, un modèle continent-île déformera la distribution normalisée des MLD dans la même direction qu'un déclin de population (comme montré dans le chapitre Testing for population decline using maximal linkage disequilibrium blocks). L'effet de la structure de la population n'a pas encore été étudié sur D_0 et D_1. Cependant D_0 et D_1 étant influencés par la topologie et les longueurs de branches de l'arbre de coalescence, sur lesquels la structure a un impact, la structure modifiera certainement les distributions de D_0 et de D_1.

5.2.2 Autres écarts aux hypothèses du modèle

5.2.2.1 Sélection

Une autre hypothèse du modèle standard neutre est la neutralité (ou quasi-neutralité) des mutations survenant au sein de la population. Il est cependant certain qu'il existe des mutations sélectionnées et qu'elles entraînent des modifications de la distribution des génomes dans la population, notamment en faisant augmenter les allèles liés physiquement (auto-stop génétique). Même si ces évènements sont peu fréquents et restent localisés sur le génome, ils doivent, par exemple, entraîner des blocs MLD plus longs qui, dans des cas extrêmes, pourraient avoir un impact sur la distribution des longueurs de blocs MLD.

La sélection peut prendre différentes formes, certaines sont plus ou moins faciles à détecter. Pour certaines, un sous-échantillonage des données utilisées pour les inférences permet de ne garder que les régions du génome dont l'histoire évolutive est neutre ou quasi-neutre. Il est possible, par exemple, de considérer seulement les parties non codantes des génomes quand celles-ci sont connues. Dans le cas du SFS, il est également possible de considérer seulement les mutations non biaisées (A-T ou C-G) pour contrer l'effet du biais de GC (Marchi and Excoffier 2020).

5.2.2.2 Variation du taux de recombinaison

Lors de nos études sur les évènements de recombinaison et l'information qu'ils portent, nous avons considéré un taux de recombinaison constant dans le temps et sur le génome. Il est pourtant connu que le taux de recombinaison est variable le long du génome, dépendant par exemple de la présence de PRDM9 chez les vertébrés (Baudat et al. 2010). Pourtant, en considérant seulement une portion du génome où les évènements de recombinaison ont une densité uniforme, nous n'avons pas observé de changement dans la distribution des blocs MLD (voir chapitre Comment expliquer les distributions de longueurs de blocs MLD observées ?). La variabilité du taux de recombinaison semble ne pas avoir d'effet dans ce cas précis mais ce n'est peut être pas toujours le cas. Quand les blocs MLD servent au découpage du génome pour mesurer d'autres statistiques, un taux de recombinaison variable n'aurait pas d'impact sur la statistique en découlant, excepté dans le cas où la détectabilité des évènements de recombinaison aurait gravement chuté.

L'écart à cette hypothèse doit également être pris en compte dans l'estimation du taux de recombinaison à l'échelle populationnelle. En plus de devoir considérer l'interaction entre démographie et effet d'un évènement de recombinaison sur le génome (voir chapitre Impact de la démographie sur les évènements de recombinaison), il est important d'autoriser le taux à être variable le long du génome.

5.2.2.3 Ratio taux de mutation sur taux de recombinaison

Impact du ratio. Le nombre d'évènements de mutation ayant lieu par évènement de recombinaison a un fort impact sur l'estimation de paramètres utilisant le SMC. En effet, plus il y a de mutation par évènement de recombinaison, plus il y a d'information sur la généalogie de chaque MRF, ce qui permet de mieux inférer ses caractéristiques. Les méthodes d'inférence utilisant le SMC sont plus précises quand le ratio taux de mutation sur taux de recombinaison est élevé (Sellinger et al. 2021).

Comme l'article Testing for population decline using maximal linkage disequilibrium blocks le montre, plus ce ratio augmente plus il est possible de détecter des évènements de recombinaison. Une meilleure détection des évènements de recombinaison permet de récolter plus d'informations sur les changements topologiques et permet également un découpage plus fin du génome en blocs MLD. Plus le nombre de blocs MLD augmente, plus la distribution des longueurs de blocs est précise, rendant plus précises les inférences utilisant la distribution des longueurs de blocs MLD.

Estimation du ratio. La plupart des méthodes utilise une approximation du taux de mutation par site μ, son équivalent populationnel : $\theta = 4N_e\mu$ (N_e étant la taille effective de la population) et le taux de recombinaison par site r par l'équivalent populationnel $\rho = 4N_e r$. Les taux moléculaires μ et r étant difficilement mesurables. Il existe cependant des déviations pour lesquelles $\frac{\mu}{r} \neq \frac{\theta}{\rho}$, comme dans

le cas de systèmes reproductifs différents d'un modèle de Wright-Fisher diploïde à deux sexes, qui causent des erreurs d'inférences (Sellinger et al. 2020).

Ce ratio et son estimation auront donc un effet sur la l'exactitude et la précision des inférences démographiques, il est indispensable de le prendre en compte.

5.2.3 Limites techniques

Hormis les écarts aux hypothèses du modèle standard neutre, une autre limitation pour inférer l'histoire évolutive d'une population existe : celle de l'accessibilité des données, des mesures nécessaires pour les inférences.

5.2.3.1 Problèmes de séquençage

Le séquençage de l'ADN devient de plus en plus abordable et de plus en plus performant, mais il comporte tout de même certaines limitations pouvant avoir un impact sur l'inférence de l'histoire évolutive.

Erreurs de séquençage. Un allèle peut être lu et reporté à la place d'un autre ce qui provoque des erreurs de séquençage. Ces erreurs dépendent de la qualité du séquençage et sont en grande partie des singletons. Il est donc possible d'éviter leur impact sur l'inférence démographique en ne considérant pas les singletons, à la fois pour le SFS ou pour le calcul de D (Achaz 2008). De nombreuses études comme Boitard et al. 2016 ne considèrent que les mutations ayant une fréquence allélique minimale, considérant les allèles peu fréquents comme peu fiables, pouvant provenir d'une erreur de séquençage.

Les singletons ne permettent pas de détecter un évènement de recombinaison : pour le test des 4 gamètes, un allèle doit au moins être présent deux fois, pour faire partie de deux gamètes. Ils n'ont donc aucun impact sur la distribution des longueurs de blocs MLD, cette méthode est donc robuste aux erreurs de séquençage.

Dans leur étude sur des méthodes d'inférence démographique utilisant les propriétés du SMC, Sellinger et al. montrent que s'il existe plus de 10% de faux SNPs, les logiciels utilisant l'information du IICR (comme MSMC et MSMC2) ont tendance à surestimer fortement la taille de la population pour des temps récents. L'absence de séquençage de SNPs n'a, quant à lui, pas d'effet sur les inférences dans les temps lointains et qu'un faible effet pour les temps récents.

Couverture de séquençage. Un chromosome n'est pas séquencé d'un bout à l'autre, il existe des parties non ou difficilement accessibles, comme les télomères, le centromère ou les parties du génomes comportant beaucoup d'éléments répétés. La couverture de séquençage n'est pas du tout considérée pour les méthodes utilisant la seule information des mutations, mais est très importante pour les méthodes utilisant l'information de liaison physique ou l'information des évènements de recombinaison. Elle est même limitante pour l'utilisation de la distribution normalisée des blocs MLD pour inférer une démographie. Même si, suivant la méthode utilisée, le fait de retirer aléatoirement certaines parties du génome n'a pas toujours

un impact sur l'estimation des variations de tailles de la population (Sellinger et al. 2021).

Grâce aux avancées techniques, il sera possible de séquencer des chromosomes quasi-entiers. Cependant, certaines régions du génome ne seront jamais accessibles car trop endommagées ou fragiles (comme les télomères).

Phasage des génomes diploïdes. Dans le cas de génomes diploïdes, ou polyploïdes, chaque cellule contient plusieurs jeux complets de chromosomes homologues. En séquençant l'ADN de la cellule globalement, il est difficile de différencier ces chromosomes, de savoir quel allèle est sur quel chromosome. On appelle phasage, le fait d'associer ensemble les allèles présents sur le même chromosome. Cela n'a pas d'impact sur les études de fréquences alléliques, mais est très important pour les études d'association d'allèles. Par exemple, des données phasées sont nécessaires pour utiliser les logiciels de détection de segments IBD.

Il existe des études de trios « père-mère-enfant » pour distinguer si l'allèle est sur le chromosome paternel ou maternel, mais elles demandent un effort de séquençage beaucoup plus important. Des logiciels ont été développés pour phaser statistiquement des génomes (Browning and Browning 2010).

Bien que le phasage soit important pour les études d'association, il est tout de même possible de mesurer certaines statistiques sur des données non phasées. Le test des 4 gamètes est facilement applicable sur des génomes non phasés (Kerdoncuff et al. 2020), perdant seulement un peu de puissance de détection. Le D est également mesurable sur des données non phasées grâce à une approximation (Ragsdale and Gravel 2019).

5.2.3.2 Problèmes d'accessibilité

Des évènements de recombinaison invisibles. L'utilisation des blocs MRF pour effectuer des inférences démographiques nécessiterait la détection de tous les évènements de recombinaison. Cependant, comme défini dans le chapitre Impact de la démographie sur les évènements de recombinaison, certains évènements de recombinaison sont invisibles. Par exemple, s'il n'y a aucun changement entre l'arbre de coalescence avant l'évènement et après l'évènement de recombinaison, ce dernier ne pourra jamais être détecté. Il est toutefois possible d'imaginer l'utilisation des autres types de recombinaison (définies dans Impact de la démographie sur les évènements de recombinaison), à la condition que ces dernières soient détectables, pour définir des blocs afin d'inférer l'histoire évolutive d'une population.

Ancestralité des allèles. Le plus souvent, pour inférer lequel des deux allèles observés est ancestral, on compare la séquence de la population avec celle d'une espèce sœur (*outgroup*). Si une seule mutation a eu lieu, l'espèce sœur porte l'allèle ancestral. Cependant, il est possible que deux mutations se produisent sur le même site. L'espèce sœur pourrait donc porter un troisième allèle, ou le même que l'espèce étudiée, provoquant une erreur dans l'ancestralité de l'allèle. Il est possible de prendre en compte ces erreurs d'ancestralité comme dans le chapitre

U-shaped genome site frequency spectra : challenging the reference model of molecular evolution ?. Si l'espèce sœur porte un troisième allèle, il n'est pas possible de connaître l'ancestralité de l'allèle. De plus, le génome d'une espèce sœur pour inférer l'ancestralité des allèles n'est pas toujours disponible.

Des variations des statistiques ont été mise en place pour utiliser les fréquences alléliques sans connaître l'ancestralité de la mutation, comme le SFS plié ou le calcul du D à partir de l'allèle minoritaire, *Minor Allele Frequency* (MAF) en anglais. La connaissance de l'ancestralité des allèles permet d'obtenir une puissance statistique plus grande pour inférer l'histoire démographique des espèces, pour, par exemple, différencier plusieurs scénarios.

Taille de la population au temps présent. Le temps de coalescence est exprimé en fonction de la taille de la population. La taille de la population est calculée en prenant en compte un modèle d'histoire évolutive. Notre modèle de changement soudain de taille de population (utilisé dans Inférences de changement de taille de population à partir de SFS,Testing for population decline using maximal linkage disequilibrium blocks et Inférences utilisant D_0 et D_1) est exprimé en fonction de N_0, la taille de la population au temps présent. Or, il n'est pas possible d'inférer la taille de population au temps présent avec notre méthode. Nos méthodes permettent donc d'inférer une intensité de changement de taille de population, mais pas la taille de la population. L'information de la taille de la population au temps présent peut être connue grâce, notamment, à des comptages. Cependant, la taille de population N_0 correspond à la taille efficace de la population au temps présent, qui ne peut être estimée (sauf si la population est restée constante).

5.3 Application à la conservation ?

5.3.1 Difficultés rencontrées

La plupart des méthodes d'inférences démographiques utilisant les données génétiques ont été conçues et appliquées à des données humaines. Elles considèrent souvent des génomes diploïdes phasés possédant une bonne couverture de séquencage et dont l'ancestralité des allèles est connue. Comme décrit précédemment, le non respect de ces conditions peut avoir un fort impact sur la qualité de l'inférence, voire totalement fausser l'inférence.

Les espèces en voie d'extinction, ou celles dont on ne connaît pas le statut de conservation sont rarement des organismes modèles. Il existe donc rarement un génome de référence pour ces espèces rendant le séquençage de génome complet plus compliqué. Il est pourtant nécessaire d'avoir accès à des séquences ADN non interrompues longues pour pouvoir utiliser l'information portée par les évènements de recombinaison. Par exemple, la distribution de blocs MLD n'est pas accessibles sans une très bonne couverture de séquencage. Pour l'utilisation de PSMC (Li and Durbin 2011) il est conseillé d'avoir une profondeur de séquencage moyenne

sur le génome de 18x, au moins 10 reads par site et moins de 25% de données manquantes (Nadachowska-Brzyska et al. 2016). La méthode permettant d'utiliser le plus facilement l'information portée par les recombinaisons quand le séquençage de génomes complets est difficile, est la mesure du déséquilibre de liaison par bloc MLD.

Une autre difficulté est que les organismes non-modèles ne possèdent pas souvent d'espèce proche séquencée, l'ancestralité des allèles n'est pas souvent accessible. Cela rend l'information des évènements de mutation moins précise. Par exemple, il est nécessaire d'utiliser un SFS plié à la place d'un déplié et donc de diviser le nombre de « cases » de la distribution par deux.

Le modèle le plus utilisé pour le développement de méthodes d'inférence est le coalescent de Kingman. Cependant, beaucoup d'espèces diffèrent de ce modèle, il existe des méthodes permettant de prendre en compte certaines particularités comme la présence d'auto-fécondation, de banques de graines (Sellinger et al. 2020) ou de multifurcations dans la généalogie (U-shaped genome site frequency spectra : challenging the reference model of molecular evolution ?). Ce n'est pas le cas de toutes les méthodes développées. La plupart des méthodes s'appuie sur le coalescent de Kingman pour leur développement. Nos méthodes utilisent des approches basées sur des simulations. Les modèles que l'on peut prendre en compte dépendent donc du simulateur utilisé.

Dans un contexte d'application à des fins de conservation, une autre difficulté rencontrée pour l'application des méthodes d'inférence développées jusqu'à présent est la non prise en compte des spécificités des petites tailles de population. (Il existe également des spécificités pour les populations de grandes tailles comme des fragmentations spatiales fortes, des conditions environnementales variées conduisant à des pressions de sélection distinctes...) En effet, si une espèce est en déclin sa population est certainement de petite taille. Or, une petite taille de population a de nombreux effets différents de ce qui est attendu sous le modèle standard neutre. Par exemple, la reproduction entre individus apparentés étant plus importante quand la population est petite, la valeur sélective de la population diminue : il s'agit de la dépression de consanguinité. Et inversement, s'il existe un mécanisme d'auto-incompatibilité, empêchant de se reproduire avec des organismes trop similaires, il peut être plus difficile de trouver un partenaire. Même sans auto-incompatibilité, quand il y a peu d'individus de la même espèce, les partenaires potentiels sont plus rares. L'espèce se trouve alors dans un vortex d'extinction.

Le développement de méthodes utilisables avec des données de qualité moindre et prenant en compte une large gamme d'écarts au modèle standard neutre classique est donc nécessaire.

5.3.2 Précision de l'information

Cependant, les méthodes d'inférence génétique actuelles permettent d'obtenir des informations très précises sur les populations étudiées. Des scénarios comportant plusieurs changements de tailles de population, des structures de population complexes ou un mélange des deux sont inférables à l'heure actuelle. La connais-

sance fine de l'histoire de la population est un atout indéniable à sa conservation. Les dates auxquelles sont détectables ces évènements sont de l'ordre de N générations. À des fins de conservations, ces méthodes sont donc utilisables pour les populations de grande taille et/ou pour les espèces aux temps de générations courts.

Comme dit précédemment, de plus en plus de méthodes intègrent des modèles différents du modèle de Kingman classique. Il est donc possible de prendre en compte les particularités du cycle de vie/cycle reproducteur de chaque espèce. L'intégration de modèle précis s'approchant le plus possible de la réalité de l'espèce permettrait des inférences très fines de l'histoire de la population. Certains scénarios très différents peuvent produire des effets similaires sur les arbres de coalescence (comme cela peut être le cas entre des scénarios démographiques et des scénarios de structure), il est nécessaire d'avoir une bonne connaissance de la biologie de l'espèce pour inférer sa véritable histoire démographique.

Les méthodes utilisant l'information des évènements de recombinaison étudient de nombreuses histoires évolutives sur un faible nombre de génomes. Elles peuvent donc être utilisées à partir d'un petit échantillon d'individus et nous renseigner sur des populations comportant un faible nombre d'individus ou sur une population dont les individus sont difficilement échantillonnables.

À partir d'un alignement de génomes, il est possible d'utiliser de nombreuses méthodes utilisant des informations (évènement de mutation, évènement de recombinaison) et des statistiques différentes. Comme on a pu le voir au cours de cette thèse, ces informations et ces statistiques ne sont pas affectées par les mêmes facteurs. L'utilisation jointe de ces différentes méthodes permet donc une inférence robuste de l'histoire évolutive de la population.

5.3.3 Le futur de la génomique de la conservation ?

La génomique de la conservation peut être décrite comme l'application d'analyses génomiques pour la préservation des populations et de la biodiversité. Les méthodes génomiques peuvent être utilisées pour aider à la délimitation d'espèces, connaître leur diversité génétique, leur histoire démographique ainsi que la taille de leur population. Actuellement, ce n'est pas tout à fait le cas. S'il existe des méthodes génomiques utilisées pour aider aux décisions de conservation, elles sont majoritairement centrées sur la délimitation d'espèce et sur la description de la diversité génétique de la population (Supple and Shapiro 2018; Rossetto et al. 2021). Peu de méthodes génomiques sont utilisées pour inférer l'histoire démographique de populations à des fins de conservation.

Les termes *conservation genetics* et *conservation genomics* sont principalement retrouvés dans des articles traitant de génétique et d'hérédité ou d'écologie et moins souvent dans des articles de biologie de la conservation (Fig 5.1). La génomique de la conservation semble être encore un concept principalement théorique.

Cependant, les travaux en inférences démographiques s'élargissent de plus en plus à des espèces diverses, aux cycles de vie de plus en plus éloignés du modèle standard neutre et essayent de prendre en compte les différentes limites techniques

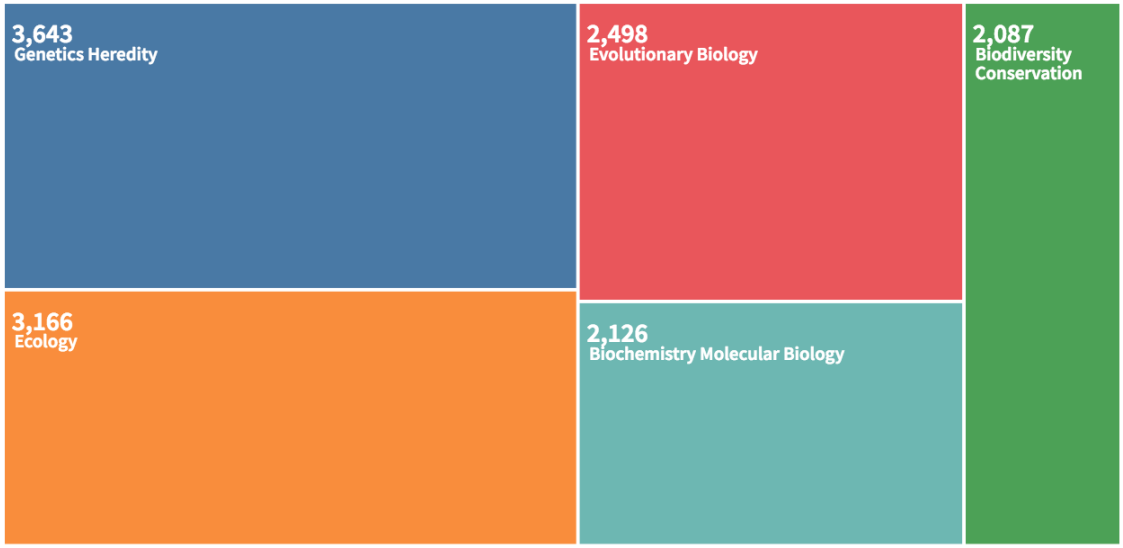


Figure 5.1: Nombres d'articles par catégorie Web of Science possédant les termes *conservation genetics* ou *conservation genomics* dans leur titre, résumé ou sujet, parmi 13,009 publications comptabilisées. Chiffres récoltés sur webofscience.com le 15 septembre 2021.

que le séquençage d'ADN peut rencontrer. Celui-ci voit son coût baisser et sa qualité s'améliorer d'années en années.

De plus, de nombreux projets et centres de recherche avec le but de mettre la génomique au service de la conservation, se sont crées ces dernières années.

La génomique de la conservation a donc le potentiel pour devenir un des outils majeurs à la prise de décision en matière de conservation.

Chapitre 6

Annexes

6.1 Expectation and variance of linkage disequilibrium in a fixed tree

Expectation and variance of linkage disequilibrium in a fixed tree

Amaury Lambert

November 11, 2020

1 Introduction and notation

Let T be a (not necessarily binary) tree with n leaves labelled by $\{1,\dots,n\}$ and let $\mathcal{E}$ be its edge set (neglecting the presence of a root edge, because mutations occurring on a root edge are not segregating). We introduce the notation $\prec$, where

$$x \prec y \iff y \text{ is in the descendance of } x,$$

for any x, y may be vertices or edges of T.

We see T as the genealogy of n haploid individuals. Differences between the DNA sequences of these n individuals are due to mutations falling on T. We assume that each site of the sequence can be hit by **at most one mutation** (infinite-site model). Then a site that has been hit by a mutation is bi-allelic and the carriers of the mutant base form a **subtree of T, which is the set $S(e)$ of leaves descending from the edge e where the mutation has fallen**:

$$S(e) := \{i \in \{1,\dots,n\} : e \prec i\}.$$

We assume that for each site hit by a mutation, the mutation falls on a random edge of $\mathcal{E}$ according to **the same probability distribution** $p : \mathcal{E} \to [0,1]$. From now on, it will be convenient to think of $p(e)$ **as the length of** e, as when mutations occur as homogeneous Poisson point processes on the tree. In particular, **the total length of T (sum of edge lengths) is equal to 1** and T is endowed with the natural distance between vertices induced by edge lengths. We also assume that **mutations occurring at different sites are independent** (conditional on T).

We are interested in a **pair of bi-allelic sites**, where the ancestral state is a for the first site, b for the second site and the mutant state is A for the first site and B for the second site. **We let $f_A(T)$ (resp. $f_B(T)$, resp. $f_{AB}(T)$) denote the frequency of carriers of A (resp. B, resp. AB) in the leaf set of T.**

The **linkage disequilibrium** between these two sites is measured by the statistic

$$D(T) := f_{AB}(T) - f_A(T) f_B(T).$$

Because mutations at different sites are independent conditional on T and identically distributed, the expected linkage disequilibrium (LD) is

$$\mathbb{E}(D(T)) = \mathbb{E}(f_{AB}(T)) - (q(T))^2,$$

where $q(T) := \mathbb{E}(f_A(T))$.

For any leaf $i = 1, \ldots, n$, we denote by H_i **the height of** i, or **distance between the root and** i, that is

$$H_i := \sum_{e \in \mathcal{E}: e \prec i} p(e).$$

We will say that T is **ultrametric** when all leaves are at the same distance from the root, that is when all the H_i's are equal, in which case we denote their common value by $H(T)$**, called height of** T.

2 Expectation of LD

Proposition 2.1. *Recall that $q(T)$ is the expected frequency of a mutation falling on T according to p. Then*

$$q(T) = \frac{1}{n} \sum_{i=1}^{n} H_i. \tag{1}$$

In the ultrametric case, $q(T) = H(T)$.

Proof. The frequency of a mutation falling on edge e is $\#S(e)/n$, so that

$$q(T) = \sum_{e \in \mathcal{E}} p(e) \frac{\#S(e)}{n} = \frac{1}{n} \sum_{e \in \mathcal{E}} \sum_{i: e \prec i} p(e) = \frac{1}{n} \sum_{i=1}^{n} \sum_{e: e \prec i} p(e) = \frac{1}{n} \sum_{i=1}^{n} H_i,$$

which ends the proof. □

Proposition 2.2. *Recall that $f_{AB}(T)$ is the frequency of the double mutant, for two mutations falling independently on T according to p. Then*

$$\mathbb{E}(f_{AB}(T)) = \frac{1}{n} \sum_{i=1}^{n} H_i^2. \tag{2}$$

In the ultrametric case, $\mathbb{E}(f_{AB}(T)) = H(T)^2$.

Proof. If the mutation at the first site and the mutation at the second site fall on edges e_1 and e_2 respectively, then

$$f_{AB}(T) = \begin{cases} \#S(e_1)/n & \text{if } e_2 \prec e_1 \\ \#S(e_2)/n & \text{if } e_1 \prec e_2 \\ 0 & \text{otherwise.} \end{cases}$$

As a consequence, reasoning similarly as in the proof of the previous statement,

$$\begin{aligned}\mathbb{E}(f_{AB}(T)) &= \sum_{e\in\mathcal{E}} p(e)^2\,\frac{\#S(e)}{n} + 2\sum_{e\in\mathcal{E}} p(e) \sum_{e'\neq e, e'\prec e} p(e')\frac{\#S(e)}{n} \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{e\in\mathcal{E}:e\prec i} p(e)^2 + 2\sum_{e\in\mathcal{E}:e\prec i} p(e) \sum_{e'\neq e, e'\prec e} p(e')\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{e\in\mathcal{E}:e\prec i} p(e) \sum_{e'\in\mathcal{E}:e'\prec i} p(e')\right) = \frac{1}{n}\sum_{i=1}^{n} H_i^2,\end{aligned}$$

which ends the proof. □

Corollary 2.3. *As a consequence of the previous two propositions, the expected linkage disequilibrium in tree T is equal to*

$$\mathbb{E}(D(T)) = \mathbb{E}(f_{AB}(T)) - q(T)^2 = \frac{1}{n}\sum_{i=1}^{n} H_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} H_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(H_i - \overline{H}\right)^2, \quad (3)$$

where $\overline{H} = \frac{1}{n}\sum_{j=1}^{n} H_j$. Note that $\mathbb{E}(D(T)) > 0$ except in the ultrametric case where

$$\mathbb{E}(D(T)) = 0. \quad (4)$$

3 Variance of LD

We have to compute three quantities:

$$\mathbb{E}(f_A^2), \quad \mathbb{E}(f_{AB}^2) \quad \text{and } \mathbb{E}(f_{AB}f_A f_B).$$

We will need to introduce, for every pair of leaves i and j, the **height $H_{i\wedge j}$ of the vertex $i\wedge j$ which denotes the most recent common ancestor to i and j**:

$$H_{i\wedge j} := \sum_{e:e\prec i,j} p(e).$$

Note that

$$H_{i\wedge j} = \frac{1}{2}(H_i + H_j - d(i,j)). \quad (5)$$

Also set

$$H_{i\wedge j\wedge k} := \sum_{e:e\prec i,j,k} p(e).$$

Lemma 3.1. *We have*

$$\mathbb{E}(f_A^2) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} H_{i\wedge j},$$

$$\mathbb{E}(f_{AB}^2) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} H_{i\wedge j}^2,$$

and

$$\mathbb{E}(f_{AB}f_Af_B) = \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} H_{i\wedge j\wedge k}(2H_{i\wedge j} - H_{i\wedge j\wedge k}),$$

which simplifies, in the ultrametric case, into

$$\mathbb{E}(f_{AB}f_Af_B) = \frac{1}{n^3}\sum_{i=1}^{n}\left(\sum_{j=1}^{n} H_{i\wedge j}\right)^2.$$

Proof. First,

$$\mathbb{E}(f_A^2) = \sum_{e\in\mathcal{E}} p(e)\left(\frac{\#S(e)}{n}\right)^2 = \frac{1}{n^2}\sum_{e\in\mathcal{E}}\sum_{i:e\prec i}\sum_{j:e\prec j} p(e) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{e:e\prec i,j} p(e),$$

which yields the first result. Similarly,

$$\begin{aligned}
\mathbb{E}(f_{AB}^2) &= \sum_{e\in\mathcal{E}} p(e)^2\left(\frac{\#S(e)}{n}\right)^2 + 2\sum_{e\in\mathcal{E}} p(e)\sum_{e'\neq e, e'\prec e} p(e')\left(\frac{\#S(e)}{n}\right)^2 \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\sum_{e:e\prec i,j} p(e)^2 + 2\sum_{e:e\prec i,j} p(e)\sum_{e'\neq e,e'\prec e} p(e')\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\sum_{e:e\prec i,j} p(e)\sum_{e':e'\prec i,j} p(e')\right),
\end{aligned}$$

which yields the second result. Finally,

$$\begin{aligned}
\mathbb{E}(f_{AB}f_Af_B) &= \sum_{e\in\mathcal{E}} p(e)^2\left(\frac{\#S(e)}{n}\right)^3 + 2\sum_{e\in\mathcal{E}} p(e)\sum_{e'\neq e, e'\prec e} p(e')\left(\frac{\#S(e)}{n}\right)^2\frac{\#S(e')}{n} \\
&= \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\left(\sum_{e:e\prec i,j,k} p(e)^2 + 2\sum_{e:e\prec i,j} p(e)\sum_{e'\neq e,e'\prec e,e'\prec k} p(e')\right) \\
&= \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\left(2\sum_{e:e\prec i,j} p(e)\sum_{e':e'\prec i,j,k} p(e') - \sum_{e:e\prec i,j,k} p(e)\sum_{e':e'\prec i,j,k} p(e')\right),
\end{aligned}$$

which yields

$$\mathbb{E}(f_{AB}f_Af_B) = \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} H_{i\wedge j\wedge k}(2H_{i\wedge j} - H_{i\wedge j\wedge k}).$$

Let us try to simplify this last expression in the ultrametric case. First note that for any three leaves labelled i, j, k, if we set

$$a_{ijk} := \min\{H_{i\wedge j}, H_{i\wedge k}, H_{j\wedge k}\} \quad \text{and} \quad A_{ijk} := \max\{H_{i\wedge j}, H_{i\wedge k}, H_{j\wedge k}\},$$

then the three quantities $H_{i\wedge j}$, $H_{i\wedge k}$ and $H_{j\wedge k}$ take their values in $\{a_{ijk}, A_{ijk}\}$, and two of these three quantities take the value a_{ijk}, which is also equal to $H_{i\wedge j\wedge k}$. Now set

$$\Sigma_1 := \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n H_{i\wedge j} H_{j\wedge k} = \frac{1}{3} \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n (H_{i\wedge j} H_{i\wedge k} + H_{i\wedge j} H_{j\wedge k} + H_{i\wedge k} H_{j\wedge k})$$

and

$$\Sigma_2 := \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n H_{i\wedge j} H_{i\wedge j\wedge k} = \frac{1}{3} \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n H_{i\wedge j\wedge k} \left(H_{i\wedge j} + H_{j\wedge k} + H_{i\wedge k}\right).$$

Now observe that

$$H_{i\wedge j} H_{i\wedge k} + H_{i\wedge j} H_{j\wedge k} + H_{i\wedge k} H_{j\wedge k} = a_{ijk}(a_{ijk} + 2A_{ijk})$$

and

$$H_{i\wedge j\wedge k} \left(H_{i\wedge j} + H_{j\wedge k} + H_{i\wedge k}\right) = a_{ijk}(2a_{ijk} + A_{ijk}).$$

As a consequence,

$$2\Sigma_2 - \Sigma_1 = \frac{1}{3} \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n 3a_{ijk}^2 = \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n H_{i\wedge j\wedge k}^2.$$

This shows that

$$\sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n H_{i\wedge j\wedge k}(2H_{i\wedge j} - H_{i\wedge j\wedge k}) = \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n H_{i\wedge j} H_{j\wedge k} = \sum_{j=1}^n \left(\sum_{i=1}^n H_{i\wedge j}\right)^2,$$

which finishes the proof. □

We can now use the lemma to compute the second moment of D.

Proposition 3.2. *In the general case, we have*

$$\mathbb{E}(D^2) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n H_{i\wedge j}^2 + \left(\frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n H_{i\wedge j}\right)^2 - \frac{2}{n^3} \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n H_{i\wedge j\wedge k}(2H_{i\wedge j} - H_{i\wedge j\wedge k}).$$

In the case when T is ultrametric, this expression becomes (recall $\mathbb{E}(D) = 0$ in this case)

$$\mathbb{E}(D^2) = \mathit{Var}(D) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n H_{i\wedge j}^2 + \left(\frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n H_{i\wedge j}\right)^2 - \frac{2}{n^3} \sum_{i=1}^n \left(\sum_{j=1}^n H_{i\wedge j}\right)^2, \quad (6)$$

which also reads

$$Var(D) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(H_{i\wedge j} - \overline{H}_i)(H_{i\wedge j} - \overline{H}_j), \tag{7}$$

where we have put for each $i \in \{1, \dots, n\}$,

$$\overline{H}_i := \frac{1}{n}\sum_{j=1}^{n} H_{i\wedge j}.$$

Remark 3.3. *Thanks to* (5)*, in the ultrametric case we have* $H_{i\wedge j} = H - \frac{1}{2}d(i,j)$*. Injecting this into* (7)*, we get*

$$Var(D) = \frac{1}{4n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(d(i,j) - \overline{d}_i)(d(i,j) - \overline{d}_j), \tag{8}$$

where $\overline{d}_i = \frac{1}{n}\sum_j d(i,j)$*. Note that* $d(i,j)$ *is equal to twice the coalescence time between leaves* i *and* j *when time is scaled so that the total length of the tree is 1. This may indicate a lead to compute the variance of* D *integrating over the tree* T*, when* T *is a time-inhomogeneous coalescent.*

Proof. By definition,

$$\mathbb{E}(D^2) = \mathbb{E}((f_{AB} - f_A f_B)^2) = \mathbb{E}(f_{AB}^2) + \mathbb{E}((f_A f_B)^2) - 2\mathbb{E}(f_{AB} f_A f_B).$$

Since

$$\mathbb{E}((f_A f_B)^2) = \mathbb{E}(f_A^2)\mathbb{E}(f_B^2) = (\mathbb{E}(f_A^2))^2,$$
$$\mathbb{E}(D^2) = \mathbb{E}(f_{AB}^2) + (\mathbb{E}(f_A^2))^2 - 2\mathbb{E}(f_{AB} f_A f_B).$$

Thanks to the three expressions stated in the lemma, we get the first two expressions displayed in the proposition. Now using the notation $\overline{H}_i$, the expression in the ultrametric case reads

$$\mathbb{E}(D^2) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} H_{i\wedge j}^2 + \left(\frac{1}{n}\sum_{i=1}^{n}\overline{H}_i\right)^2 - \frac{2}{n}\sum_{i=1}^{n}\overline{H}_i^2.$$

Now the second expression displayed in the statement equals

$$\begin{aligned}
\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(H_{i\wedge j} - \overline{H}_i)(H_{i\wedge j} - \overline{H}_j) &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} H_{i\wedge j}^2 - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\overline{H}_i H_{i\wedge j} \\
&- \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} H_{i\wedge j}\overline{H}_j + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\overline{H}_i\overline{H}_j \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} H_{i\wedge j}^2 - \frac{1}{n}\sum_{i=1}^{n}\overline{H}_i^2 \\
&- \frac{1}{n^2}\sum_{j=1}^{n}\overline{H}_j^2 + \left(\frac{1}{n}\sum_{i=1}^{n}\overline{H}_i\right)^2,
\end{aligned}$$

which indeed equals $\mathbb{E}(D^2)$. □

4 Discussion around (7)

A consequence of (7) is that $\sum_{i=1}^n \sum_{j=1}^n (H_{i\wedge j} - \overline{H}_i)(H_{i\wedge j} - \overline{H}_j)$ is always non-negative. This brings the question of proving independently the following general result.

Proposition 4.1. *For any real, symmetric matrix M with generic element m_{ij},*

$$\sum_{i=1}^n \sum_{j=1}^n (m_{ij} - \overline{m}_i)(m_{ij} - \overline{m}_j) \geq 0,$$

where $\overline{m}_i = \frac{1}{n}\sum_j m_{ij}$.

Proof. Since M is real and symmetric, there is an orthogonal matrix P ($^tP = P^{-1}$) and a diagonal matrix $D = \mathrm{Diag}(\lambda_1, \ldots, \lambda_n)$ such that $M = PDP^{-1}$. Then writing $A = PD$, we get

$$a_{ij} = \sum_k p_{ik} d_{kj} = p_{ij} d_{jj} = p_{ij}\lambda_j.$$

Next, since $M = A\,{}^tP$,

$$m_{ij} = \sum_k a_{ik} p_{jk} = \sum_k p_{ik}\lambda_k p_{jk}.$$

In particular,

$$\overline{m}_i = \frac{1}{n}\sum_j m_{ij} = \frac{1}{n}\sum_j \sum_k p_{ik}\lambda_k p_{jk} = \frac{1}{n}\sum_k p_{ik}\lambda_k \sum_j p_{jk} = \sum_k p_{ik}\lambda_k q_k,$$

where we have set

$$q_k := \frac{1}{n}\sum_j p_{jk}.$$

Note that $\sum_j p_{jk}$ is the entry at row k of ${}^tPv$, where v is the vector with only ones. Then

$$n = {}^t vv = {}^t vP\ {}^tPv = \sum_k \left(\sum_j p_{jk}\right)^2 = \sum_k n^2 q_k^2,$$

which we record as

$$\sum_k n q_k^2 = 1. \tag{9}$$

Then we have

$$m_{ij} - \overline{m}_i = \sum_k p_{ik}\lambda_k (p_{jk} - q_k).$$

Next, defining

$$S := \sum_{i=1}^n \sum_{j=1}^n (m_{ij} - \overline{m}_i)(m_{ij} - \overline{m}_j),$$

we can write

$$\begin{aligned} S &= \sum_i \sum_j \sum_k p_{ik}\lambda_k(p_{jk}-q_k)\sum_\ell p_{j\ell}\lambda_\ell(p_{i\ell}-q_\ell) \\ &= \sum_k \sum_\ell \lambda_k\lambda_\ell \sum_i p_{ik}(p_{i\ell}-q_\ell)\sum_j p_{j\ell}(p_{jk}-q_k) \\ &= \sum_k \sum_\ell \lambda_k\lambda_\ell b_{k\ell}b_{\ell k}, \end{aligned}$$

where

$$b_{k\ell} := \sum_i p_{ik}(p_{i\ell}-q_\ell) = \sum_i p_{ik}p_{i\ell} - q_\ell \sum_i p_{ik} = \delta_{kl} - nq_kq_\ell,$$

because $\sum_i p_{ik}p_{i\ell}$ is the entry (k,ℓ) of ${}^tPP = I_n$. Finally we get

$$S = \sum_k \sum_\ell \lambda_k\lambda_\ell(\delta_{kl} - nq_kq_\ell)^2.$$

We can rewrite it as follows

$$S = \sum_k \sum_\ell \lambda_k\lambda_\ell(-nq_kq_\ell)^2 + \sum_k \lambda_k^2((1-nq_k^2)^2 - (nq_k^2)^2),$$

that is,

$$S = \left(\sum_k \lambda_k n q_k^2\right)^2 + \sum_k \lambda_k^2(1-2nq_k^2).$$

Now thanks to (9), writing $\alpha_k = nq_k^2$, we have

$$S = \left(\sum_k \alpha_k\lambda_k\right)^2 + \sum_k \lambda_k^2(1-2\alpha_k),$$

where the α_k's are non-negative and sum to 1. The proof ends thanks to the following result. □

Lemma 4.2. *For any integer n, for any non-negative $(\alpha_k)_{k=1,\dots,n}$ such that $\sum_{k=1}^n \alpha_k = 1$, for any real numbers $\lambda_1,\dots,\lambda_n$,*

$$\left(\sum_k \alpha_k\lambda_k\right)^2 + \sum_k \lambda_k^2(1-2\alpha_k) \geq 0.$$

Proof. We reason by induction on n. The quantity is zero when $n=1$. Now let $n \geq 2$. We are going to prove that for any α_k's such that $\sum_{k=1}^{n-1}\alpha_k = 1$, for any $t \in [0,1]$ and λ_n, we have $F_t(\lambda_n) \geq 0$, where

$$F_t(\lambda_n) = \left(\sum_{k=1}^{n-1}\alpha_k(1-t)\lambda_k + t\lambda_n\right)^2 + \sum_{k=1}^{n-1}\lambda_k^2(1-2\alpha_k(1-t)) + \lambda_n^2(1-2t).$$

Note that t plays the role of α_n and $\alpha_k(1-t)$ of α_k for $k \leq n-1$. Also note that $F_1(\lambda_n) = \sum_{k=1}^{n-1} \lambda_k^2 \geq 0$, so we can assume that $t \neq 1$. Now note that $F_t(\lambda) = a(t)\lambda^2 + 2b(t)\lambda + c(t)$, where

$$a(t) = (1-t)^2,$$

$$b(t) = t(1-t)\sum_{k=1}^{n-1} \alpha_k \lambda_k,$$

and

$$c(t) = \left(\sum_{k=1}^{n-1} \alpha_k(1-t)\lambda_k\right)^2 + \sum_{k=1}^{n-1} \lambda_k^2(1 - 2\alpha_k(1-t)).$$

Since $t \neq 1$, $a(t) > 0$ and we only need to show that for any $t \in [0,1)$, $b(t)^2 - a(t)c(t) \leq 0$. Writing $A = \sum_{k=1}^{n-1} \alpha_k \lambda_k$, we have

$$\begin{aligned} b(t)^2 - a(t)c(t) &= t^2(1-t)^2A^2 - (1-t)^2\left[(1-t)^2A^2 + \sum_{k=1}^{n-1} \lambda_k^2(1 - 2\alpha_k(1-t))\right] \\ &= (1-t)^2 g(t), \end{aligned}$$

where

$$g(t) = (2t-1)A^2 - \sum_{k=1}^{n-1} \lambda_k^2(1 - 2\alpha_k(1-t)).$$

Then we only need to show that $g(t) \leq 0$ for all $t \in [0,1)$. Since g is affine, it is sufficient to show that $g(0) \leq 0$ and $g(1) \leq 0$. Now

$$g(0) = -\left(\sum_{k=1}^{n-1} \alpha_k \lambda_k\right)^2 - \sum_{k=1}^{n-1} \lambda_k^2(1 - 2\alpha_k),$$

which is nonpositive by the induction hypothesis. Finally,

$$g(1) = \left(\sum_{k=1}^{n-1} \alpha_k \lambda_k\right)^2 - \sum_{k=1}^{n-1} \lambda_k^2.$$

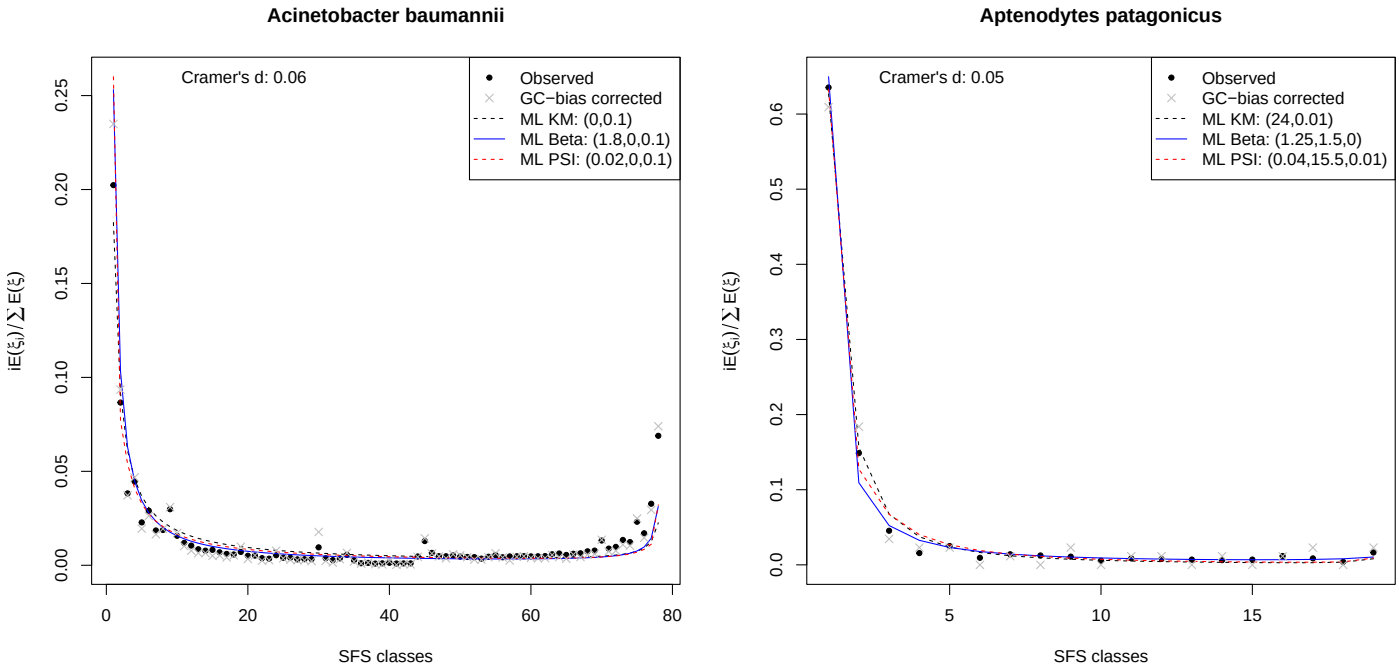
By Jensen's inequality,

$$\left(\sum_{k=1}^{n-1} \alpha_k \lambda_k\right)^2 \leq \sum_{k=1}^{n-1} \alpha_k \lambda_k^2 \leq \sum_{k=1}^{n-1} \lambda_k^2,$$

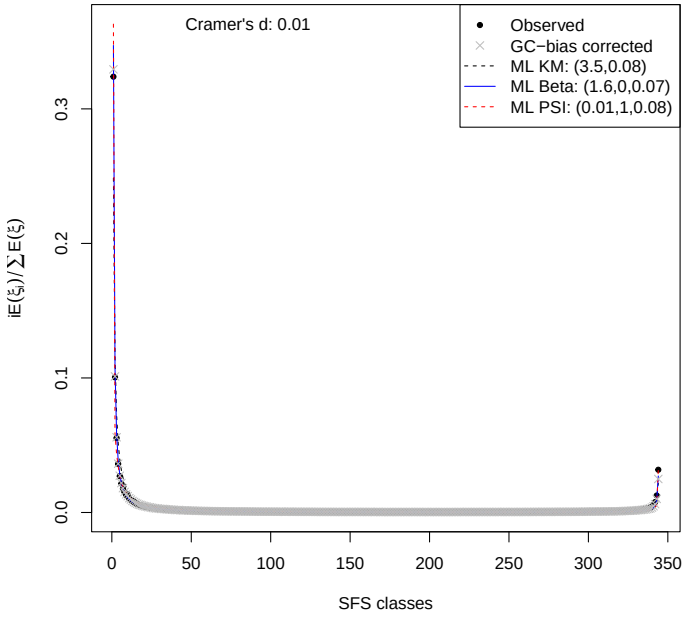
which shows that $g(1) \leq 0$ and terminates the proof. □

6.2 Supplementary files of *U-shaped genome site frequency spectra : challenging the reference model of molecular evolution ?*

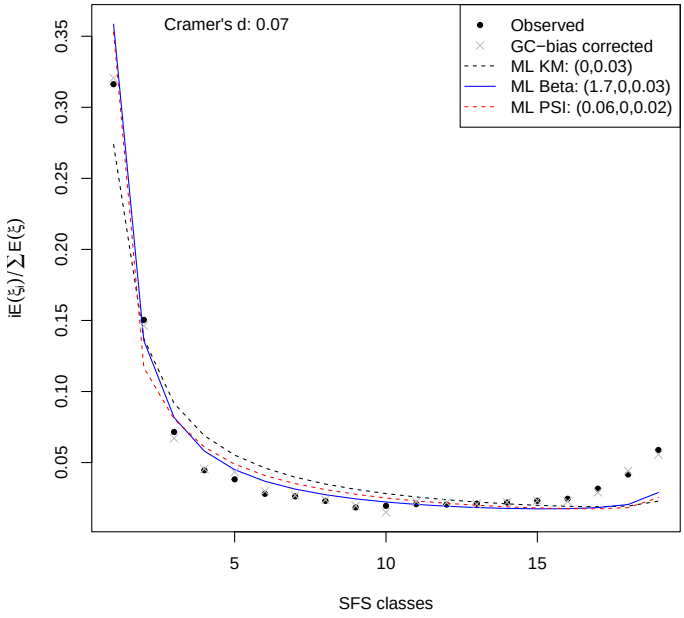
6.2.1 Supplementary file 1



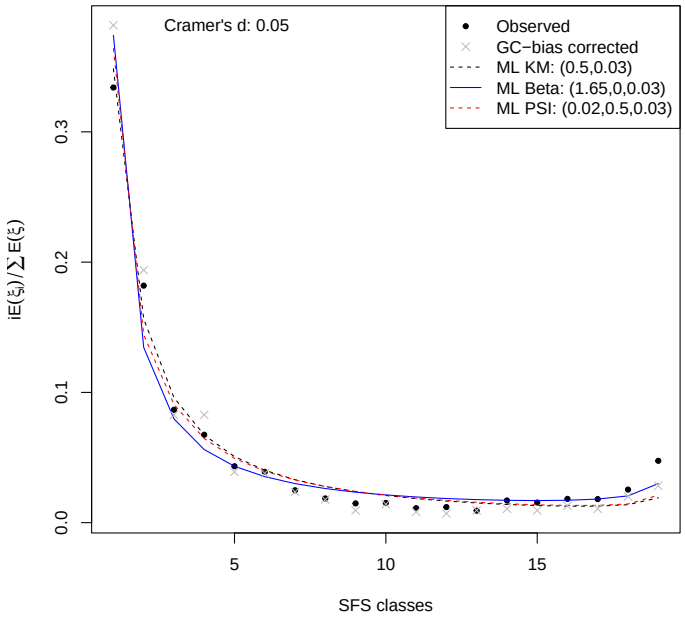
Arabidopsis thaliana



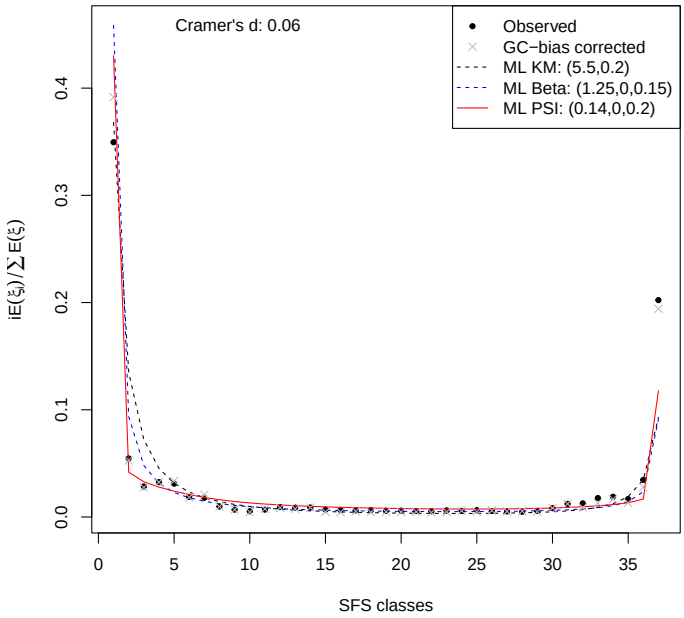
Armadillidium vulgare



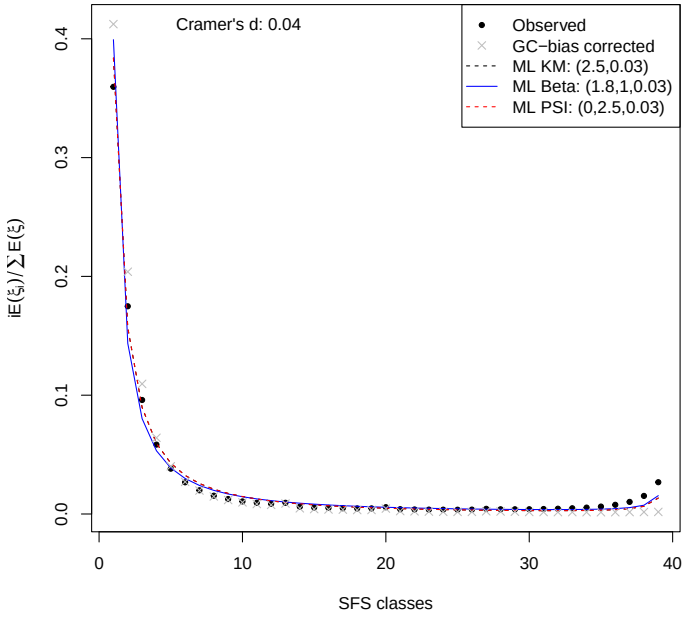
Artemia franciscana



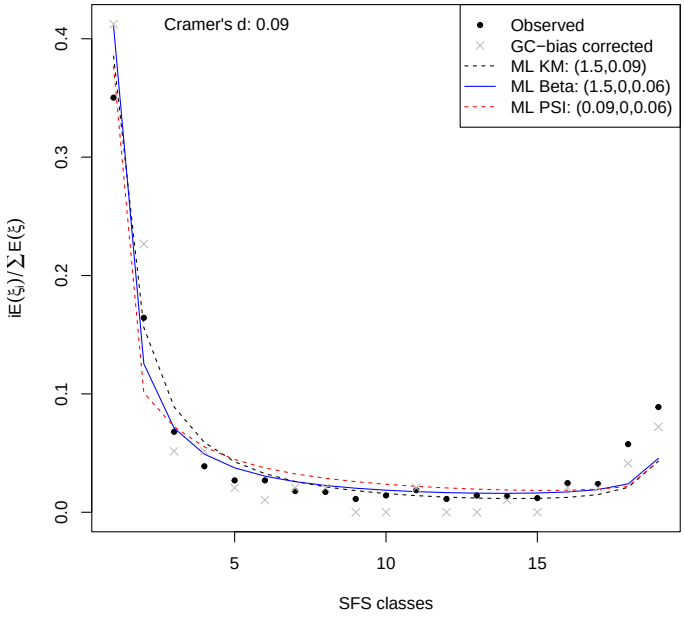
Bacillus subtilis



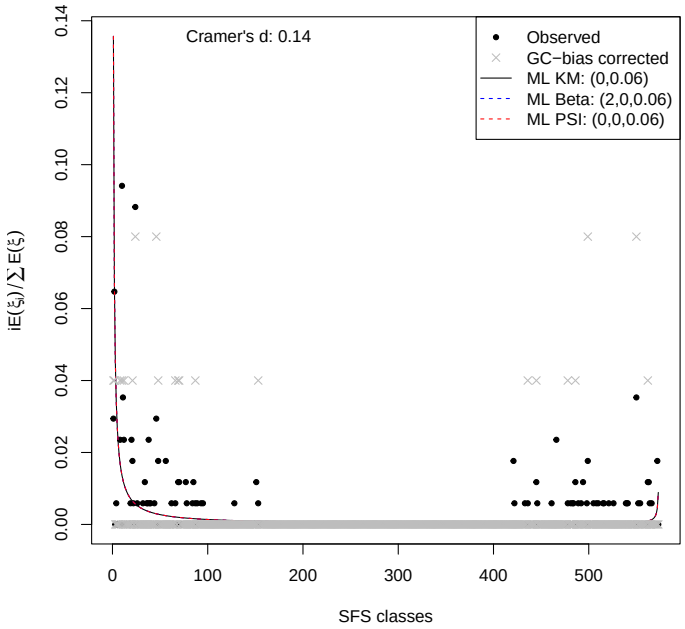
Athene cunicularia



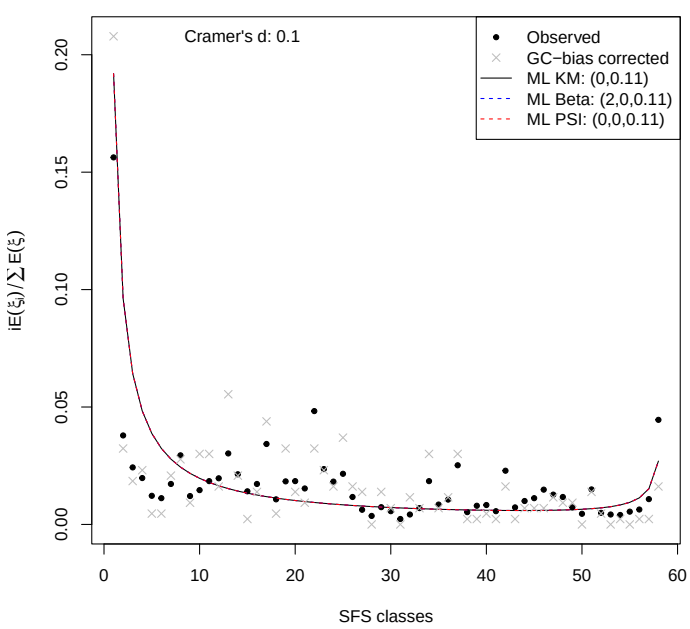
Caenorhabditis brenneri

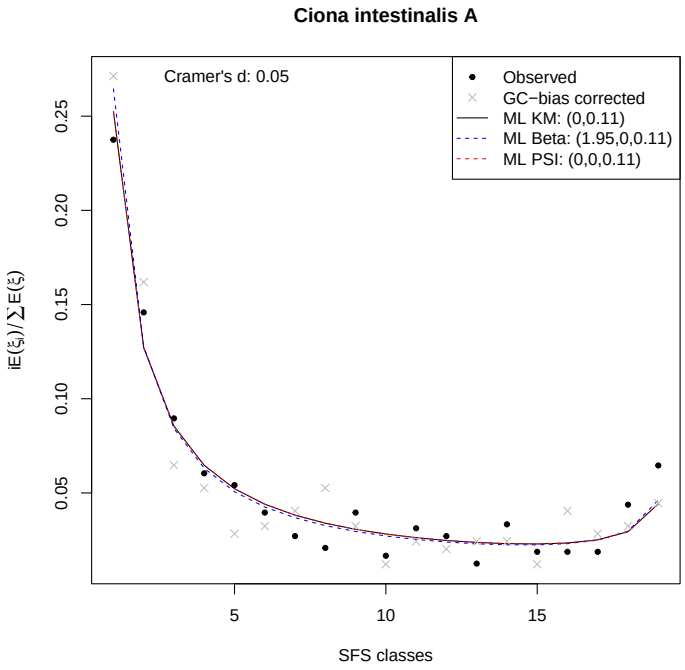


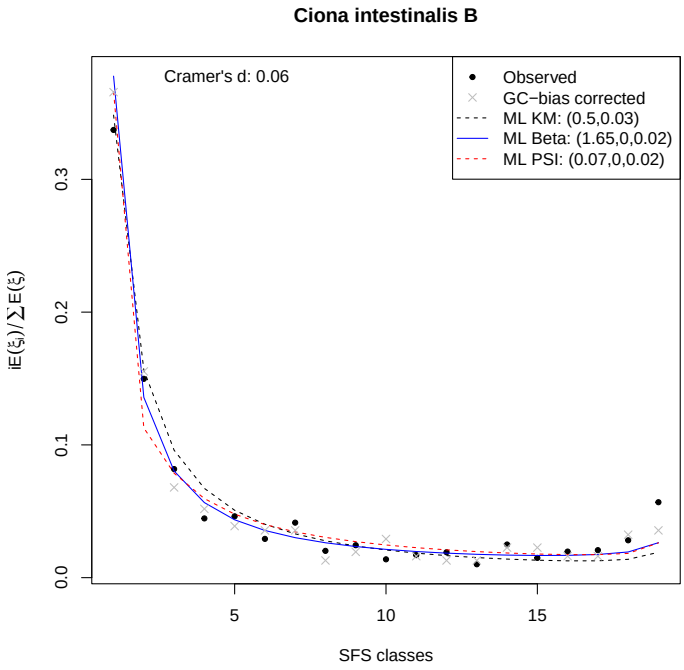
Caenorhabditis elegans

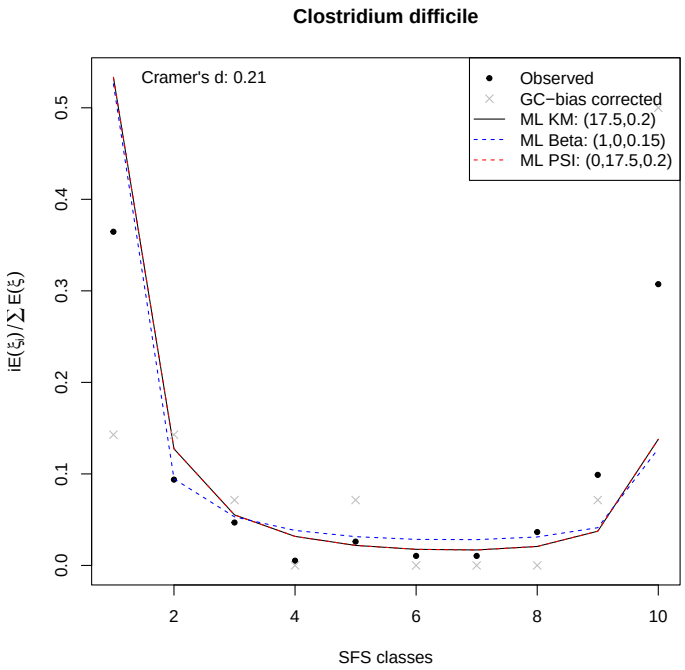


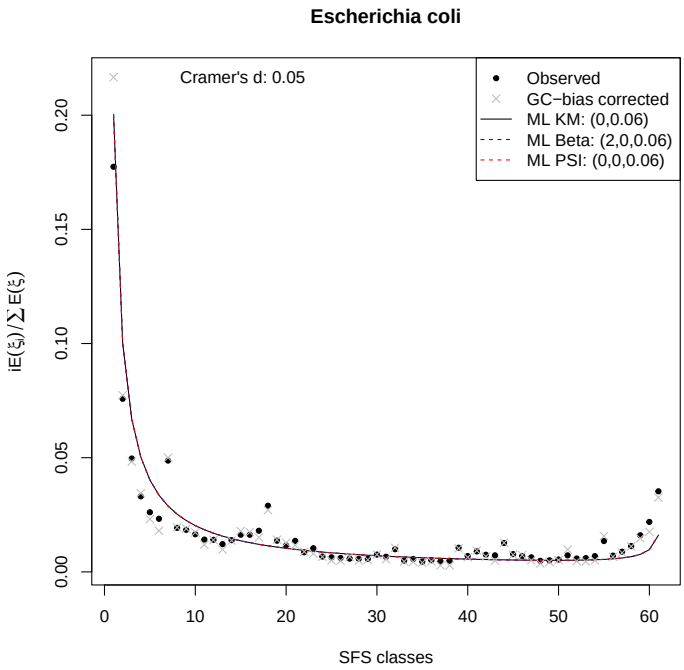
Chlamydia trachomatis



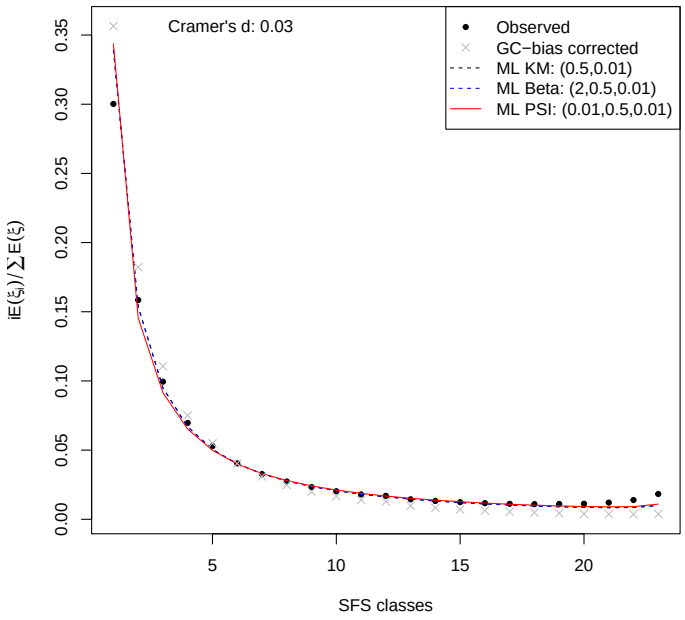

Ciona intestinalis A
Cramer's d: 0.05
Observed
GC−bias corrected
ML KM: (0,0.11)
ML Beta: (1.95,0,0.11)
ML PSI: (0,0,0.11)
iE(ξi)/∑E(ξ)
SFS classes


Ciona intestinalis B
Cramer's d: 0.06
Observed
GC−bias corrected
ML KM: (0.5,0.03)
ML Beta: (1.65,0,0.02)
ML PSI: (0.07,0,0.02)
iE(ξi)/∑E(ξ)
SFS classes

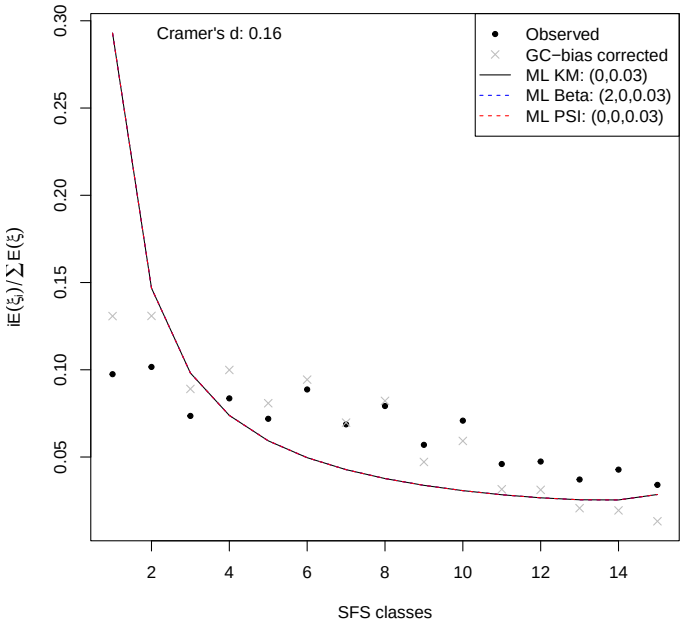

Clostridium difficile
Cramer's d: 0.21
Observed
GC−bias corrected
ML KM: (17.5,0.2)
ML Beta: (1,0,0.15)
ML PSI: (0,17.5,0.2)
iE(ξi)/∑E(ξ)
SFS classes


Escherichia coli
Cramer's d: 0.05
Observed
GC−bias corrected
ML KM: (0,0.06)
ML Beta: (2,0,0.06)
ML PSI: (0,0,0.06)
iE(ξi)/∑E(ξ)
SFS classes

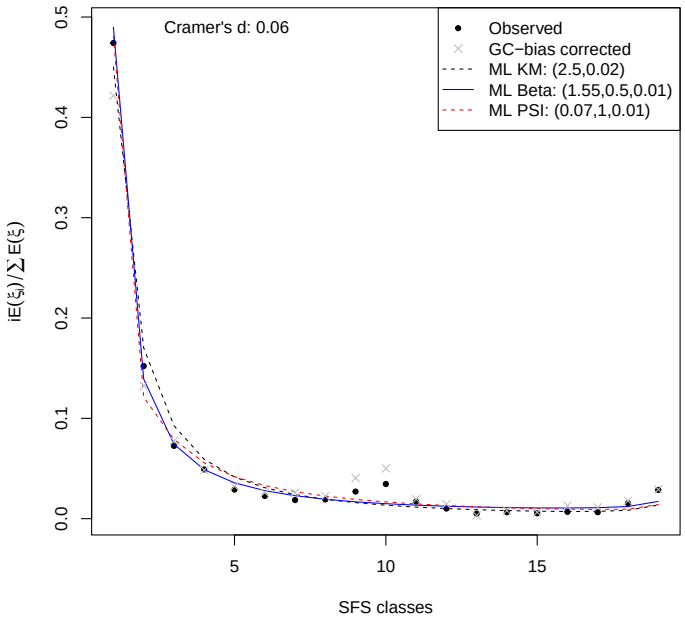
Ficedula albicollis



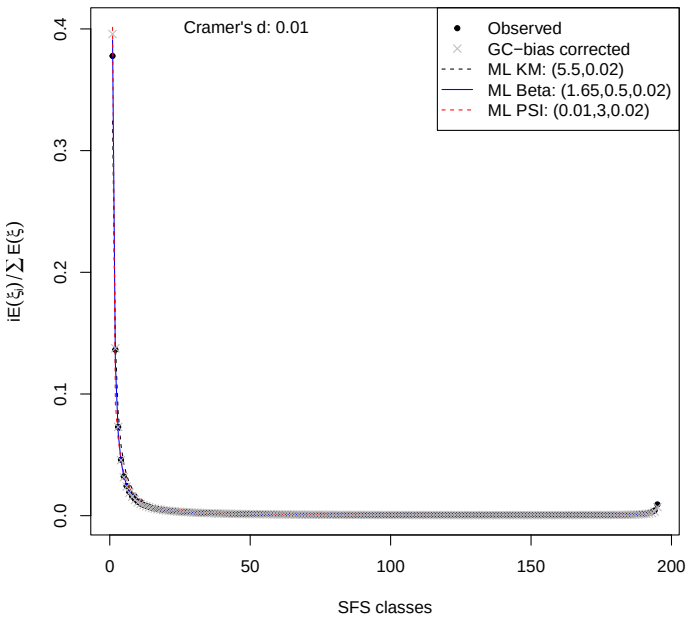
Nipponia nippon

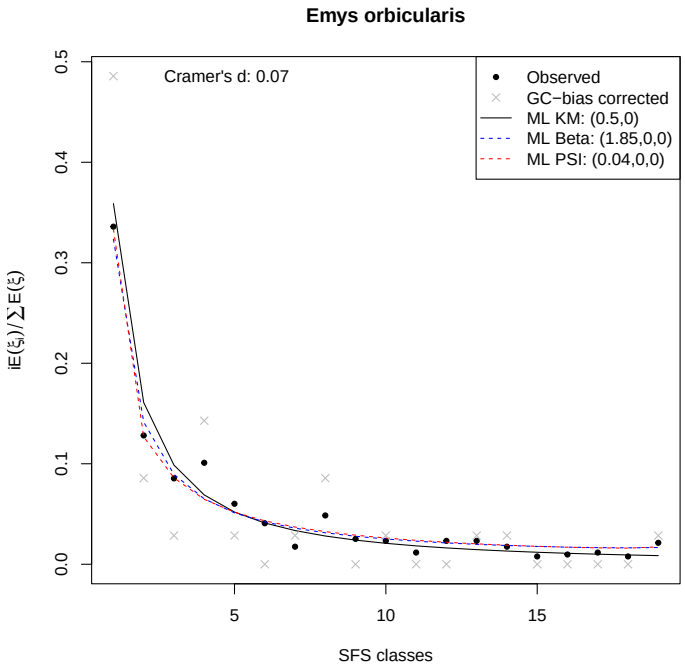


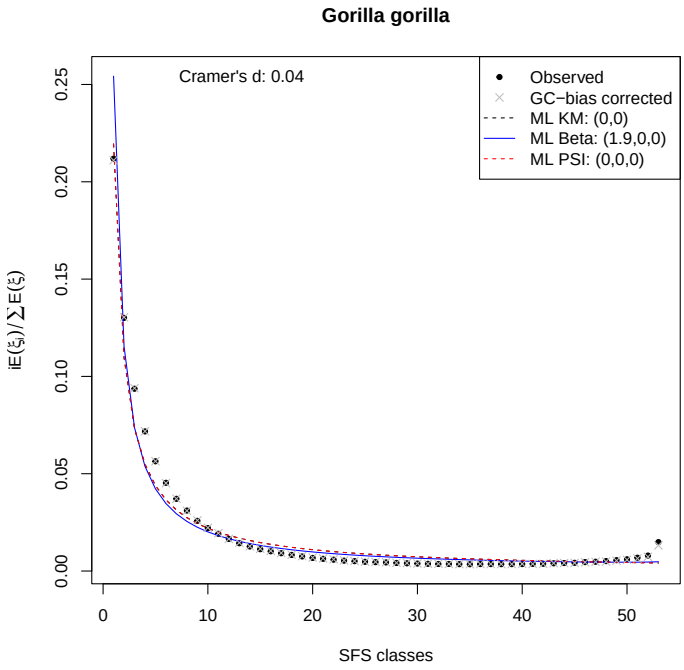
Culex pipiens

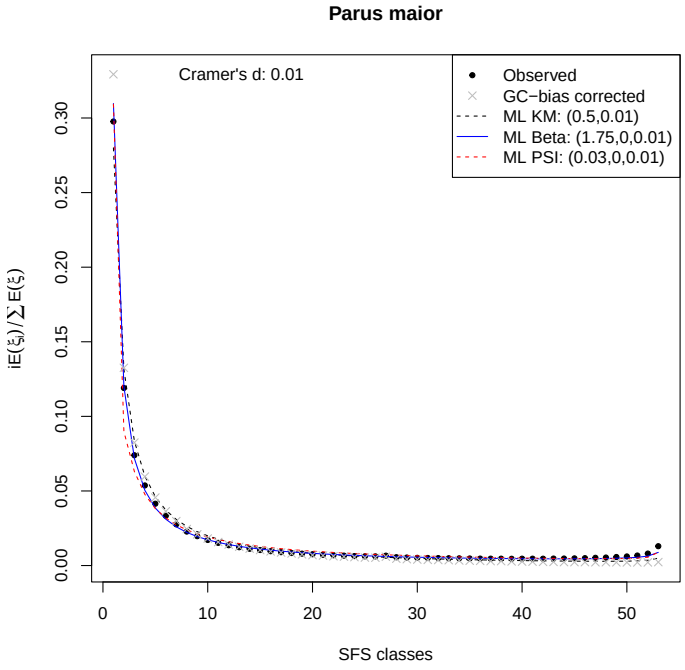


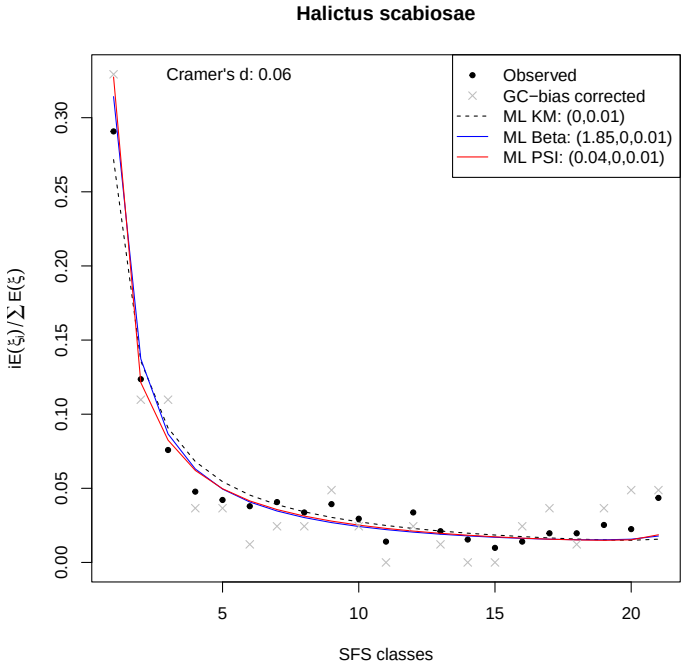
Drosophila melanogaster

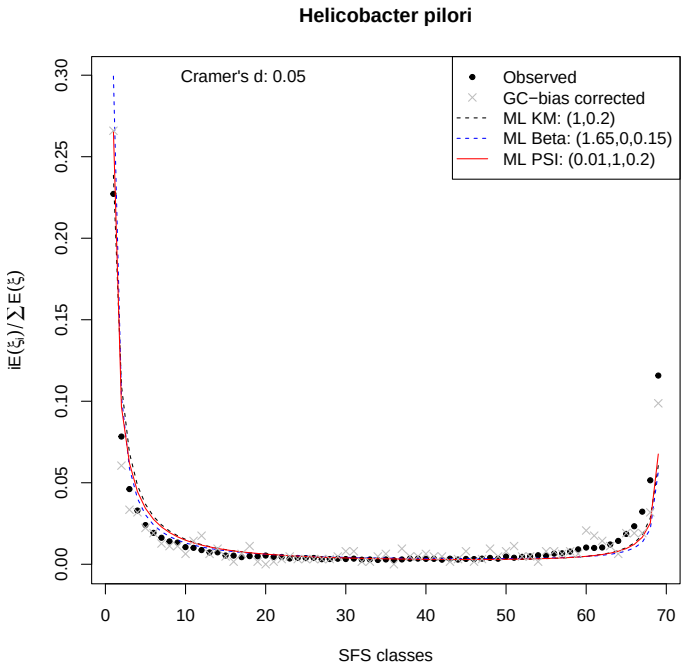


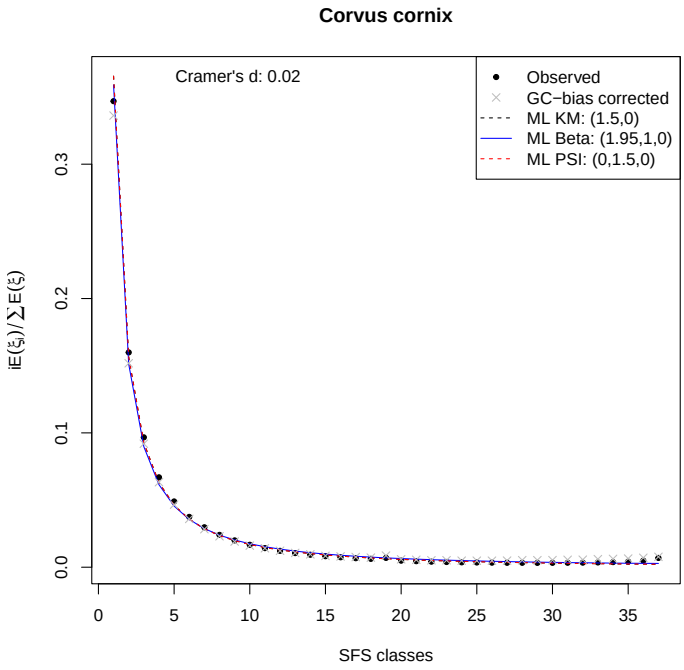

Emys orbicularis
Cramer's d: 0.07
Observed
GC−bias corrected
ML KM: (0.5,0)
ML Beta: (1.85,0,0)
ML PSI: (0.04,0,0)
iE(ξi)/∑E(ξ)
SFS classes

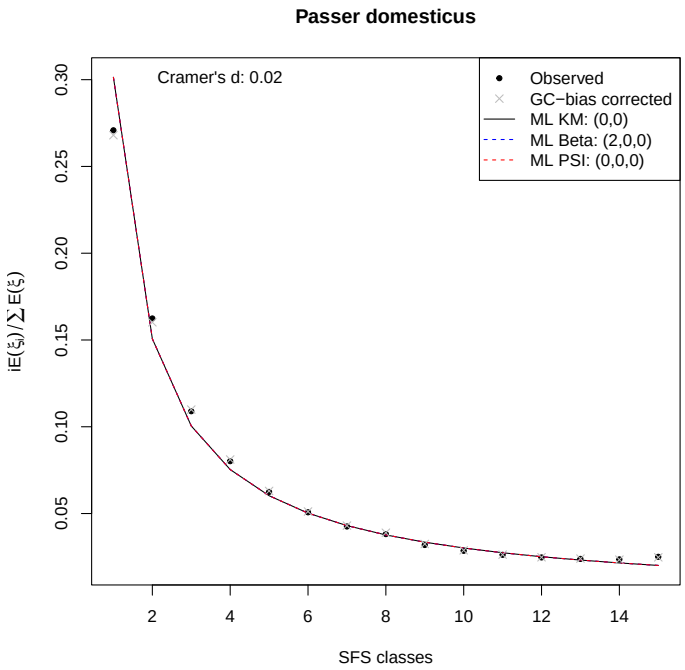

Gorilla gorilla
Cramer's d: 0.04
Observed
GC−bias corrected
ML KM: (0,0)
ML Beta: (1.9,0,0)
ML PSI: (0,0,0)
iE(ξi)/∑E(ξ)
SFS classes

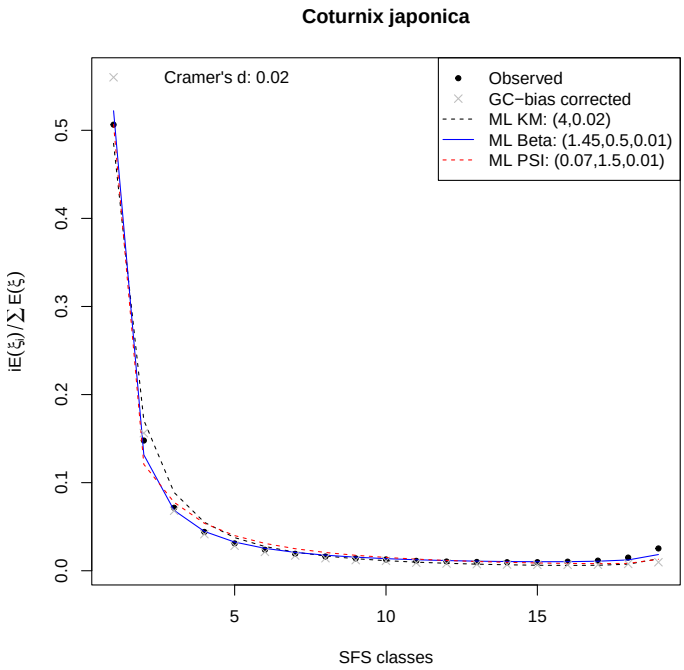

Parus maior
Cramer's d: 0.01
Observed
GC−bias corrected
ML KM: (0.5,0.01)
ML Beta: (1.75,0,0.01)
ML PSI: (0.03,0,0.01)
iE(ξi)/∑E(ξ)
SFS classes


Halictus scabiosae
Cramer's d: 0.06
Observed
GC−bias corrected
ML KM: (0,0.01)
ML Beta: (1.85,0,0.01)
ML PSI: (0.04,0,0.01)
iE(ξi)/∑E(ξ)
SFS classes

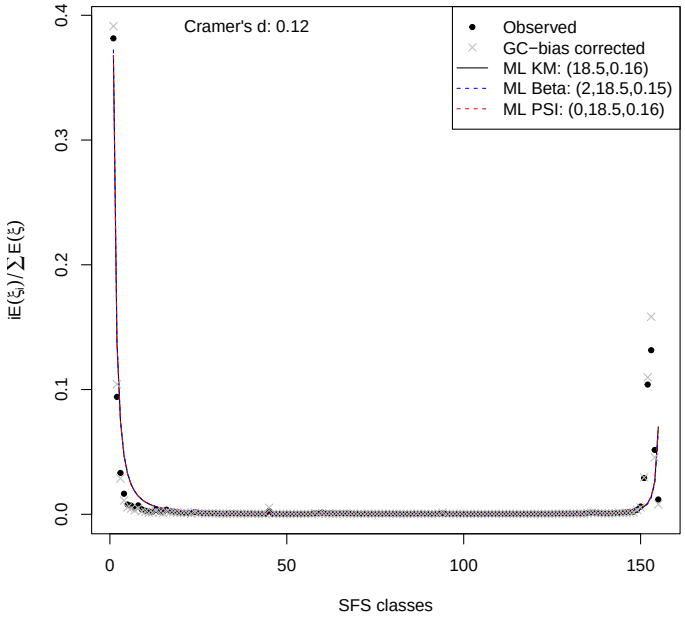

Helicobacter pilori
Cramer's d: 0.05
Observed
GC−bias corrected
ML KM: (1,0.2)
ML Beta: (1.65,0,0.15)
ML PSI: (0.01,1,0.2)
iE(ξi)/∑E(ξ)
SFS classes


Corvus cornix
Cramer's d: 0.02
Observed
GC−bias corrected
ML KM: (1.5,0)
ML Beta: (1.95,1,0)
ML PSI: (0,1.5,0)
iE(ξi)/∑E(ξ)
SFS classes

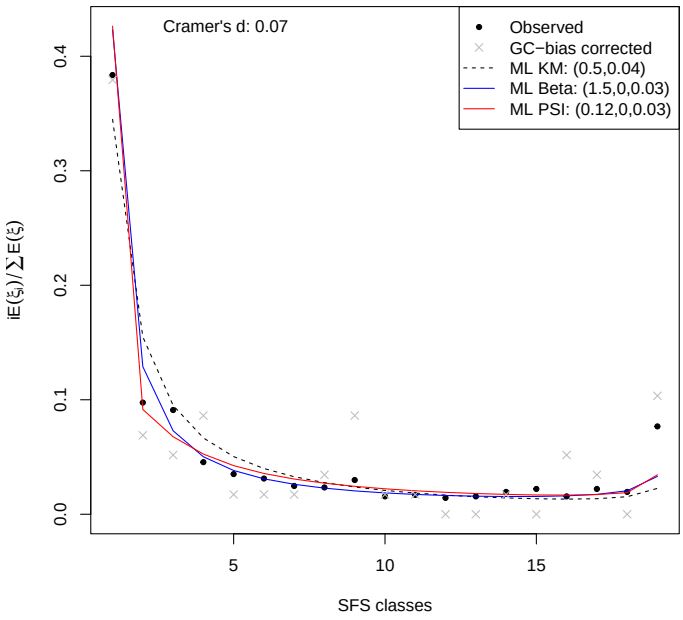

Passer domesticus
Cramer's d: 0.02
Observed
GC−bias corrected
ML KM: (0,0)
ML Beta: (2,0,0)
ML PSI: (0,0,0)
iE(ξi)/∑E(ξ)
SFS classes


Coturnix japonica
Cramer's d: 0.02
Observed
GC−bias corrected
ML KM: (4,0.02)
ML Beta: (1.45,0.5,0.01)
ML PSI: (0.07,1.5,0.01)
iE(ξi)/∑E(ξ)
SFS classes

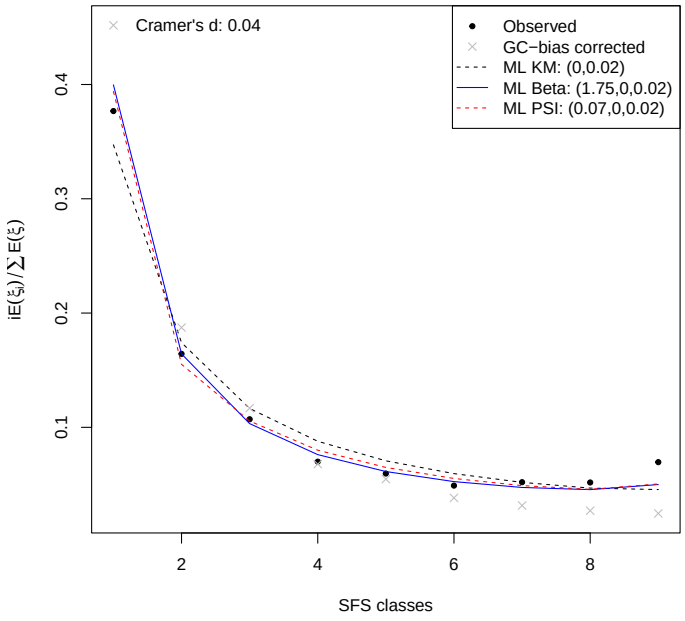
Klebsiella pneumoniae



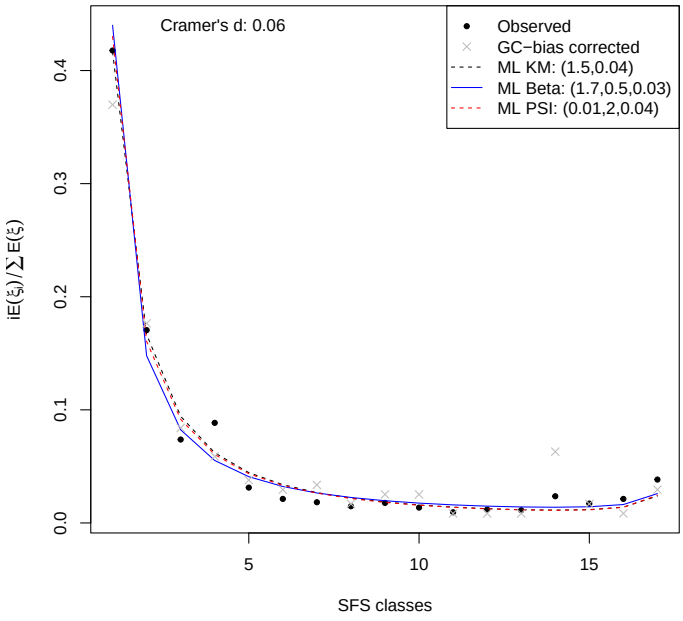
Lepus granatensis

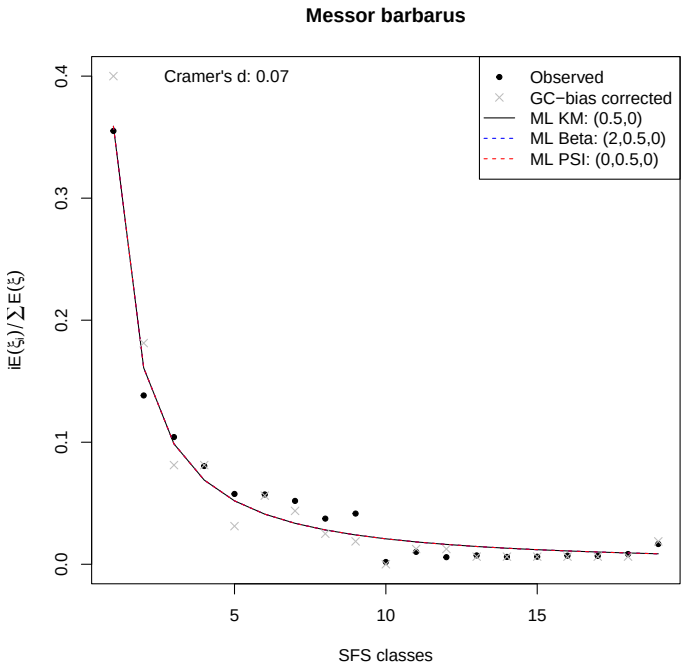


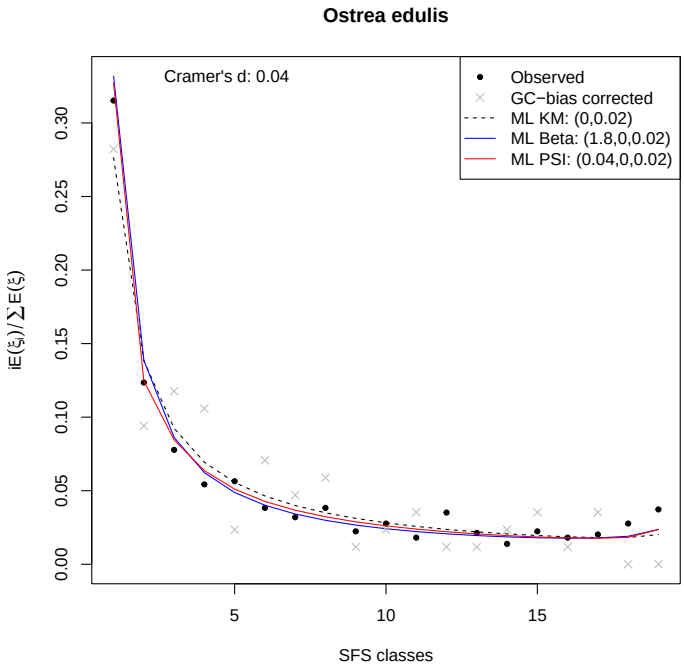
Egretta garzetta

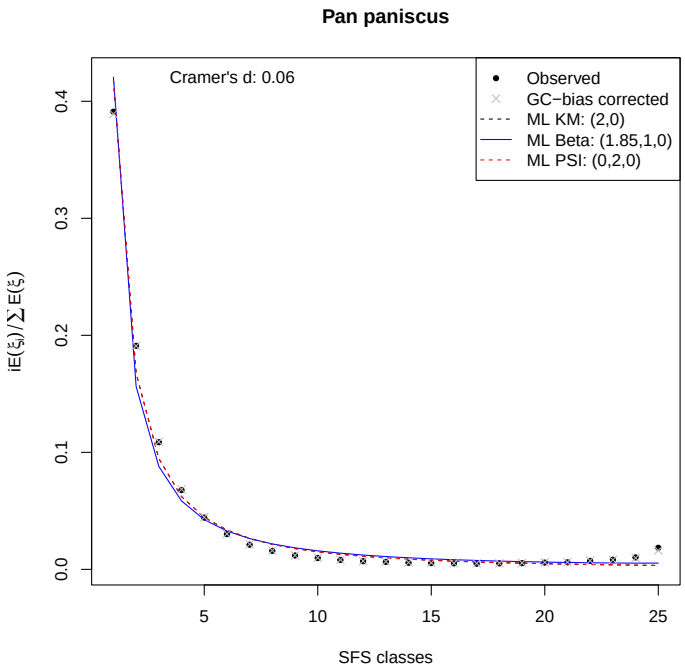


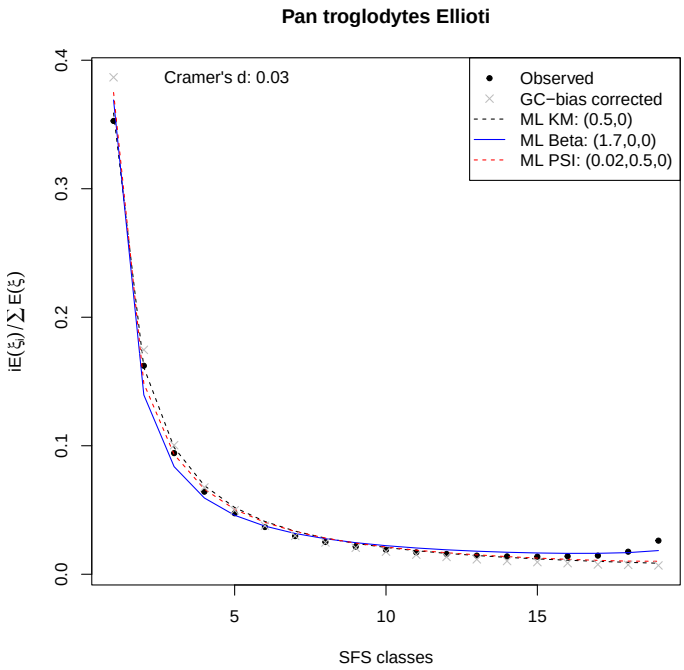
Melitaea cinxia



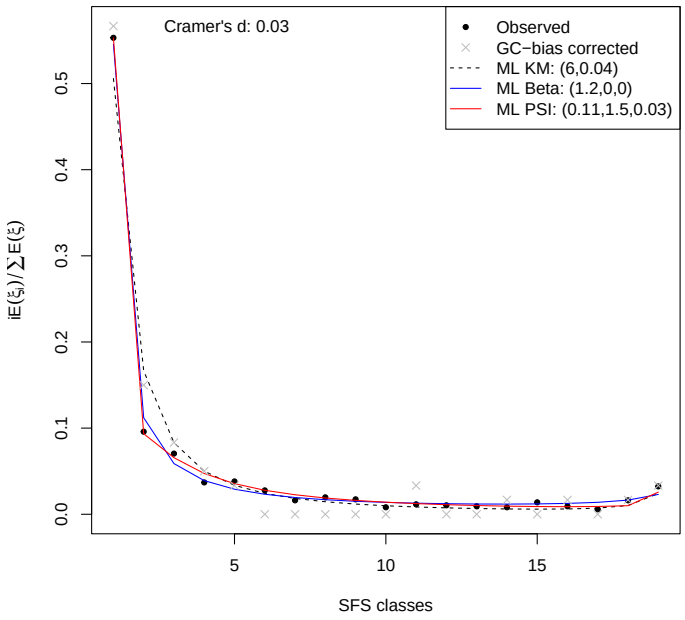

Messor barbarus
Cramer's d: 0.07
Observed
GC−bias corrected
ML KM: (0.5,0)
ML Beta: (2,0.5,0)
ML PSI: (0,0.5,0)
iE(ξi)/∑E(ξ)
SFS classes


Ostrea edulis
Cramer's d: 0.04
Observed
GC−bias corrected
ML KM: (0,0.02)
ML Beta: (1.8,0,0.02)
ML PSI: (0.04,0,0.02)
iE(ξi)/∑E(ξ)
SFS classes

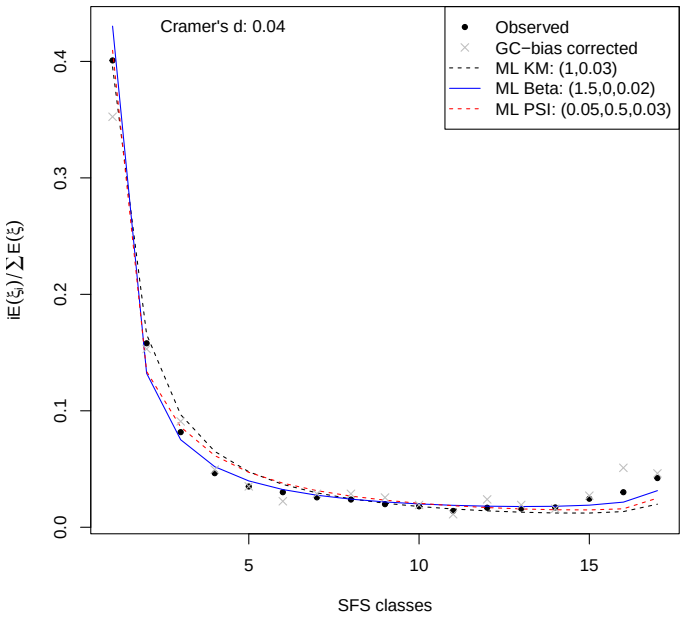

Pan paniscus
Cramer's d: 0.06
Observed
GC−bias corrected
ML KM: (2,0)
ML Beta: (1.85,1,0)
ML PSI: (0,2,0)
iE(ξi)/∑E(ξ)
SFS classes


Pan troglodytes Ellioti
Cramer's d: 0.03
Observed
GC−bias corrected
ML KM: (0.5,0)
ML Beta: (1.7,0,0)
ML PSI: (0.02,0.5,0)
iE(ξi)/∑E(ξ)
SFS classes

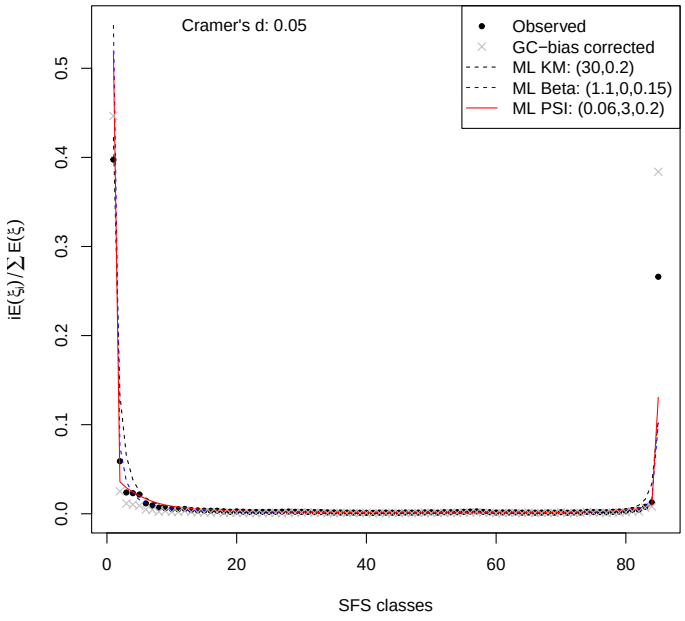
Parus caeruleus



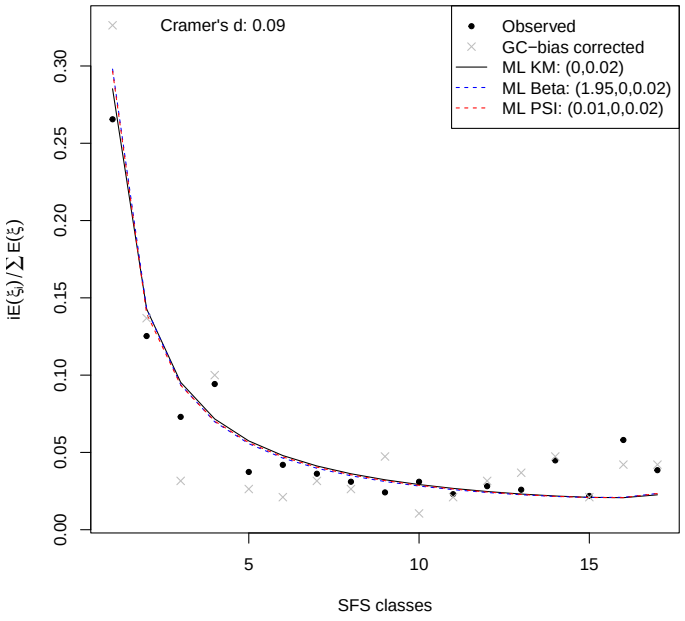
Physa acuta



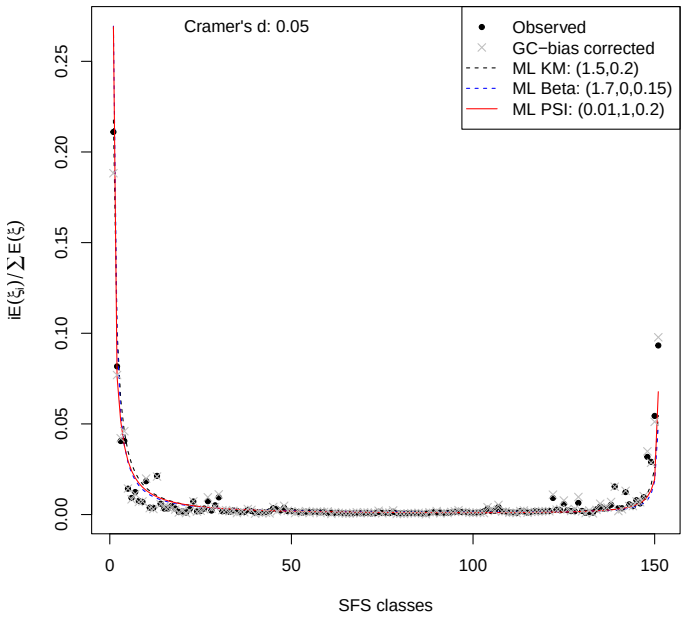
Pseudomonas aeruginosa



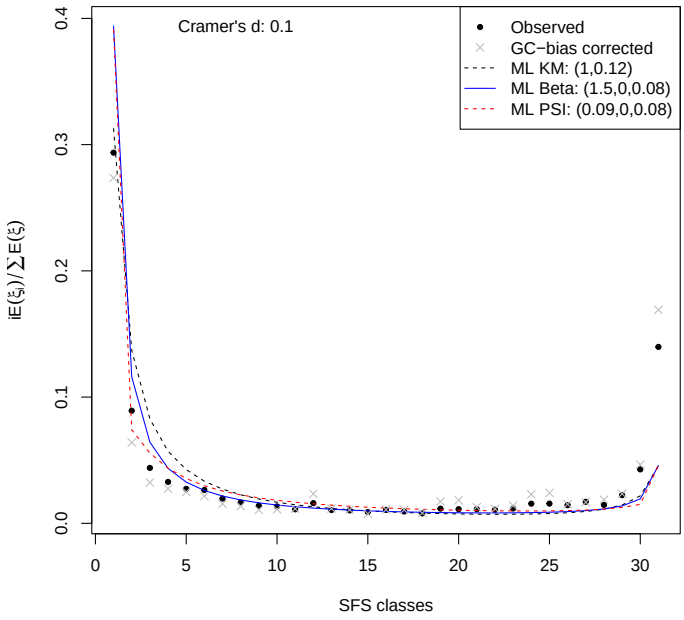
Sepia officinalis



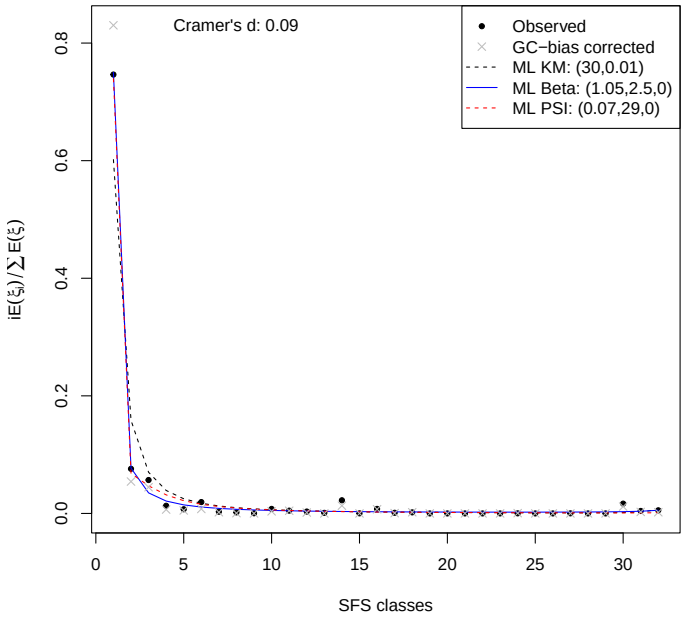
Staphylococcus aureus



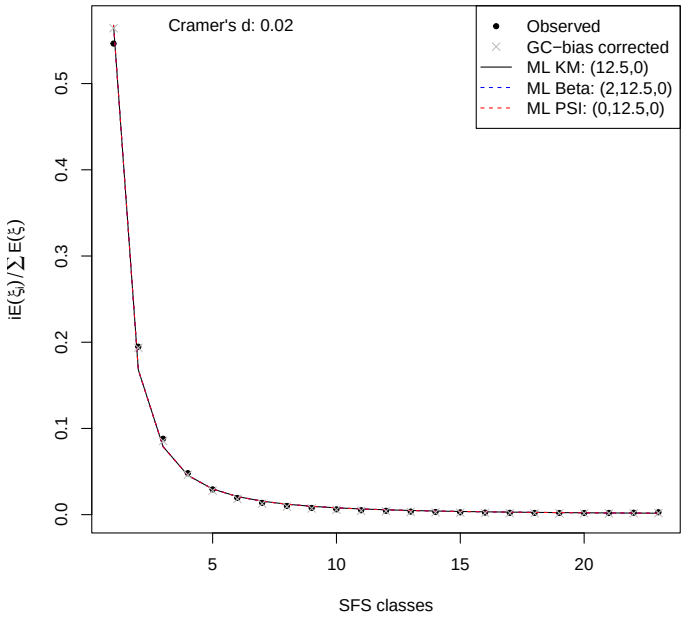
Streptococcus pneumoniae



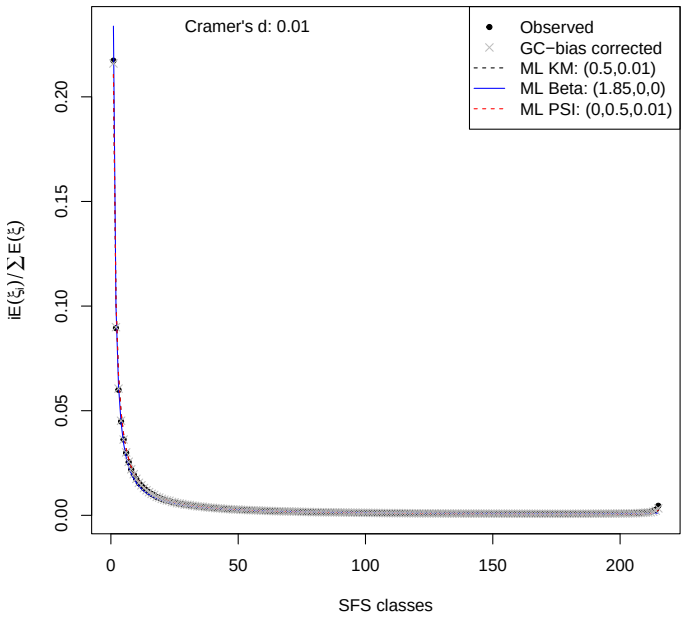
Mycobacterium tubercolosis



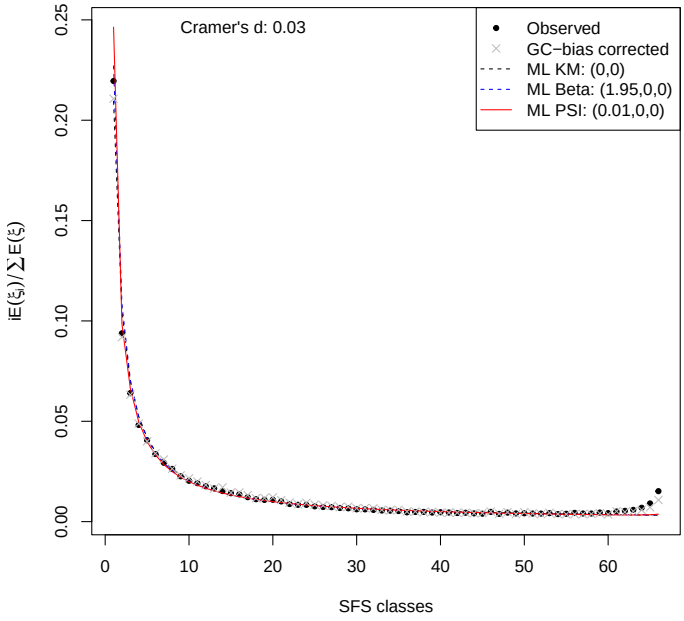
Phylloscopus trochilus



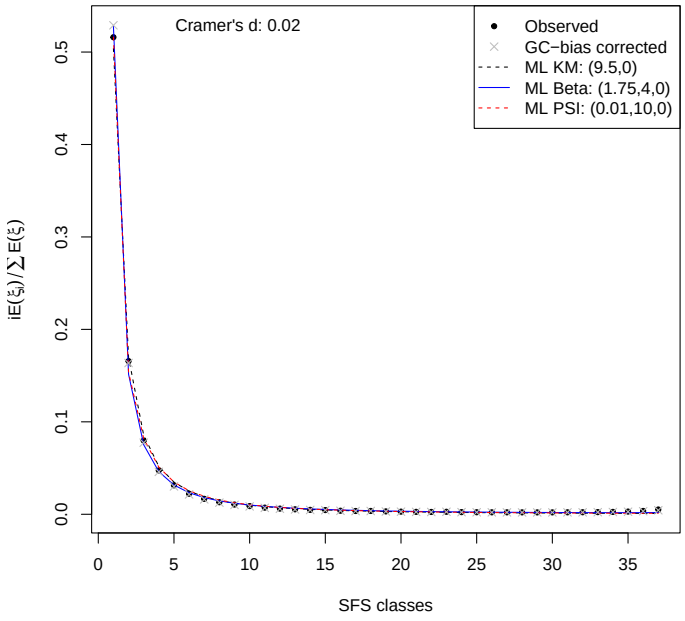
Homo sapiens



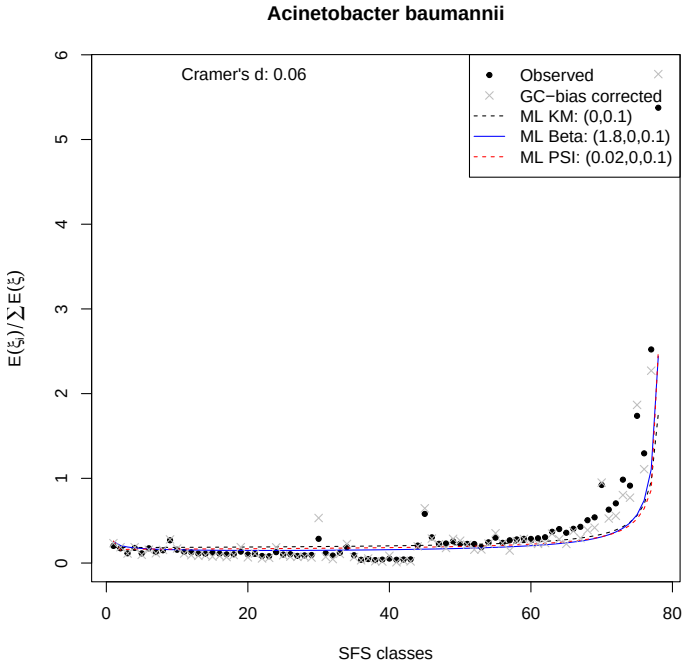
Zea Mays

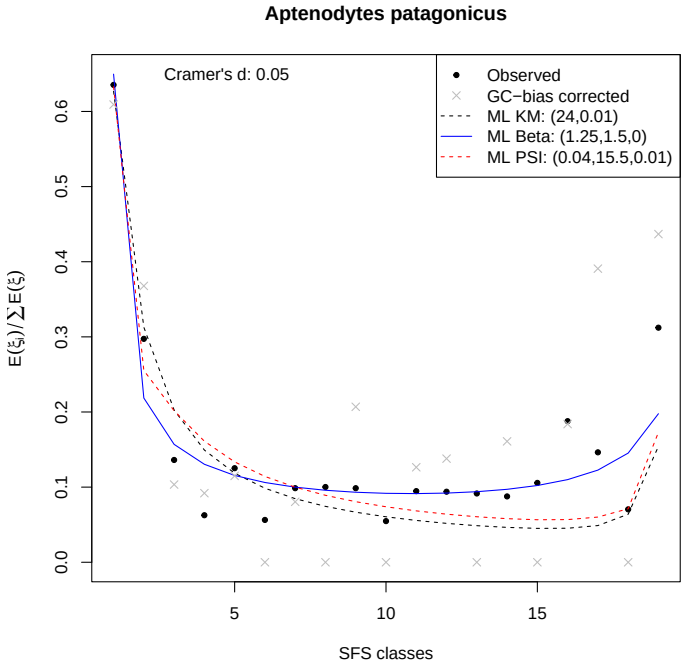


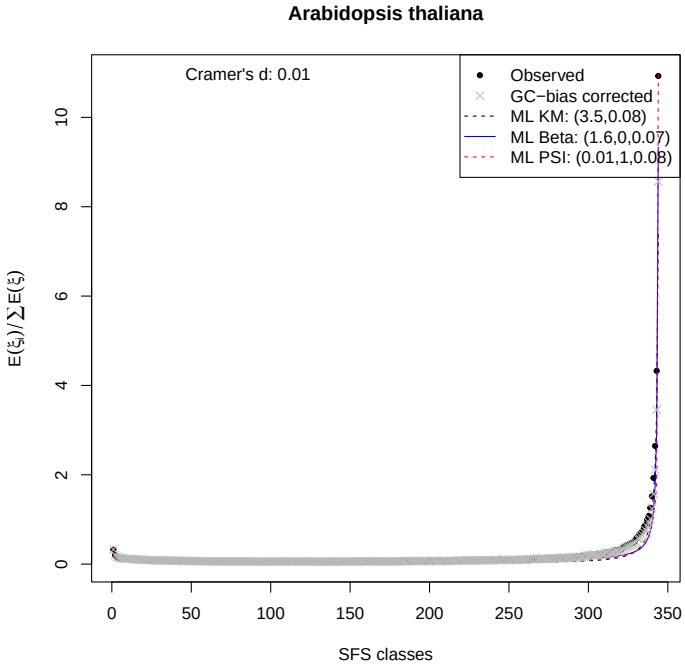
Taeniopygia guttata

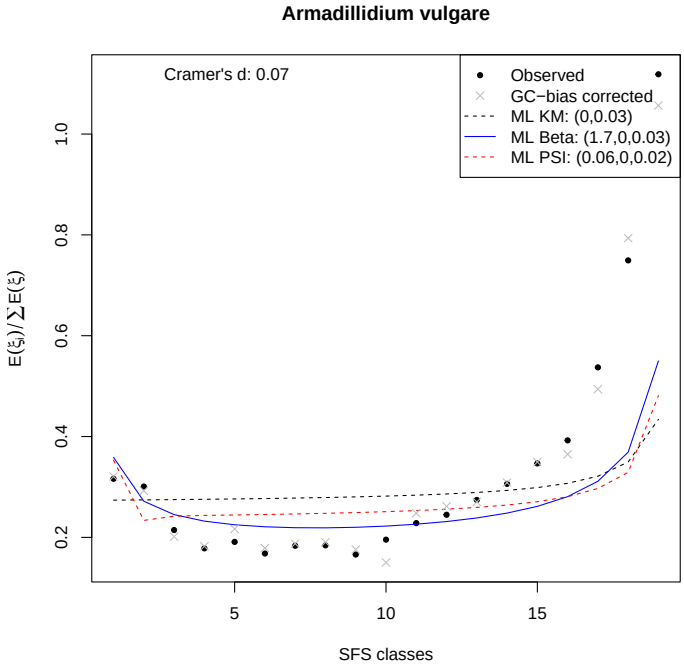


6.2.2 Supplementary file 2

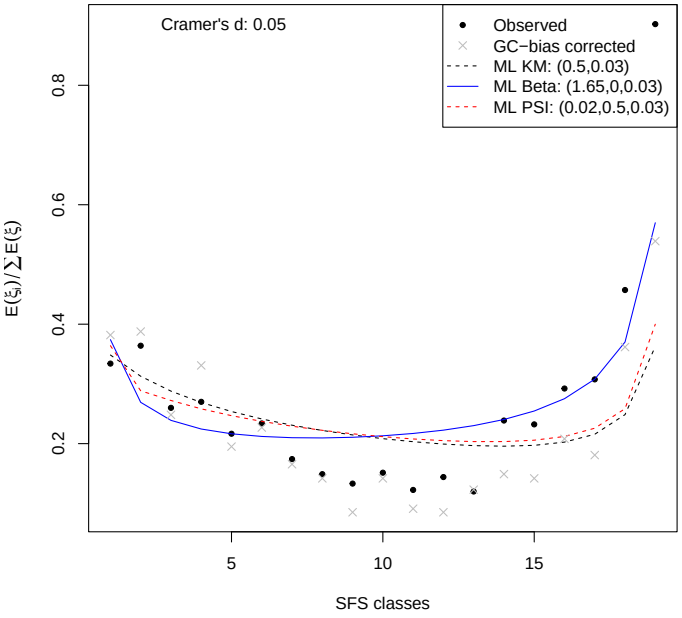




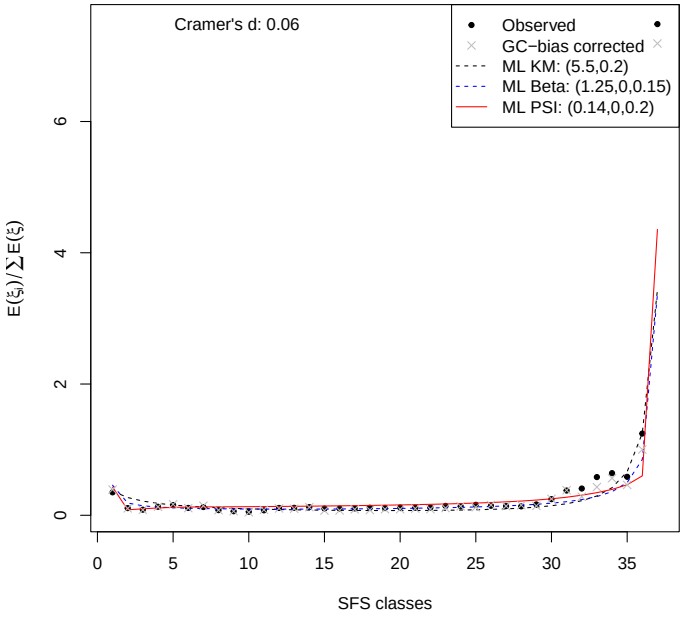




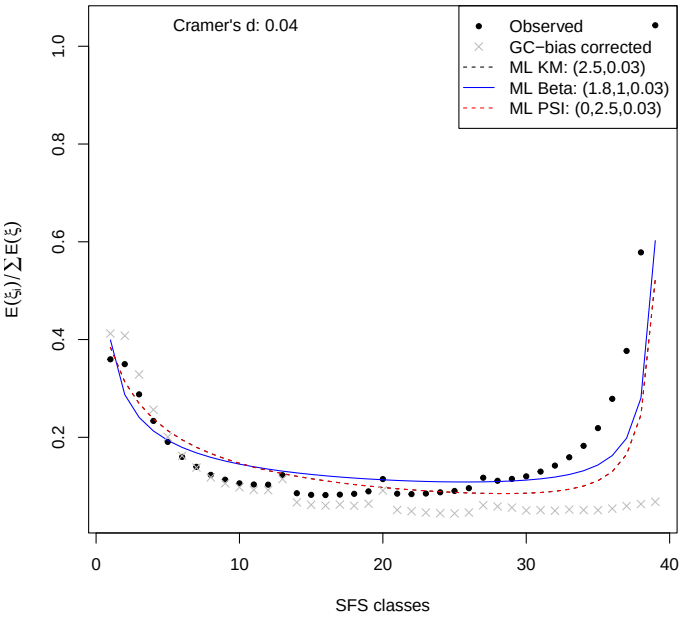
Artemia franciscana



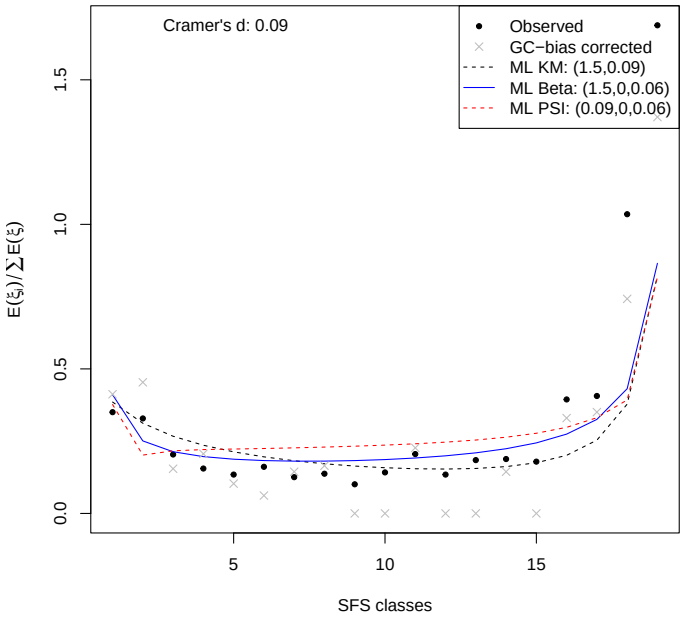
Bacillus subtilis



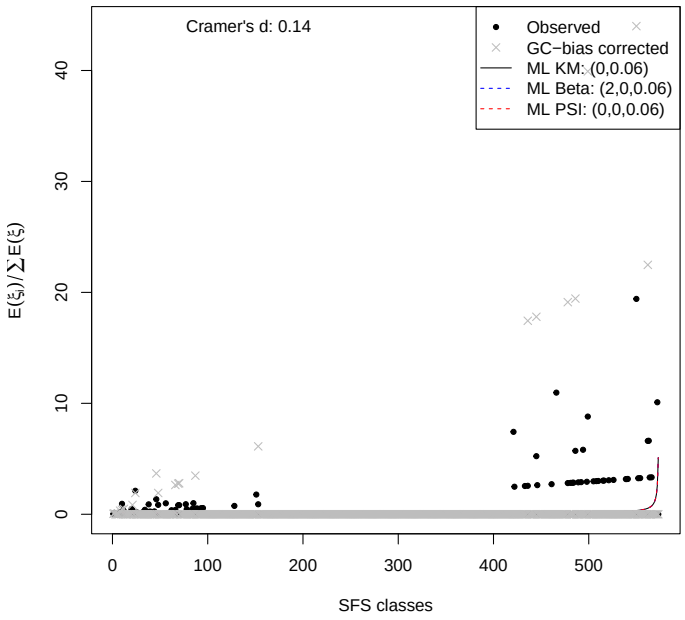
Athene cunicularia



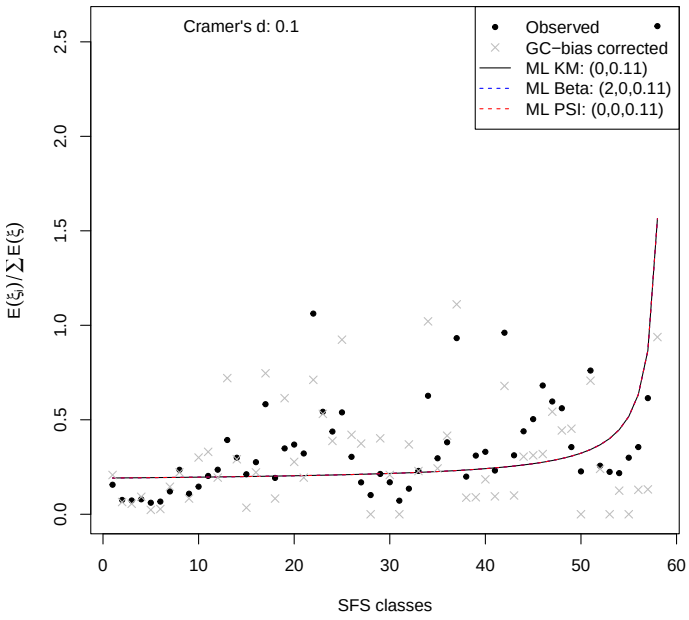
Caenorhabditis brenneri



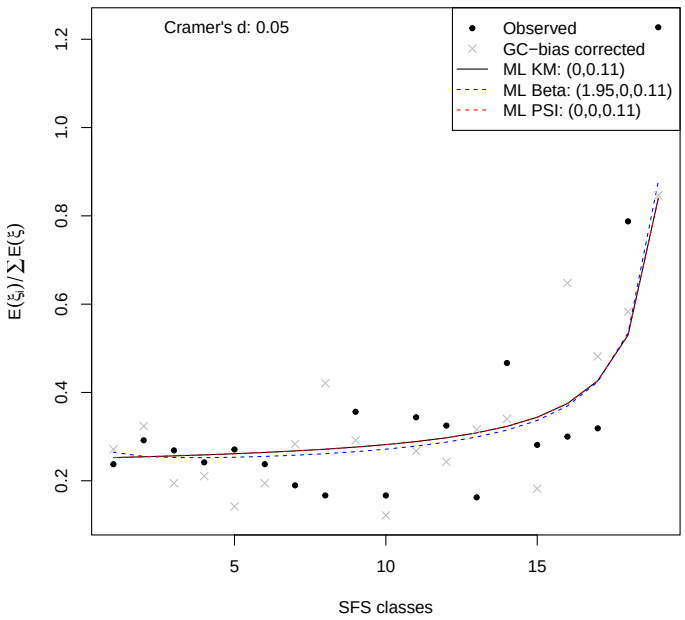
Caenorhabditis elegans



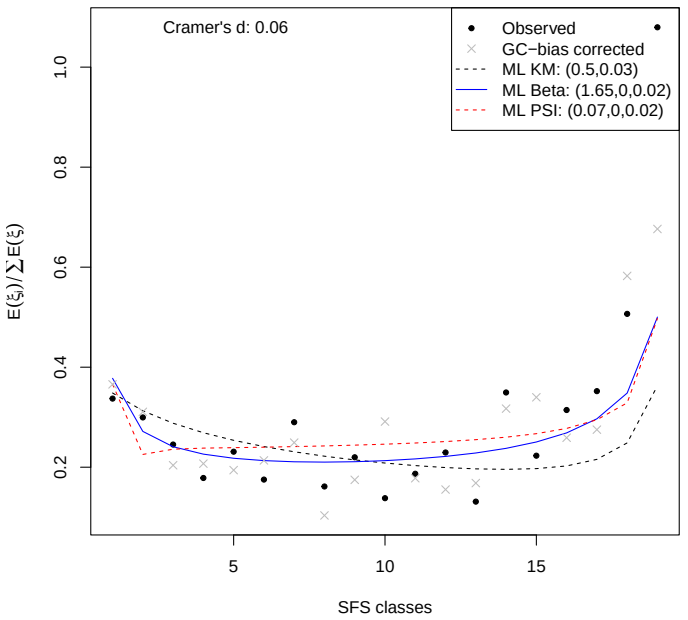
Chlamydia trachomatis



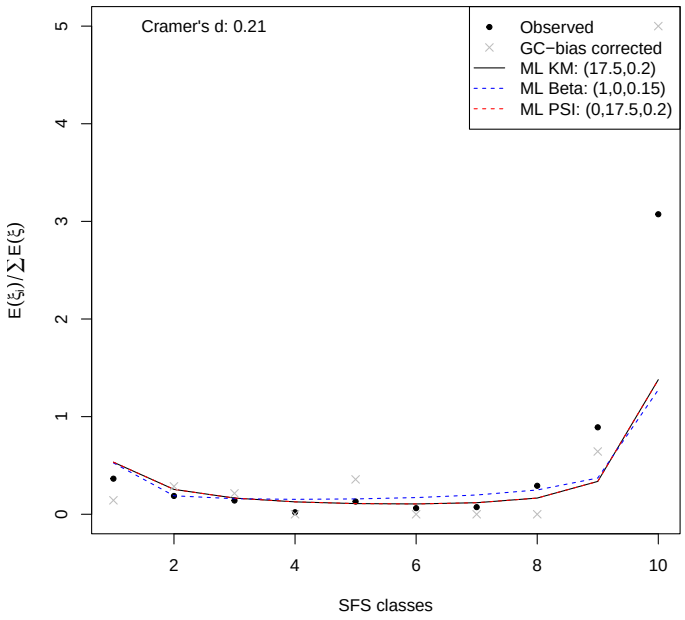
Ciona intestinalis A



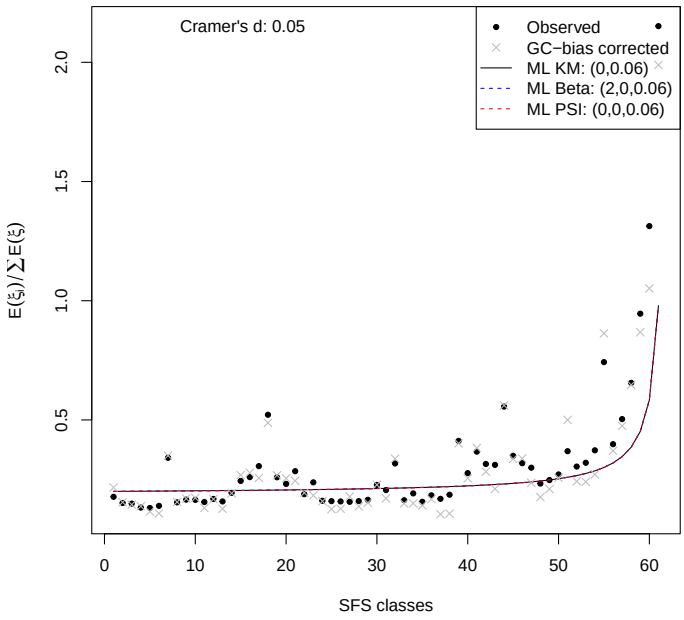
Ciona intestinalis B



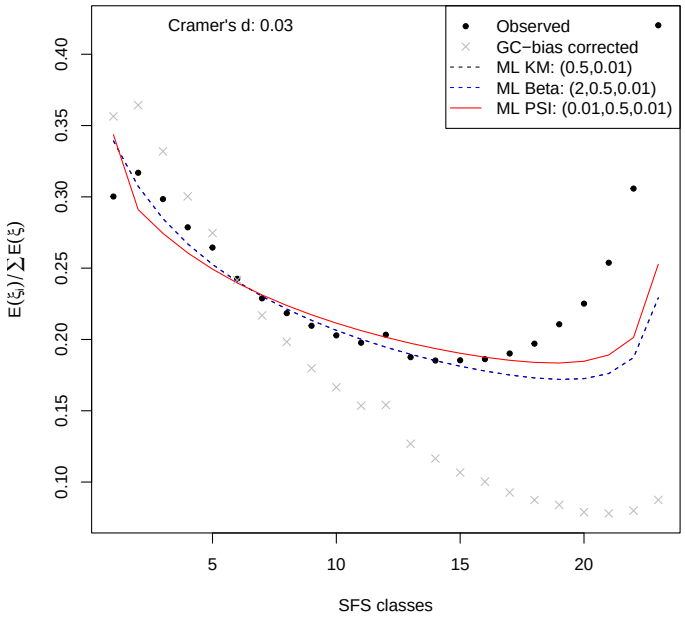
Clostridium difficile



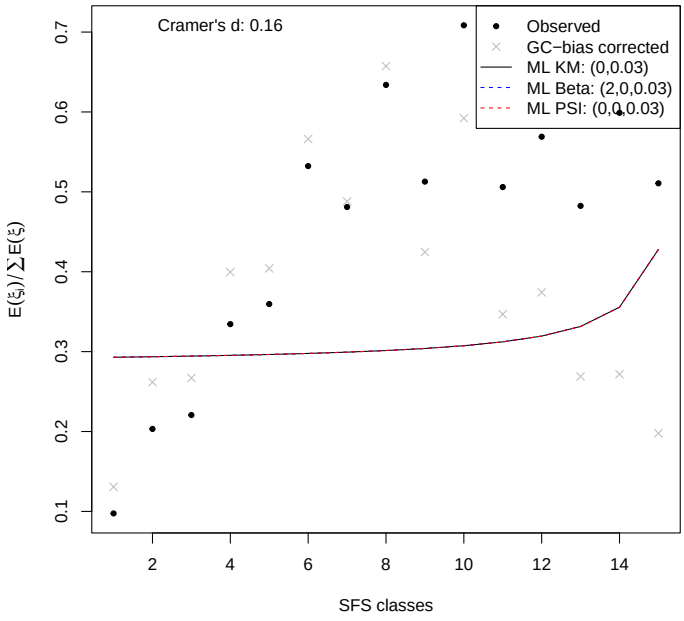
Escherichia coli



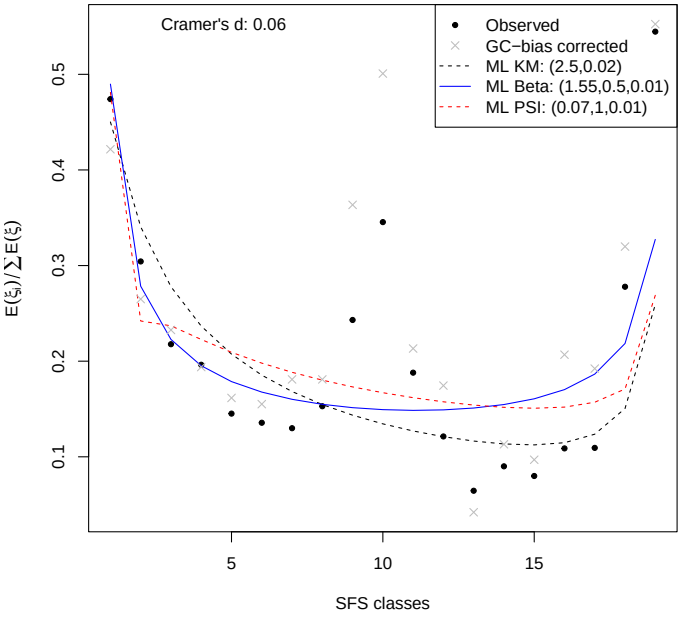
Ficedula albicollis



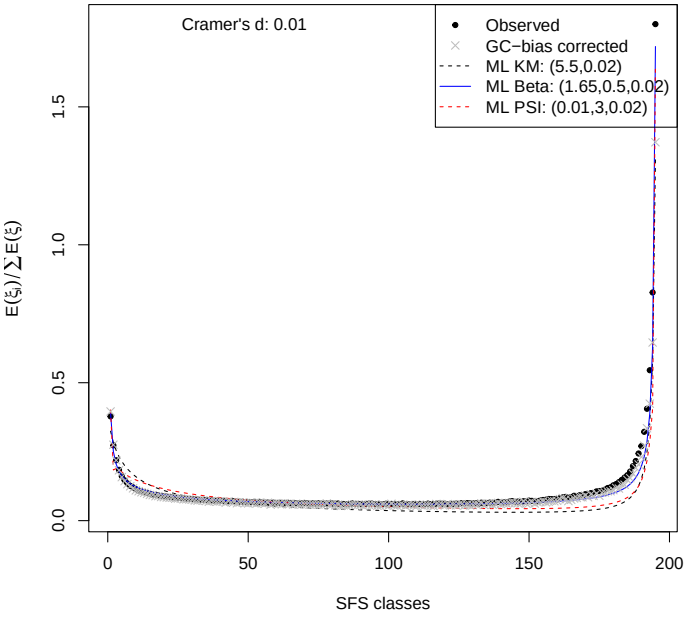
Nipponia nippon



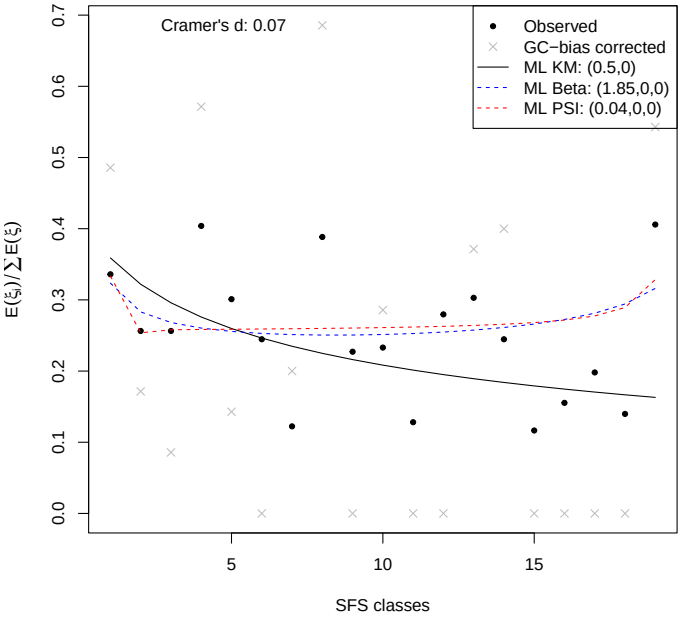
Culex pipiens



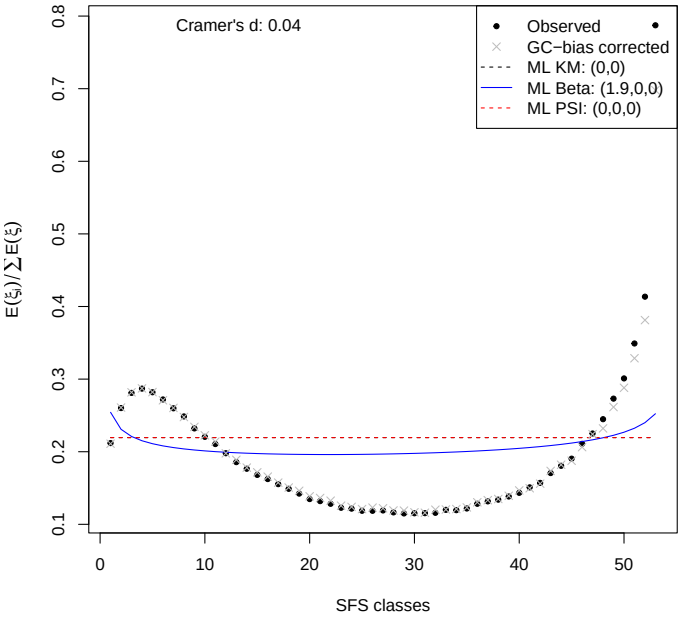
Drosophila melanogaster

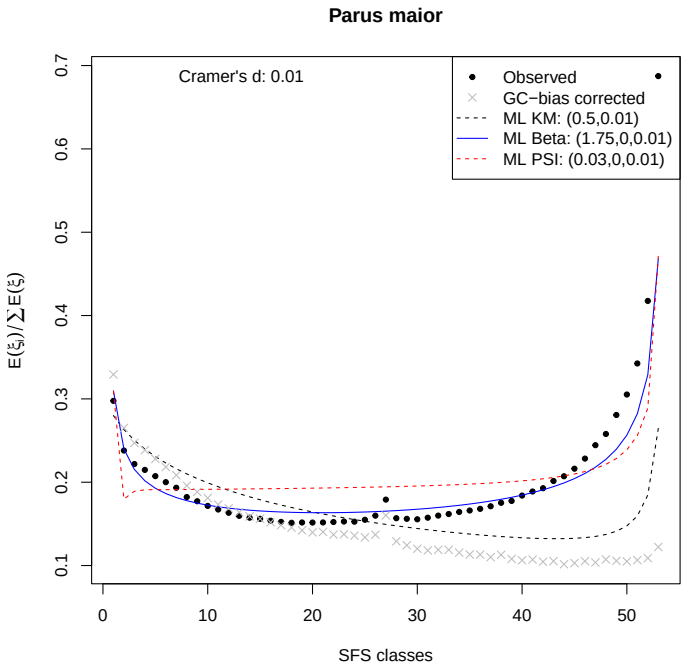


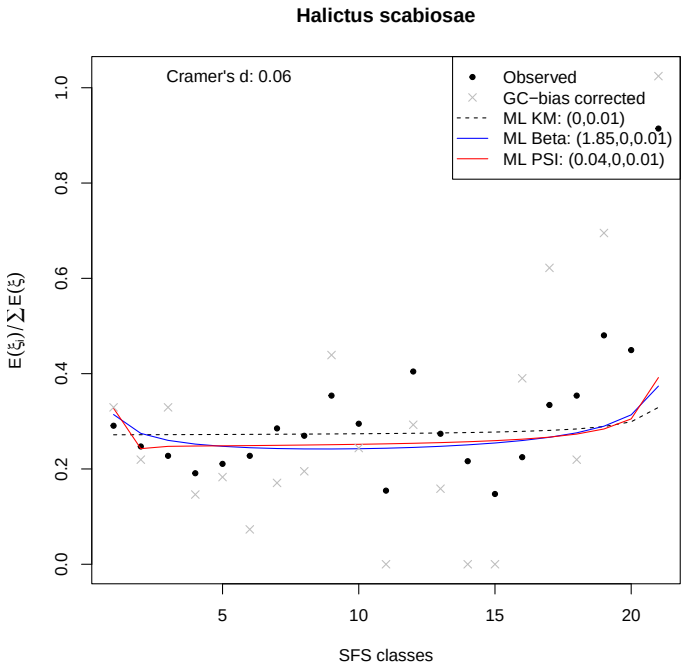
Emys orbicularis

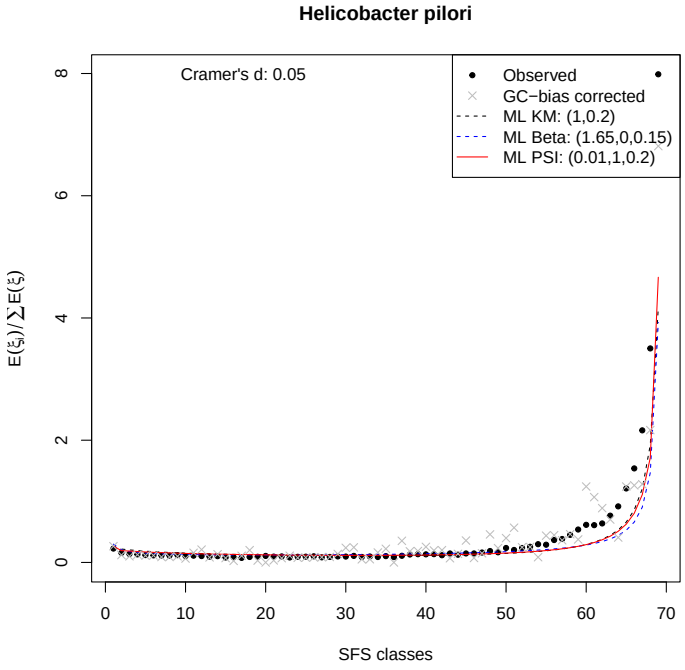


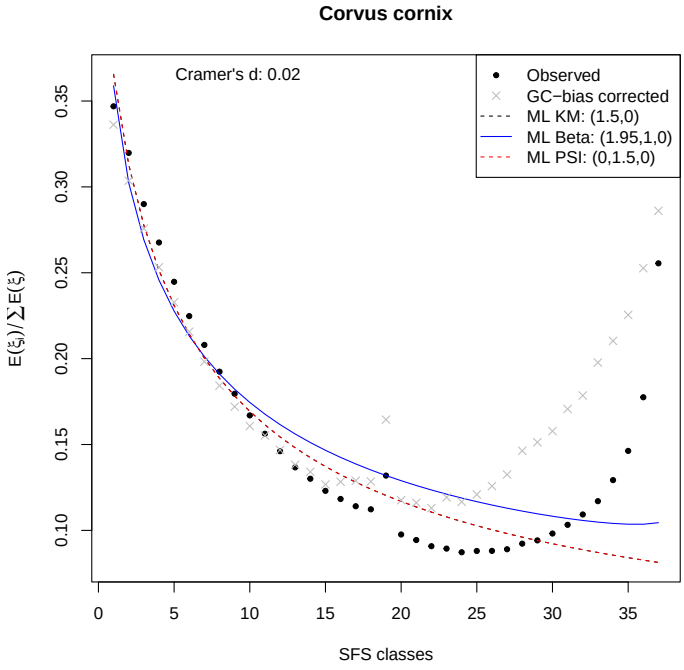
Gorilla gorilla

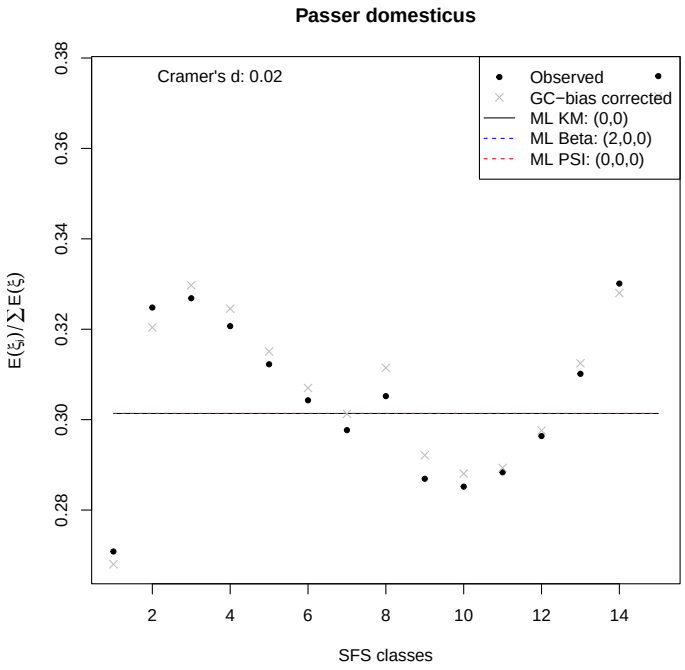


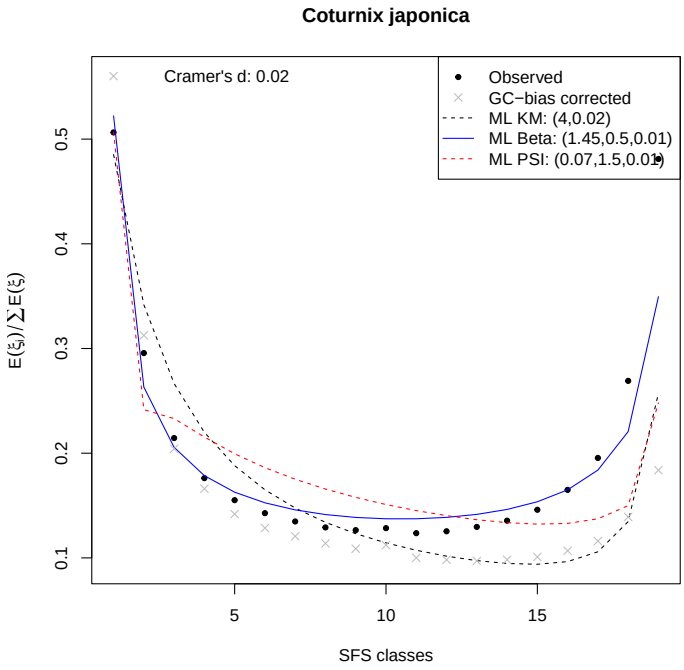

Parus maior
Cramer's d: 0.01
Observed
GC−bias corrected
ML KM: (0.5,0.01)
ML Beta: (1.75,0,0.01)
ML PSI: (0.03,0,0.01)
E(ξi)/∑E(ξ)
SFS classes

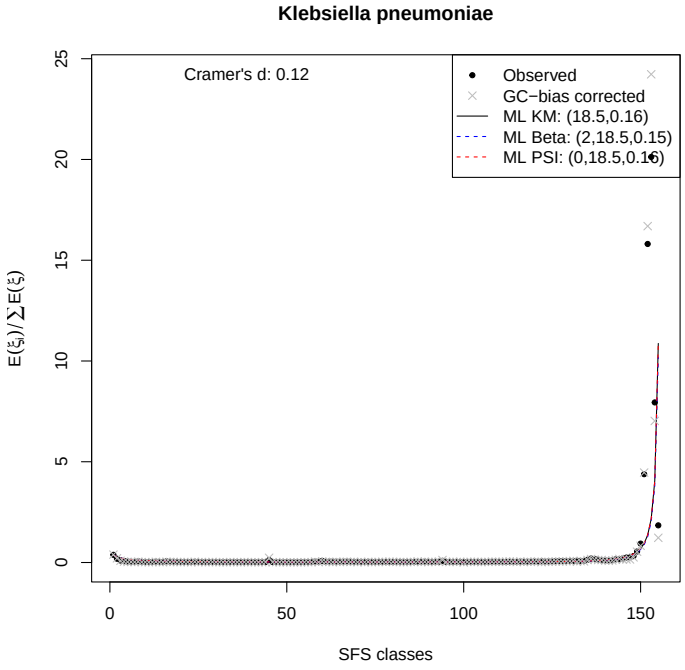

Halictus scabiosae
Cramer's d: 0.06
Observed
GC−bias corrected
ML KM: (0,0.01)
ML Beta: (1.85,0,0.01)
ML PSI: (0.04,0,0.01)
E(ξi)/∑E(ξ)
SFS classes

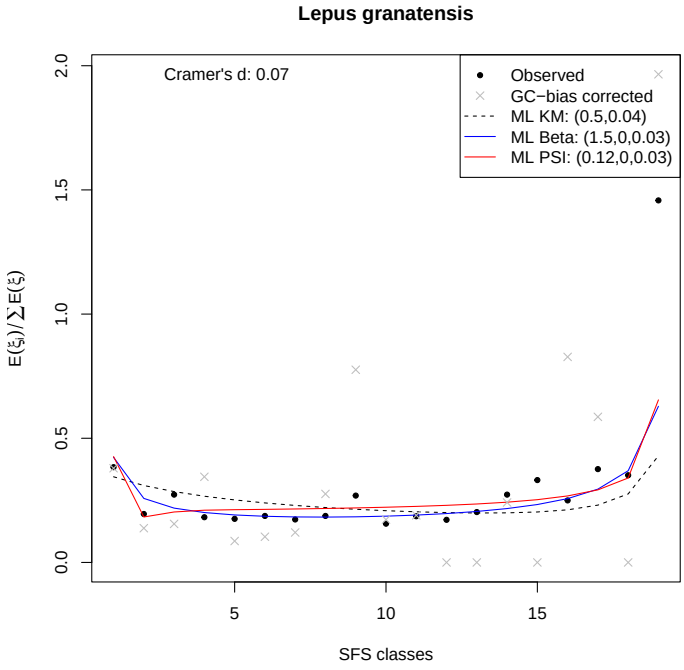

Helicobacter pilori
Cramer's d: 0.05
Observed
GC−bias corrected
ML KM: (1,0.2)
ML Beta: (1.65,0,0.15)
ML PSI: (0.01,1,0.2)
E(ξi)/∑E(ξ)
SFS classes

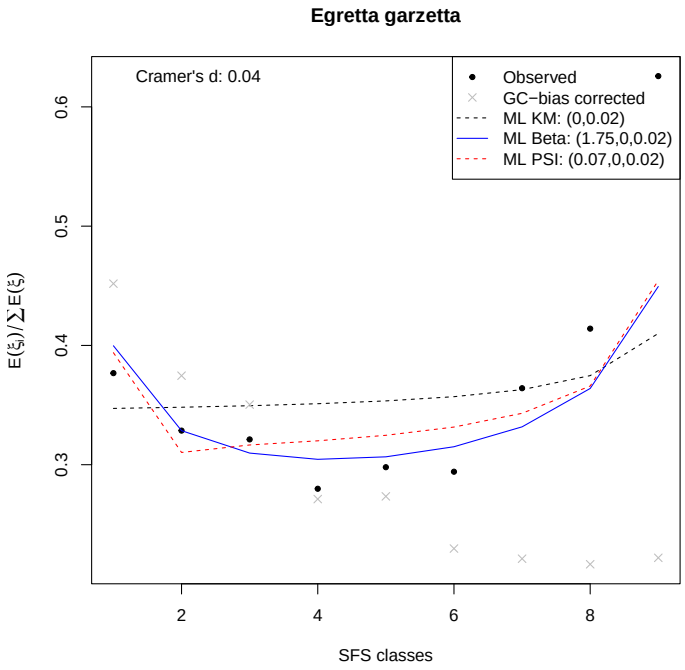

Corvus cornix
Cramer's d: 0.02
Observed
GC−bias corrected
ML KM: (1.5,0)
ML Beta: (1.95,1,0)
ML PSI: (0,1.5,0)
E(ξi)/∑E(ξ)
SFS classes

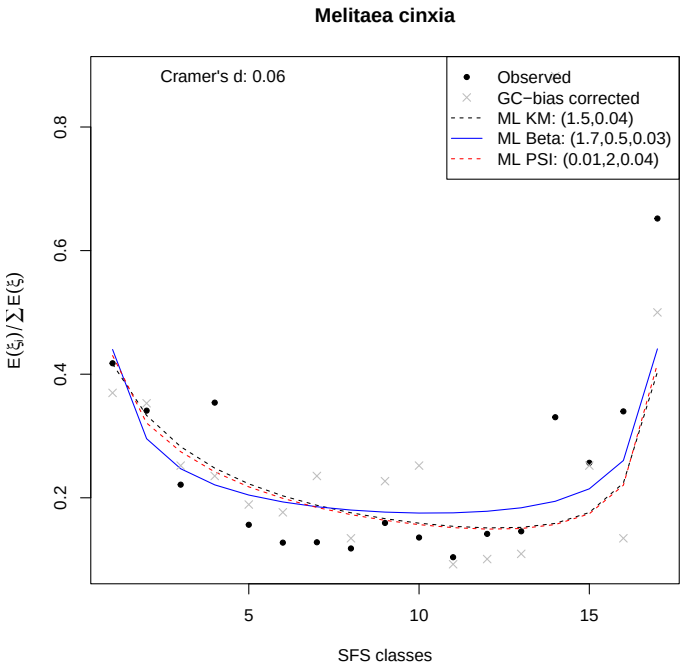

Passer domesticus
Cramer's d: 0.02
Observed
GC−bias corrected
ML KM: (0,0)
ML Beta: (2,0,0)
ML PSI: (0,0,0)
E(ξi)/∑E(ξ)
SFS classes

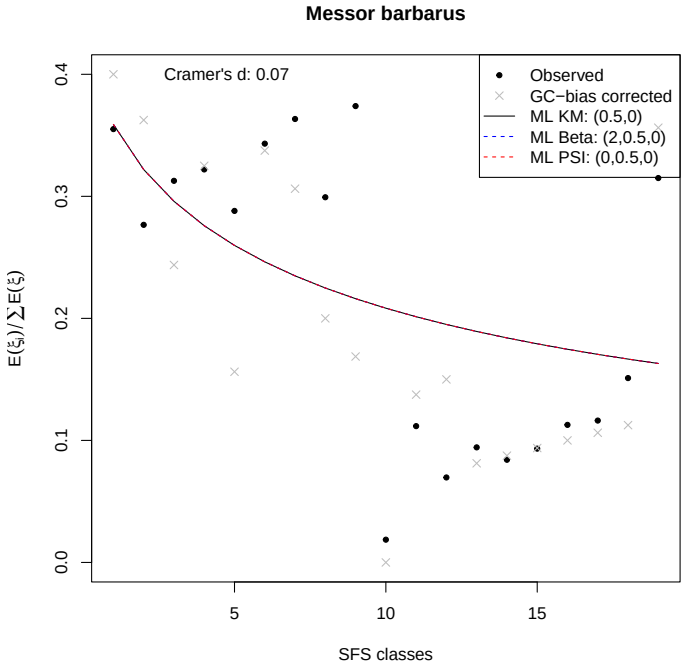

Coturnix japonica
Cramer's d: 0.02
Observed
GC−bias corrected
ML KM: (4,0.02)
ML Beta: (1.45,0.5,0.01)
ML PSI: (0.07,1.5,0.01)
E(ξi)/∑E(ξ)
SFS classes

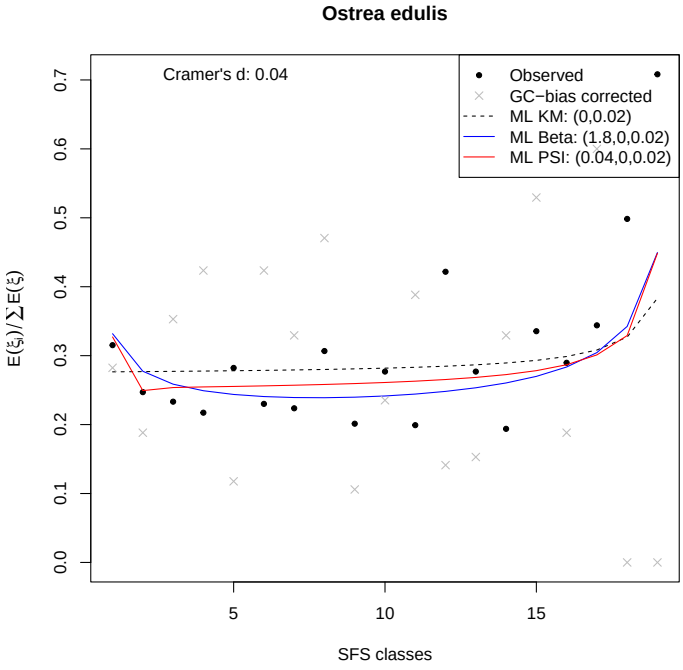

Klebsiella pneumoniae
Cramer's d: 0.12
Observed
GC−bias corrected
ML KM: (18.5,0.16)
ML Beta: (2,18.5,0.15)
ML PSI: (0,18.5,0.16)
E(ξi)/∑E(ξ)
SFS classes


Lepus granatensis
Cramer's d: 0.07
Observed
GC−bias corrected
ML KM: (0.5,0.04)
ML Beta: (1.5,0,0.03)
ML PSI: (0.12,0,0.03)
E(ξi)/∑E(ξ)
SFS classes

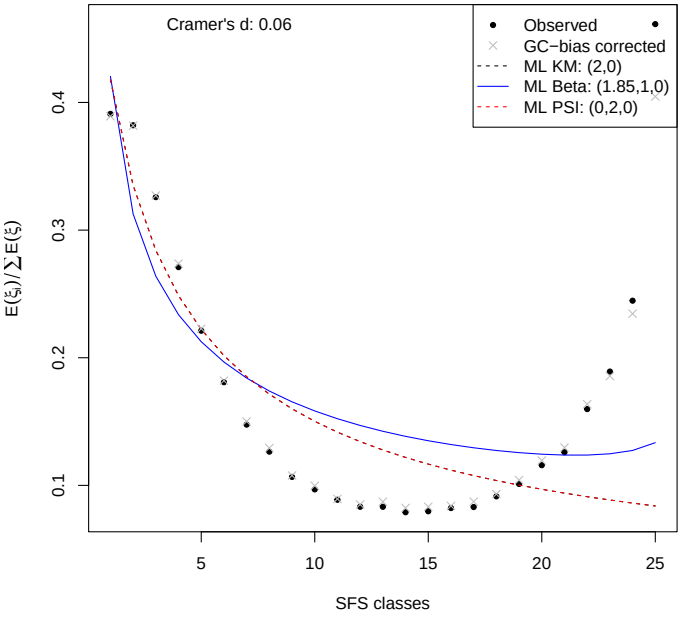

Egretta garzetta
Cramer's d: 0.04
Observed
GC−bias corrected
ML KM: (0,0.02)
ML Beta: (1.75,0,0.02)
ML PSI: (0.07,0,0.02)
E(ξi)/∑E(ξ)
SFS classes


Melitaea cinxia
Cramer's d: 0.06
Observed
GC−bias corrected
ML KM: (1.5,0.04)
ML Beta: (1.7,0.5,0.03)
ML PSI: (0.01,2,0.04)
E(ξi)/∑E(ξ)
SFS classes

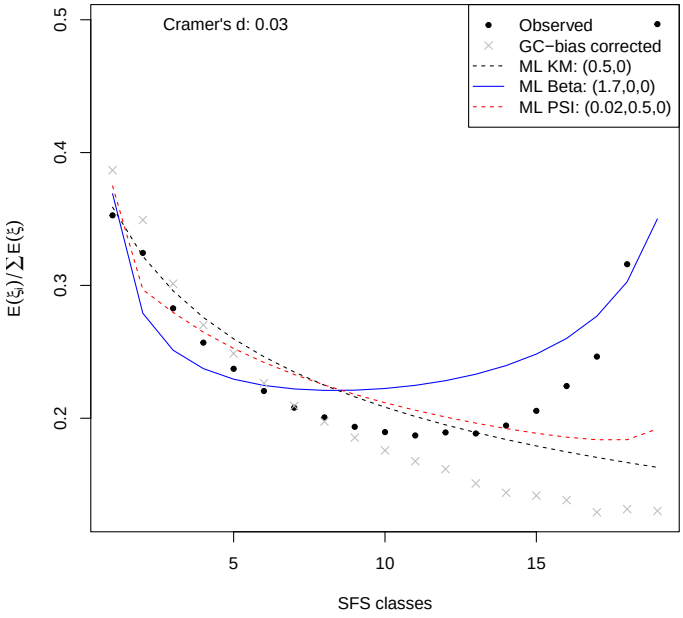

Messor barbarus
Cramer's d: 0.07
Observed
GC−bias corrected
ML KM: (0.5,0)
ML Beta: (2,0.5,0)
ML PSI: (0,0.5,0)
E(ξi)/∑E(ξ)
SFS classes


Ostrea edulis
Cramer's d: 0.04
Observed
GC−bias corrected
ML KM: (0,0.02)
ML Beta: (1.8,0,0.02)
ML PSI: (0.04,0,0.02)
E(ξi)/∑E(ξ)
SFS classes

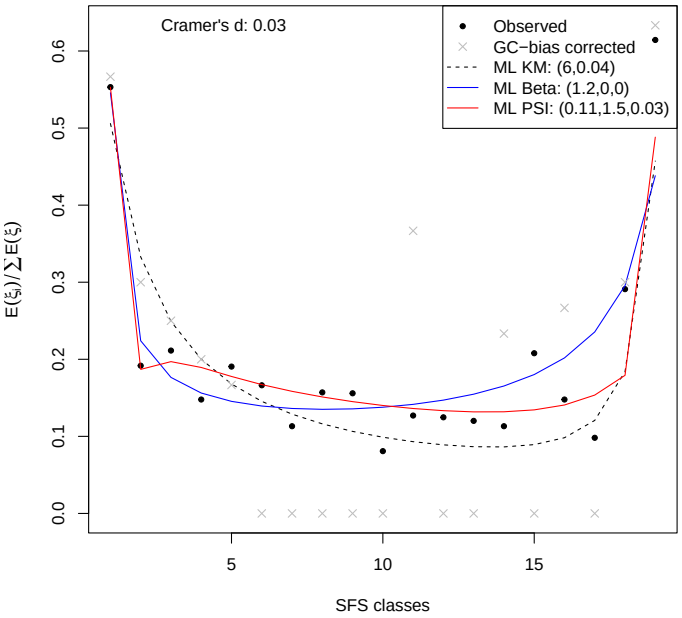
Pan paniscus



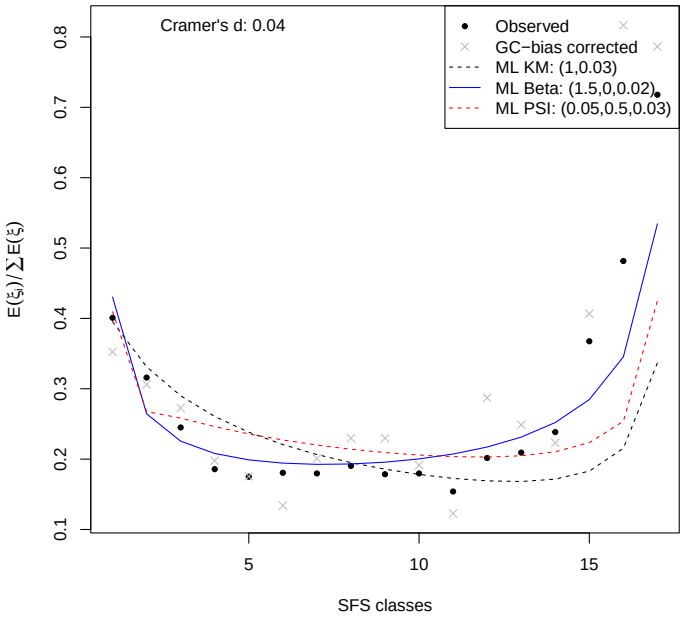
Pan troglodytes Ellioti



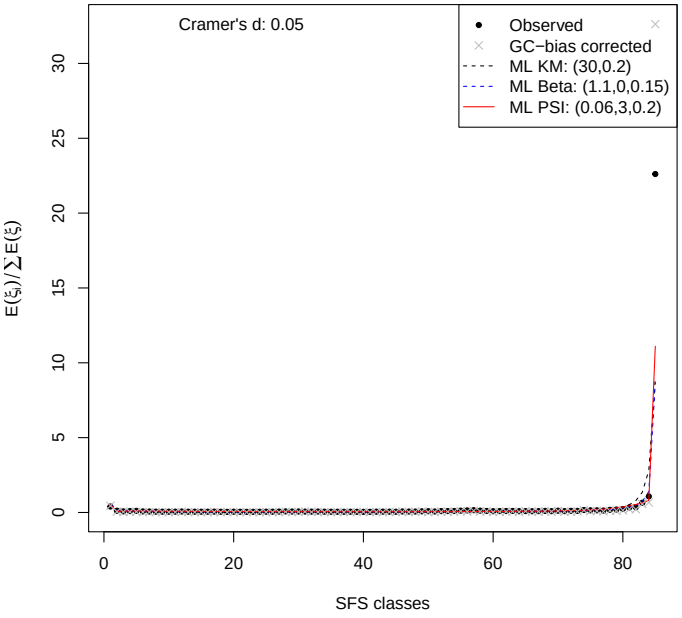
Parus caeruleus



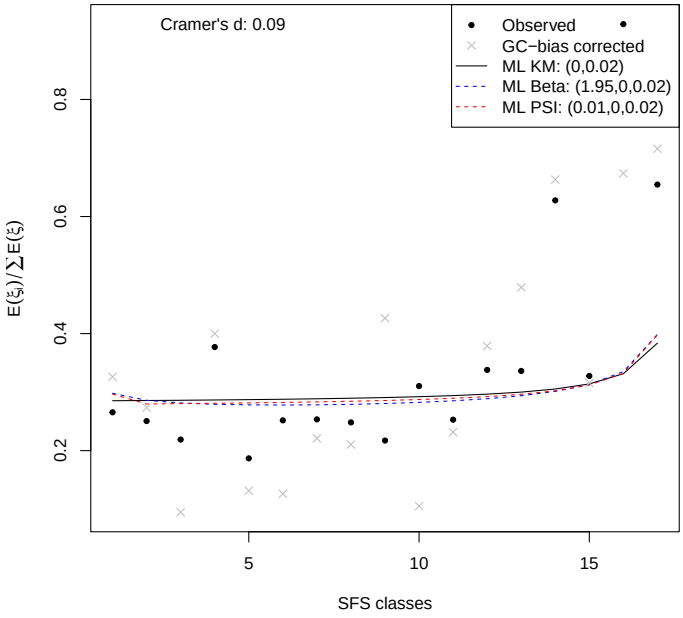
Physa acuta



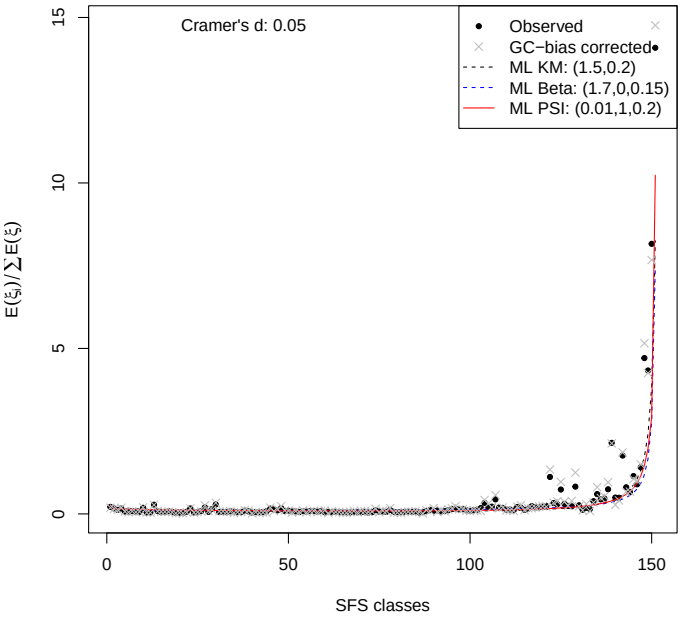
Pseudomonas aeruginosa



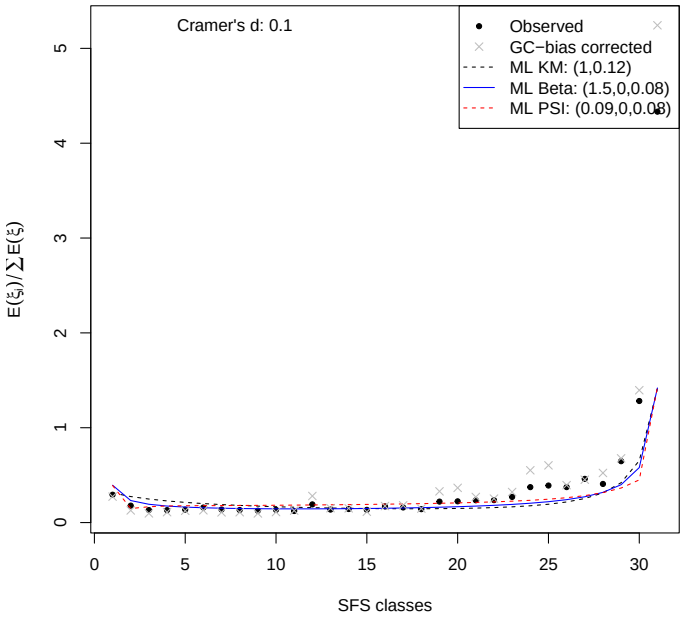
Sepia officinalis



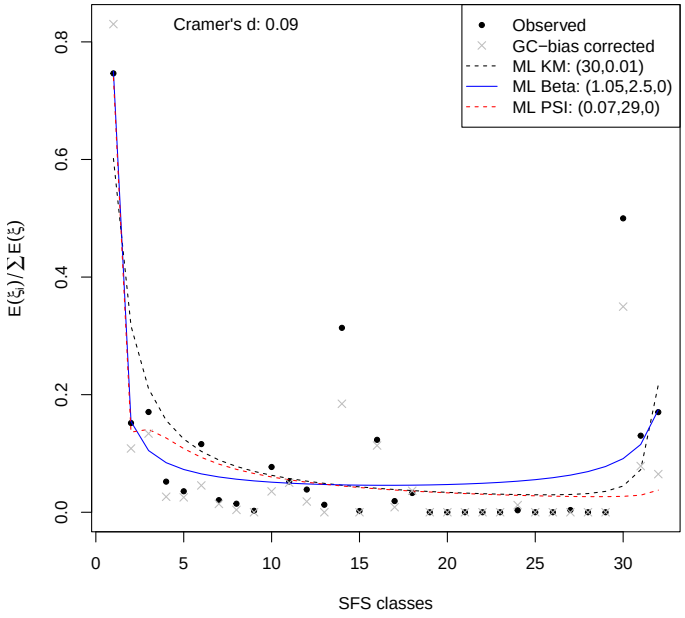
Staphylococcus aureus



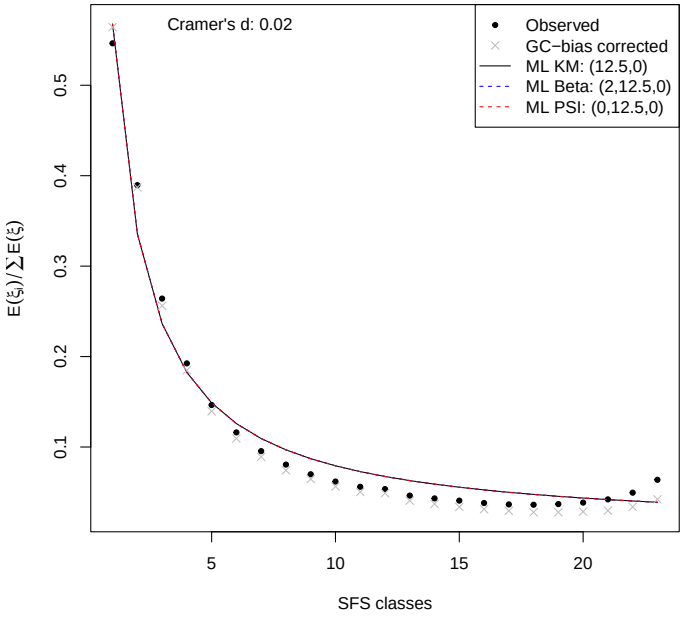
Streptococcus pneumoniae



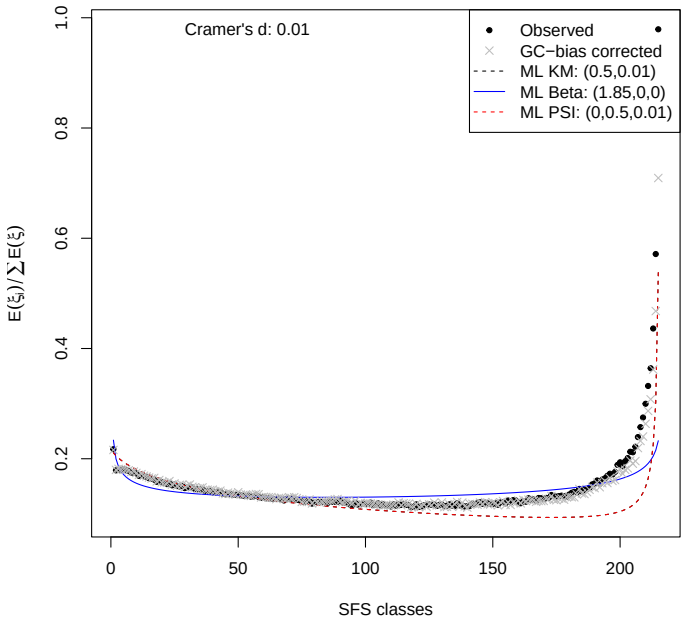
Mycobacterium tubercolosis



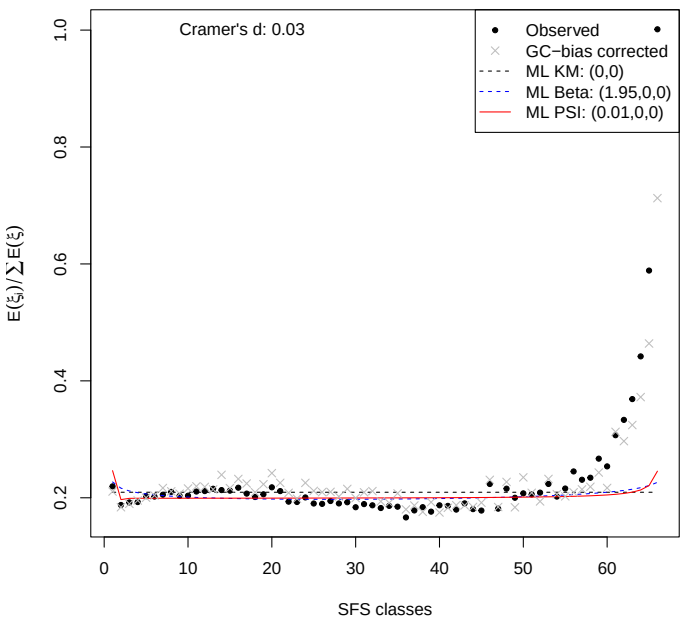
Phylloscopus trochilus

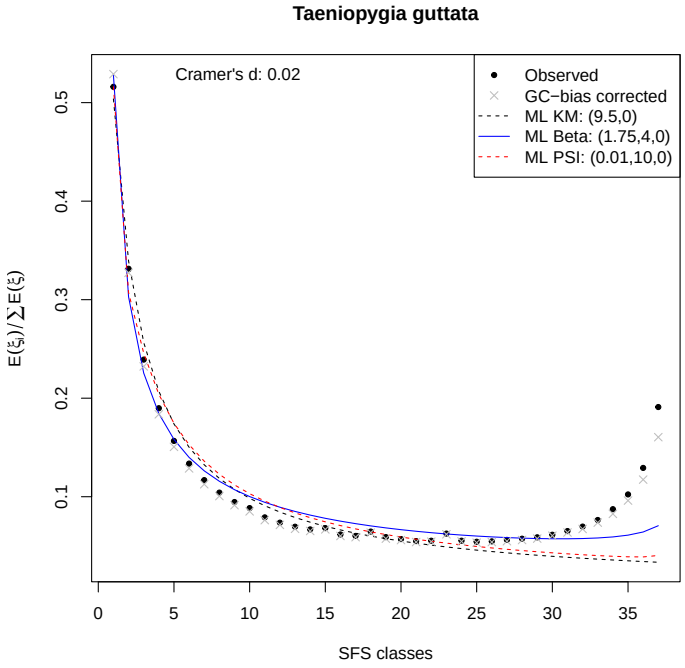


Homo sapiens

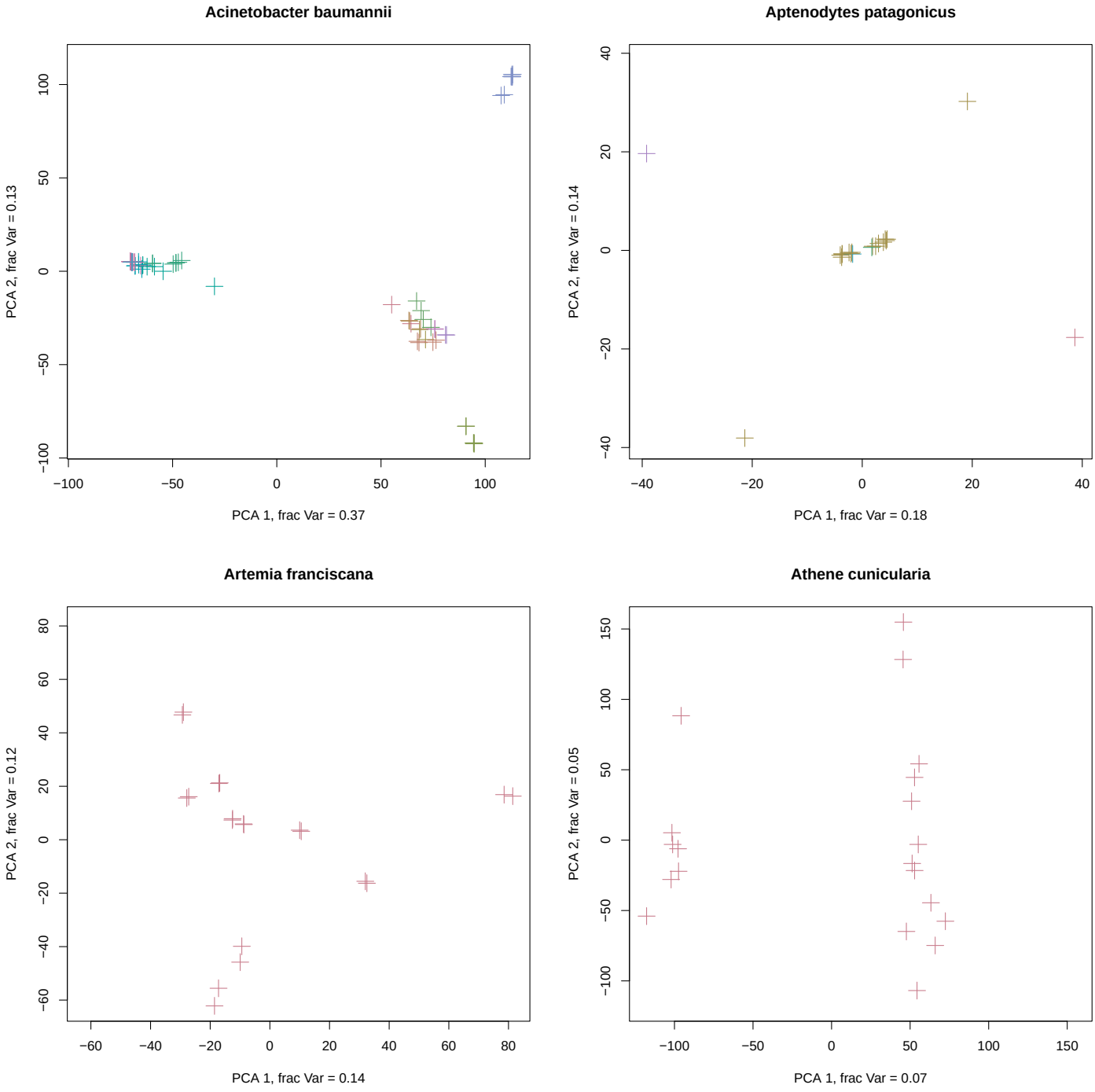


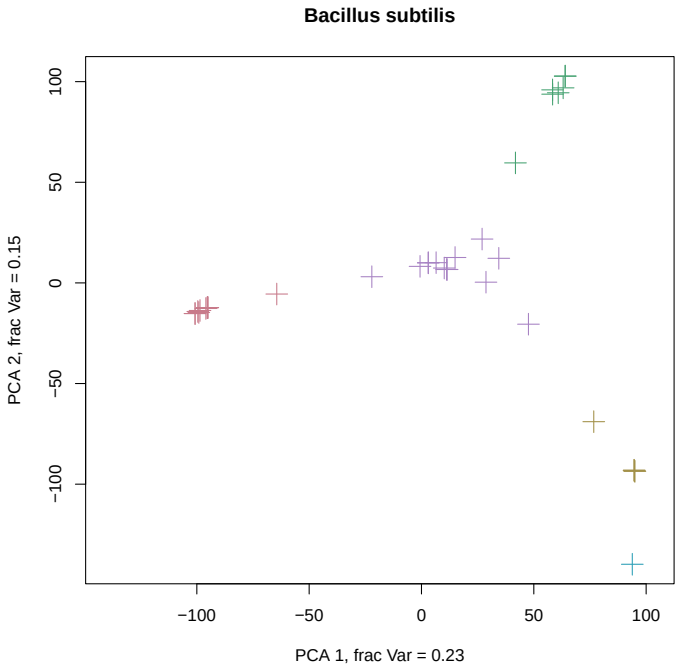
Zea Mays

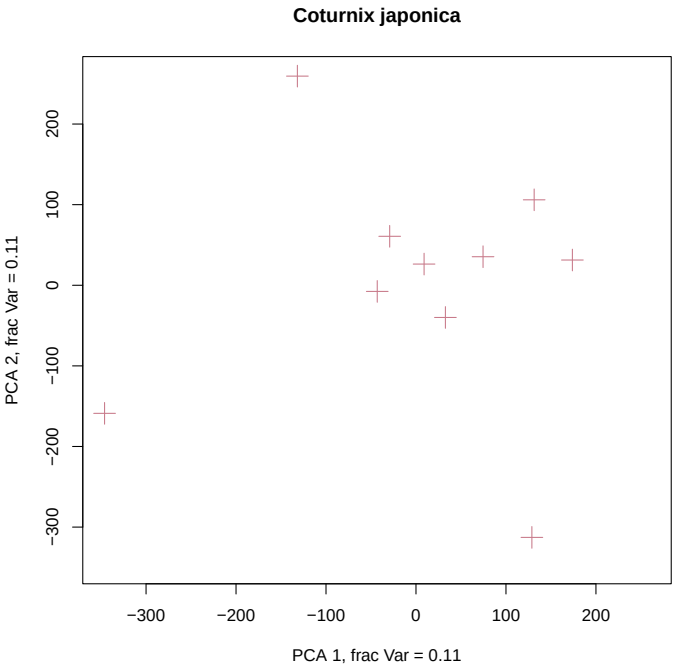


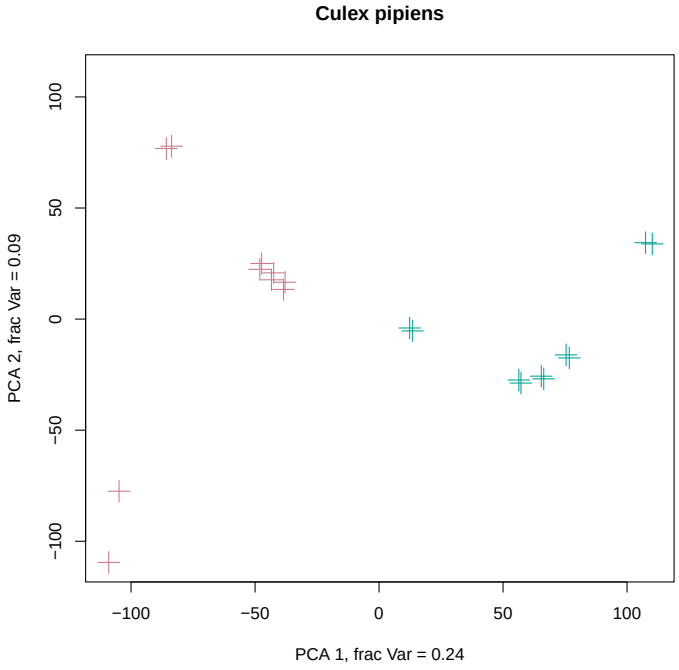

Taeniopygia guttata
Cramer's d: 0.02
Observed
GC−bias corrected
ML KM: (9.5,0)
ML Beta: (1.75,4,0)
ML PSI: (0.01,10,0)
E(ξi)/∑E(ξ)
0.1
0.2
0.3
0.4
0.5
0
5
10
15
20
25
30
35
SFS classes

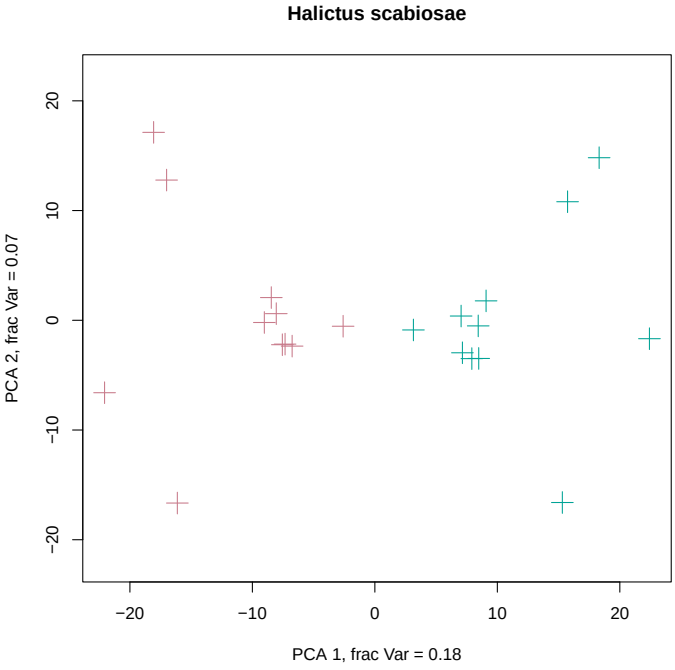
6.2.3 Supplementary file 3

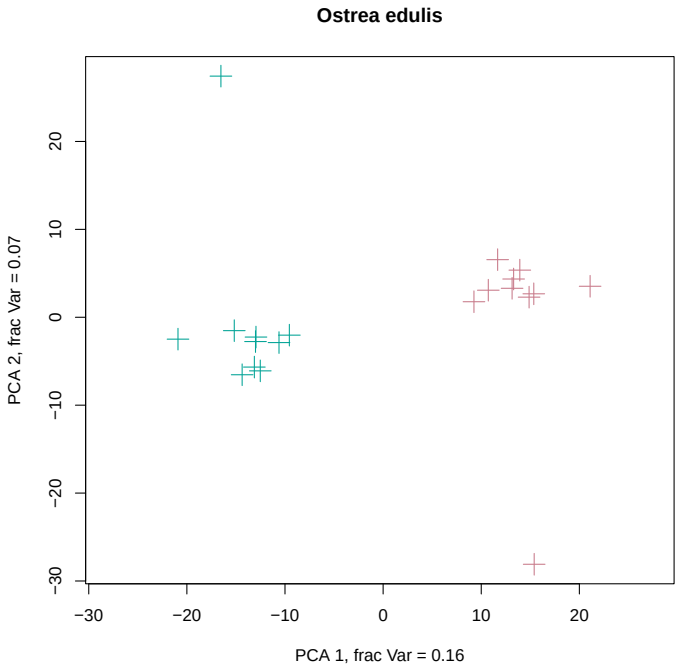


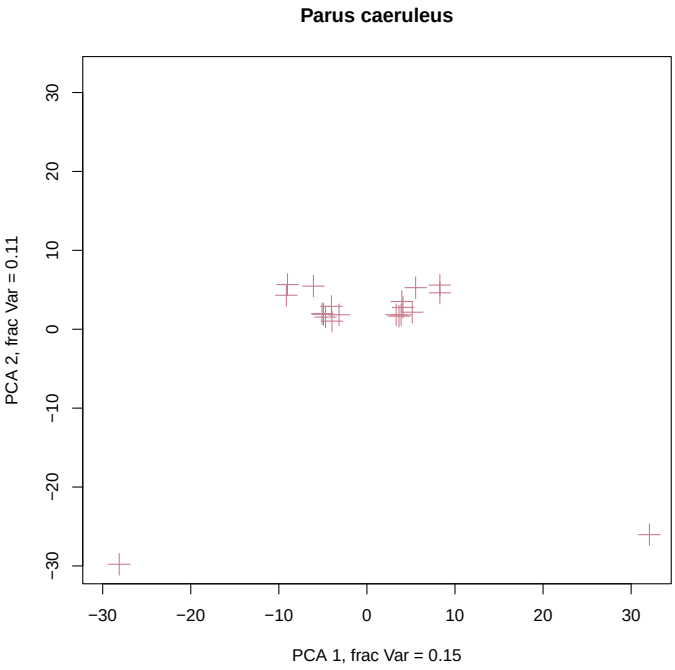

Bacillus subtilis
PCA 2, frac Var = 0.15
PCA 1, frac Var = 0.23
−100
−50
0
50
100

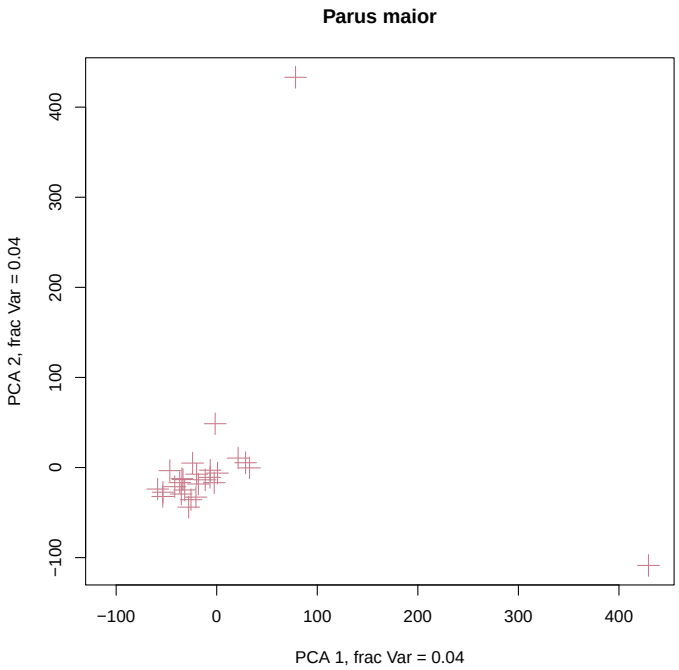

Coturnix japonica
PCA 2, frac Var = 0.11
PCA 1, frac Var = 0.11
−300
−200
−100
0
100
200

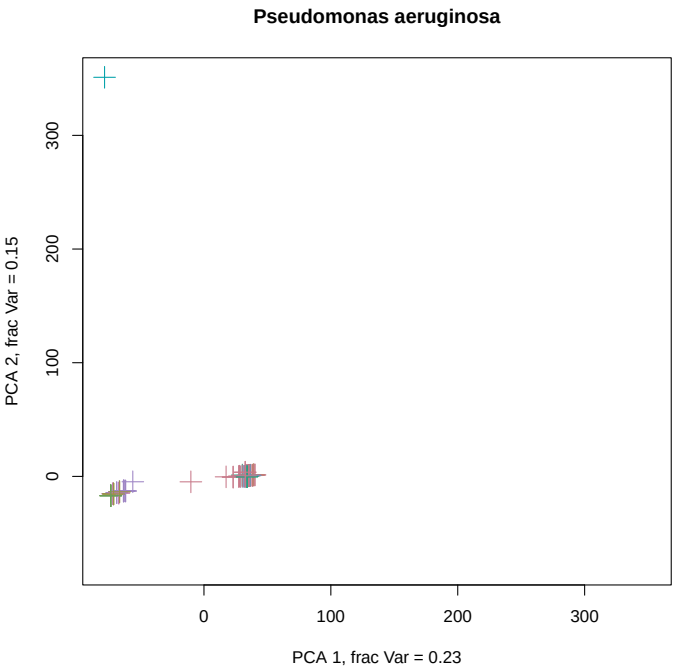

Culex pipiens
PCA 2, frac Var = 0.09
PCA 1, frac Var = 0.24
−100
−50
0
50
100


Halictus scabiosae
PCA 2, frac Var = 0.07
PCA 1, frac Var = 0.18
−20
−10
0
10
20


Ostrea edulis
PCA 2, frac Var = 0.07
PCA 1, frac Var = 0.16
−30
−20
−10
0
10
20


Parus caeruleus
PCA 2, frac Var = 0.11
PCA 1, frac Var = 0.15
−30
−20
−10
0
10
20
30


Parus maior
PCA 2, frac Var = 0.04
PCA 1, frac Var = 0.04
−100
0
100
200
300
400


Pseudomonas aeruginosa
PCA 2, frac Var = 0.15
PCA 1, frac Var = 0.23
0
100
200
300

Bibliographie

Guillaume Achaz. Testing for neutrality in samples with sequencing errors. *Genetics*, 179(3) :1409–1424, July 2008. ISSN 0016-6731. doi : 10.1534/genetics.107.082198.

Guillaume Achaz. Frequency spectrum neutrality tests : one for all and all for one. *Genetics*, 183(1) :249–58, Sep 2009.

Alison M Adams and Richard R Hudson. Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. *Genetics*, 168(3) :1699–712, Nov 2004.

Anthony D Barnosky, Nicholas Matzke, Susumu Tomiya, Guinevere O U Wogan, Brian Swartz, Tiago B Quental, Charles Marshall, Jenny L McGuire, Emily L Lindsey, Kaitlin C Maguire, Ben Mersey, and Elizabeth A Ferrer. Has the earth's sixth mass extinction already arrived ? *Nature*, 471(7336) :51–7, Mar 2011.

W. Bateson. Mendel's principles of heredity in mice. *Nature*, 67(1742) :462–463, Mar 1903. ISSN 1476-4687. doi : 10.1038/067462c0. URL http://dx.doi.org/10.1038/067462c0.

William Bateson et al. The progress of genetic research. In *Report of the Third International Conference on Genetics (1906)(ed. W. Wilks). London : Royal Horticultural Society*, pages 90–97, 1906.

Frédéric Baudat, Jérôme Buard, Corinne Grey, Adi Fledel-Alon, Carole Ober, Molly Przeworski, Graham Coop, and Bernard De Massy. Prdm9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science*, 327(5967) : 836–840, 2010.

Mark A Beaumont, Wenyang Zhang, and David J Balding. Approximate bayesian computation in population genetics. *Genetics*, 162(4) :2025–35, Dec 2002.

Peter Beerli and Joseph Felsenstein. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent

approach. *PNAS ; Proceedings of the National Academy of Sciences*, 98(8) : 4563–4568, 2001.

Annabel C. Beichman, Emilia Huerta-Sanchez, and Kirk E. Lohmueller. Using genomic data to infer historic population dynamics of nonmodel organisms. *Annual Review of Ecology, Evolution, and Systematics*, 49(1) :433–456, Nov 2018. ISSN 1545-2069.

Anand Bhaskar, Y.X. Rachel Wang, and Yun S. Song. Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. *Genome Research*, 25(2) :268–279, Jan 2015. ISSN 1549-5469. doi : 10.1101/gr.178756.114. URL http://dx.doi.org/10.1101/gr.178756.114.

Simon Boitard, Willy Rodríguez, Flora Jay, Stefano Mona, and Frédéric Austerlitz. Inferring population size history from large samples of genome-wide molecular data - an approximate bayesian computation approach. *PLOS Genetics*, 12(3) : e1005877, Mar 2016. ISSN 1553-7404. doi : 10.1371/journal.pgen.1005877. URL http://dx.doi.org/10.1371/journal.pgen.1005877.

Brian L Browning and Sharon R Browning. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*, 194(2) :459–71, Jun 2013.

Sharon R. Browning and Brian L. Browning. High-resolution detection of identity by descent in unrelated individuals. *Am J Hum Genet*, 86(4) :526 – 539, 2010. ISSN 0002-9297.

Sharon R Browning and Brian L Browning. Accurate non-parametric estimation of recent effective population size from segments of identity by descent. *Am J Hum Genet*, 97(3) :404–18, Sep 2015.

Bradley J Cardinale, J Emmett Duffy, Andrew Gonzalez, David U Hooper, Charles Perrings, Patrick Venail, Anita Narwani, Georgina M Mace, David Tilman, David A Wardle, Ann P Kinzig, Gretchen C Daily, Michel Loreau, James B Grace, Anne Larigauderie, Diane S Srivastava, and Shahid Naeem. Biodiversity loss and its impact on humanity. *Nature*, 486(7401) :59–67, Jun 2012.

Francisco C. Ceballos, Peter K. Joshi, David W. Clark, Michèle Ramsay, and James F. Wilson. Runs of homozygosity : windows into population history and trait architecture. *Nature Reviews Genetics*, 19(4) :220–234, Jan 2018. ISSN 1471-0064. doi : 10.1038/nrg.2017.109. URL http://dx.doi.org/10.1038/nrg.2017.109.

Gerardo Ceballos, Paul R Ehrlich, Anthony D Barnosky, Andrés García, Robert M Pringle, and Todd M Palmer. Accelerated modern human-induced species losses : Entering the sixth mass extinction. *Science advances*, 1(5) :e1400253, Jun 2015.

N.H Chapman and E.A Thompson. A model for the length of tracts of identity by descent in finite random mating populations. *Theor Popul Biol*, 64(2) :141 – 150, 2003. ISSN 0040-5809.

Lounès Chikhi, Willy Rodríguez, Simona Grusea, Patrícia Santos, Simon Boitard, and Olivier Mazet. The iicr (inverse instantaneous coalescence rate) as a summary of genomic diversity : insights into demographic inference and model choice. *Heredity*, 120(1) :13–24, 01 2018.

Donald F Conrad, Mattias Jakobsson, Graham Coop, Xiaoquan Wen, Jeffrey D Wall, Noah A Rosenberg, and Jonathan K Pritchard. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nature Genetics*, 38(11) :1251–1260, 2006. doi : 10.1038/ng1911. URL https://doi.org/10.1038/ng1911.

The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571) :68–74, Oct 2015. ISSN 1476-4687.

The International HapMap Consortium. The international hapmap project. *Nature*, 426(6968) :789–796, Dec 2003. ISSN 1476-4687. doi : 10.1038/nature02168. URL http://dx.doi.org/10.1038/nature02168.

Graham Coop and Robert C. Griffiths. Ancestral inference on gene trees under selection. *Theoretical Population Biology*, 66(3) :219–232, 2004. ISSN 0040-5809. doi : https://doi.org/10.1016/j.tpb.2004.06.006. URL https://www.sciencedirect.com/science/article/pii/S0040580904000826.

Alex Coventry, Lara M. Bull-Otterson, Xiaoming Liu, Andrew G. Clark, Taylor J. Maxwell, Jacy Crosby, James E. Hixson, Thomas J. Rea, Donna M. Muzny, Lora R. Lewis, David A. Wheeler, Aniko Sabo, Christine Lusk, Kenneth G. Weiss, Humeira Akbar, Andrew Cree, Alicia C. Hawes, Irene Newsham, Robin T. Varghese, Donna Villasana, Shannon Gross, Vandita Joshi, Jireh Santibanez, Margaret Morgan, Kyle Chang, Walker Hale IV, Alan R. Templeton, Eric Boerwinkle, Richard Gibbs, and Charles F. Sing. Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nature Communications*, 1(1) :131, 2010. doi : 10.1038/ncomms1130. URL https://doi.org/10.1038/ncomms1130.

James Franklin Crow, Motoo Kimura, et al. An introduction to population genetics theory. *An introduction to population genetics theory.*, 1970.

Lucien Cuénot. La loi de mendel et l'hérédité de la pigmentation chez les souris. *Archives de Zoologie Expérimentale et Générale 3 Series*, 1902.

Charles Darwin. *On the origin of species by means of natural selection, or, The preservation of favoured races in the struggle for life.* London :John Murray, Albemarle Street,, 1859. URL https://www.biodiversitylibrary.org/item/122307. https ://www.biodiversitylibrary.org/bibliography/82303.

Hugo de Vries. *Die mutationstheorie. Versuche und beobachtungen über die entstehung von arten im pflanzenreich*. Veit & comp.„ Leipzig, 1901. doi : 10.5962/bhl.title.11336. URL https://www.biodiversitylibrary.org/bibliography/11336.

Yun Deng, Yun S. Song, and Rasmus Nielsen. The distribution of waiting distances in ancestral recombination graphs. *Theoretical Population Biology*, 2021. ISSN 0040-5809. doi : https://doi.org/10.1016/j.tpb.2021.06.003. URL https://www.sciencedirect.com/science/article/pii/S0040580921000484.

A. J. Drummond, A. Rambaut, B. Shapiro, and O. G. Pybus. Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences. *Molecular Biology and Evolution*, 22(5) :1185–1192, 02 2005. ISSN 0737-4038. doi : 10.1093/molbev/msi103. URL https://doi.org/10.1093/molbev/msi103.

Laurent Duret and Nicolas Galtier. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annual Review of Genomics and Human Genetics*, 10(1) :285–311, Sep 2009. ISSN 1545-293X. doi : 10.1146/annurev-genom-082908-150001. URL http://dx.doi.org/10.1146/annurev-genom-082908-150001.

Bjarki Eldon and John Wakeley. Coalescent Processes When the Distribution of Offspring Number Among Individuals Is Highly Skewed. *Genetics*, 172(4) : 2621–2633, 04 2006. ISSN 1943-2631. doi : 10.1534/genetics.105.052175. URL https://doi.org/10.1534/genetics.105.052175.

S. N. Ethier and R. C. Griffiths. On the two-locus sampling distribution. *Journal of Mathematical Biology*, 29(2) :131–159, 1990. doi : 10.1007/BF00168175. URL https://doi.org/10.1007/BF00168175.

W.J. Ewens. The sampling theory of selectively neutral alleles. *Theoretical Population Biology*, 3(1) :87–112, 1972. ISSN 0040-5809. doi : https://doi.org/10.1016/0040-5809(72)90035-4. URL https://www.sciencedirect.com/science/article/pii/0040580972900354.

Laurent Excoffier, Isabelle Dupanloup, Emilia Huerta-Sánchez, Vitor C. Sousa, and Matthieu Foll. Robust demographic inference from genomic and snp data. *PLoS Genetics*, 9(10) :e1003905, Oct 2013. ISSN 1553-7404. doi : 10.1371/journal.pgen.1003905. URL http://dx.doi.org/10.1371/journal.pgen.1003905.

J C Fay and C I Wu. Hitchhiking under positive darwinian selection. *Genetics*, 155 (3) :1405–1413, 07 2000. URL https://pubmed.ncbi.nlm.nih.gov/10880498.

Luca Ferretti, Filippo Disanto, and Thomas Wiehe. The effect of single recombination events on coalescent tree height and shape. *PLoS ONE*, 8(4) : e60123, Apr 2013. ISSN 1932-6203. doi : 10.1371/journal.pone.0060123. URL http://dx.doi.org/10.1371/journal.pone.0060123.

Luca Ferretti, Alice Ledda, Thomas Wiehe, Guillaume Achaz, and Sebastian E Ramos-Onsins. Decomposing the Site Frequency Spectrum : The Impact of Tree Topology on Neutrality Tests. *Genetics*, 207(1) :229–240, 07 2017. ISSN 1943-2631. doi : 10.1534/genetics.116.188763. URL https://doi.org/10.1534/genetics.116.188763.

Luca Ferretti, Alexander Klassmann, Emanuele Raineri, Sebastián E. Ramos-Onsins, Thomas Wiehe, and Guillaume Achaz. The neutral frequency spectrum of linked sites. *Theoretical Population Biology*, 123 :70–79, 2018. ISSN 0040-5809. doi : https://doi.org/10.1016/j.tpb.2018.06.001. URL https://www.sciencedirect.com/science/article/pii/S0040580917301399.

Ronald Aylmer Fisher. *The genetical theory of natural selection*. Oxford, The Clarendon Press, 1930.

Y X Fu. Statistical properties of segregating sites. *Theor Popul Biol*, 48(2) :172–97, Oct 1995.

Y X Fu and W H Li. Statistical tests of neutrality of mutations. *Genetics*, 133 (3) :693–709, 03 1993. ISSN 1943-2631. doi : 10.1093/genetics/133.3.693. URL https://doi.org/10.1093/genetics/133.3.693.

Jane Gibson, Newton E. Morton, and Andrew Collins. Extended tracts of homozygosity in outbred human populations. *Human Molecular Genetics*, 15 (5) :789–795, 01 2006. ISSN 0964-6906. doi : 10.1093/hmg/ddi493. URL https://doi.org/10.1093/hmg/ddi493.

G B Golding. The sampling distribution of linkage disequilibrium. *Genetics*, 108 (1) :257–274, 09 1984. ISSN 1943-2631. doi : 10.1093/genetics/108.1.257. URL https://doi.org/10.1093/genetics/108.1.257.

Robert C Griffiths and Paul Marjoram. An ancestral recombination graph. In *Progress in population genetics and human evolution*, pages 257 – 270. Springer, 1997. ISBN 0-387-94944-5.

Simona Grusea, Willy Rodríguez, Didier Pinchon, Lounès Chikhi, Simon Boitard, and Olivier Mazet. Coalescence times for three genes provide sufficient information to distinguish population structure from population size changes. *J Math Biol*, 78(1-2) :189–224, Jan 2019.

Alexander Gusev, Jennifer K Lowe, Markus Stoffel, Mark J Daly, David Altshuler, Jan L Breslow, Jeffrey M Friedman, and Itsik Pe'er. Whole population, genome-wide mapping of hidden relatedness. *Genome research*, 19(2) :318–26, Feb 2009.

Ryan N Gutenkunst, Ryan D Hernandez, Scott H Williamson, and Carlos D Bustamante. Inferring the joint demographic history of multiple populations from multidimensional snp frequency data. *PLoS Genet*, 5(10) :e1000695, Oct 2009.

J. B. S. Haldane. A mathematical theory of natural and artificial selection—i. *Bulletin of Mathematical Biology*, 52(1-2) :209–240, Jan 1924. ISSN 1522-9602. doi : 10.1007/bf02459574. URL http://dx.doi.org/10.1007/BF02459574.

J. B. S. Haldane. The effect of variation on fitness. *The American Naturalist*, 71 (735) :337–349, 1937. ISSN 00030147, 15375323. URL http://www.jstor.org/stable/2457289.

Godfrey H Hardy. Mendelian proportions in a mixed population. *Science*, 28(706) : 49–50, 1908.

Kelley Harris and Rasmus Nielsen. Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genet*, 9(6) :e1003521, Jun 2013.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1) :97–109, 04 1970. ISSN 0006-3444. doi : 10.1093/biomet/57.1.97. URL https://doi.org/10.1093/biomet/57.1.97.

Ben J Hayes, Peter M Visscher, Helen C McPartlan, and Mike E Goddard. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome research*, 13(4) :635–43, Apr 2003.

W. G. Hill and Alan Robertson. Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, 38(6) :226–231, Jun 1968. ISSN 1432-2242. doi : 10.1007/bf01245622. URL http://dx.doi.org/10.1007/BF01245622.

Melissa J. Hubisz, Daniel Falush, Matthew Stephens, and Jonathan K Pritchard. Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources*, 9(5) :1322–1332, 2009. doi : https://doi.org/10.1111/j.1755-0998.2009.02591.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1755-0998.2009.02591.x.

R R Hudson and N L Kaplan. Statistical properties of the number of recombination events in the history of a sample of dna sequences. *Genetics*, 111(1) :147–64, Sep 1985.

Richard R Hudson. Two-Locus Sampling Distributions and Their Application. *Genetics*, 159(4) :1805–1817, 12 2001. ISSN 1943-2631. doi : 10.1093/genetics/159.4.1805. URL https://doi.org/10.1093/genetics/159.4.1805.

Julian Huxley. *Evolution. The modern synthesis.* London : George Alien & Unwin Ltd., 1942.

IUCN. *IUCN Red List Categories and Criteria : Version 3.1*, 2012.

Flora Jay, Simon Boitard, and FrÃ©dÃ©ric Austerlitz. An ABC Method for Whole-Genome Sequence Data : Inferring Paleolithic and Neolithic Human Expansions. *Molecular Biology and Evolution*, 36(7) :1565–1579, 02 2019. ISSN 0737-4038. doi : 10.1093/molbev/msz038. URL https://doi.org/10.1093/molbev/msz038.

Wilhelm Johannsen. *Elemente der exakten Erblichkeitslehre*. Fischer, 1909.

Jerome Kelleher, Alison M Etheridge, and Gilean McVean. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Comput Biol*, 12(5) :e1004842, 05 2016.

Jerome Kelleher, Yan Wong, Anthony W. Wohns, Chaimaa Fadil, Patrick K. Albers, and Gil McVean. Inferring whole-genome histories in large population datasets. *Nature Genetics*, 51(9) :1330–1338, Sep 2019. ISSN 1546-1718. doi : 10.1038/s41588-019-0483-y. URL http://dx.doi.org/10.1038/s41588-019-0483-y.

Elise Kerdoncuff, Amaury Lambert, and Guillaume Achaz. Testing for population decline using maximal linkage disequilibrium blocks. *Theoretical Population Biology*, 134 :171–181, 2020. ISSN 0040-5809. doi : https://doi.org/10.1016/j.tpb.2020.03.004. URL https://www.sciencedirect.com/science/article/pii/S0040580920300289.

Yuseob Kim and Wolfgang Stephan. Detecting a Local Signature of Genetic Hitchhiking Along a Recombining Chromosome. *Genetics*, 160(2) :765–777, 02 2002. ISSN 1943-2631. doi : 10.1093/genetics/160.2.765. URL https://doi.org/10.1093/genetics/160.2.765.

Motoo Kimura. Evolutionary rate at the molecular level. *Nature*, 217(5129) :624–626, Feb 1968. ISSN 1476-4687. doi : 10.1038/217624a0. URL http://dx.doi.org/10.1038/217624a0.

Motoo Kimura. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Oct 1983. ISBN 9780511623486. doi : 10.1017/cbo9780511623486. URL http://dx.doi.org/10.1017/CBO9780511623486.

J.F.C. Kingman. The coalescent. *Stochastic Processes and their Applications*, 13 (3) :235–248, Sep 1982. ISSN 0304-4149.

A. Klassmann and L. Ferretti. The third moments of the site frequency spectrum. *Theoretical Population Biology*, 120 :16–28, 2018. ISSN 0040-5809. doi : https://doi.org/10.1016/j.tpb.2017.12.002. URL https://www.sciencedirect.com/science/article/pii/S004058091730028X.

Mary K Kuhner, Jon Yamato, and Joseph Felsenstein. Maximum Likelihood Estimation of Population Growth Rates Based on the Coalescent. *Genetics*, 149 (1) :429–434, 05 1998. ISSN 1943-2631. doi : 10.1093/genetics/149.1.429. URL https://doi.org/10.1093/genetics/149.1.429.

John Lamoreux, H Resit Akçakaya, Leon Bennun, Nigel J Collar, Luigi Boitani, David Brackett, Amie Bräutigam, Thomas M Brooks, Gustavo A.B da Fonseca, Russell A Mittermeier, Anthony B Rylands, Ulf Gärdenfors, Craig Hilton-Taylor, Georgina Mace, Bruce A Stein, and Simon Stuart. Value of the

iucn red list. *Trends in Ecology and Evolution*, 18(5) :214–215, 2003. ISSN 0169-5347. doi : https://doi.org/10.1016/S0169-5347(03)00090-9. URL https://www.sciencedirect.com/science/article/pii/S0169534703000909.

Marguerite Lapierre, Amaury Lambert, and Guillaume Achaz. Accuracy of demographic inferences from the site frequency spectrum : The case of the yoruba population. *Genetics*, 206(1) :439–449, 05 2017.

Hervé Le Guyader. *Penser l'évolution*. Éditions Actes Sud, 2012.

R C Lewontin. THE INTERACTION OF SELECTION AND LINKAGE. I. GENERAL CONSIDERATIONS ; HETEROTIC MODELS. *Genetics*, 49(1) :49–67, 01 1964. ISSN 1943-2631. doi : 10.1093/genetics/49.1.49. URL https://doi.org/10.1093/genetics/49.1.49.

R. C. Lewontin and Ken ichi Kojima. The evolutionary dynamics of complex polymorphisms. *Evolution*, 14(4) :458–472, 1960. ISSN 00143820, 15585646.

Heng Li and Richard Durbin. Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357) :493–496, Jul 2011. ISSN 1476-4687.

Xiaoming Liu and Yun-Xin Fu. Exploring population size changes using snp frequency spectra. *Nature Genetics*, 47(5) :555–559, Apr 2015. ISSN 1546-1718. doi : 10.1038/ng.3254. URL http://dx.doi.org/10.1038/ng.3254.

Xiaoming Liu and Yun-Xin Fu. Stairway plot 2 : demographic history inference with folded snp frequency spectra. *Genome Biology*, 21(1), Nov 2020. ISSN 1474-760X. doi : 10.1186/s13059-020-02196-9. URL http://dx.doi.org/10.1186/s13059-020-02196-9.

I M MacLeod, T H E Meuwissen, B J Hayes, and M E Goddard. A novel predictor of multilocus haplotype homozygosity : comparison with existing predictors. *Genetics research*, 91(6) :413–26, Dec 2009.

Iona M MacLeod, Denis M Larkin, Harris A Lewin, Ben J Hayes, and Mike E Goddard. Inferring demography from runs of homozygosity in whole-genome sequence, with correction for sequence errors. *Mol Biol Evol*, 30(9) :2209–23, Sep 2013.

Anna-Sapfo Malaspinas, Michael C. Westaway, Craig Muller, Vitor C. Sousa, Oscar Lao, Isabel Alves, Anders Bergström, Georgios Athanasiadis, Jade Y. Cheng, Jacob E. Crawford, and et al. A genomic history of aboriginal australia. *Nature*, 538(7624) :207–214, Sep 2016. ISSN 1476-4687. doi : 10.1038/nature18299. URL http://dx.doi.org/10.1038/nature18299.

Nina Marchi and Laurent Excoffier. Gene flow as a simple cause for an excess of high-frequency-derived alleles. *Evolutionary Applications*, 13(9) :2254–2263,

2020. doi : https://doi.org/10.1111/eva.12998. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/eva.12998.

Paul Marjoram and Jeff D Wall. Fast "coalescent" simulation. *BMC Genet*, 7 :16, Mar 2006.

Paul Marjoram, John Molitor, Vincent Plagnol, and Simon Tavaré. Markov chain monte carlo without likelihoods. *PNAS ; Proceedings of the National Academy of Sciences*, 100(26) :15324–15328, 2003.

Gabor T Marth, Eva Czabarka, Janos Murvai, and Stephen T Sherry. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics*, 166 (1) :351–72, Jan 2004.

O Mazet, W Rodríguez, S Grusea, S Boitard, and L Chikhi. On the importance of being structured : instantaneous coalescence rates and human evolution–lessons for ancestral population size inference ? *Heredity*, 116(4) :362–71, Apr 2016.

Kimberly F. McManus, Joanna L. Kelley, Shiya Song, Krishna R. Veeramah, August E. Woerner, Laurie S. Stevison, Oliver A. Ryder, Great Ape Genome Project, Jeffrey M. Kidd, Jeffrey D. Wall, Carlos D. Bustamante, and Michael F. Hammer. Inference of Gorilla Demographic and Selective History from Whole-Genome Sequence Data. *Molecular Biology and Evolution*, 32(3) : 600–612, 01 2015. ISSN 0737-4038. doi : 10.1093/molbev/msu394. URL https://doi.org/10.1093/molbev/msu394.

Ruth McQuillan, Anne-Louise Leutenegger, Rehab Abdel-Rahman, Christopher S. Franklin, Marijana Pericic, Lovorka Barac-Lauc, Nina Smolej-Narancic, Branka Janicijevic, Ozren Polasek, Albert Tenesa, Andrew K. MacLeod, Susan M. Farrington, Pavao Rudan, Caroline Hayward, Veronique Vitart, Igor Rudan, Sarah H. Wild, Malcolm G. Dunlop, Alan F. Wright, Harry Campbell, and James F. Wilson. Runs of homozygosity in european populations. *The American Journal of Human Genetics*, 83(3) :359–372, 2008. ISSN 0002-9297. doi : https://doi.org/10.1016/j.ajhg.2008.08.007. URL https://www.sciencedirect.com/science/article/pii/S000292970800445X.

G McVean. *Linkage disequilibrium, recombination and selection*. John Wiley and Sons, Ltd, 2008.

Gilean A T McVean. A Genealogical Interpretation of Linkage Disequilibrium. *Genetics*, 162(2) :987–991, 10 2002. ISSN 1943-2631. doi : 10.1093/genetics/162.2.987. URL https://doi.org/10.1093/genetics/162.2.987.

Gilean A T McVean and Niall J Cardin. Approximating the coalescent with recombination. *Philos Trans R Soc Lond B Biol Sci*, 360(1459) :1387–93, Jul 2005.

Gregor Mendel. Experiments in plant hybridization. *Verhandlungen des naturforschenden Vereins Brünn.) Available online : www. mendelweb. org/Mendel. html (accessed on 1 January 2013)*, 1865.

Valeria Montano. Coalescent inferences in conservation genetics : should the exception become the rule ? *Biology Letters*, 12(6) :20160211, Jun 2016. ISSN 1744-957X. doi : 10.1098/rsbl.2016.0211. URL http://dx.doi.org/10.1098/rsbl.2016.0211.

Camilo Mora, Derek P. Tittensor, Sina Adl, Alastair G. B. Simpson, and Boris Worm. How many species are there on earth and in the ocean ? *PLoS Biology*, 9(8) :e1001127, Aug 2011. ISSN 1545-7885. doi : 10.1371/journal.pbio.1001127. URL http://dx.doi.org/10.1371/journal.pbio.1001127.

P. A. P. Moran. Random processes in genetics. *Mathematical Proceedings of the Cambridge Philosophical Society*, 54(1) :60–71, 1958. doi : 10.1017/S0305004100033193.

Thomas Hunt Morgan, Alfred Henry Sturtevant, Hermann Joseph Muller, and Calvin Blackman Bridges. *The mechanism of Mendelian heredity*. Holt, 1915.

Krystyna Nadachowska-Brzyska, Reto Burri, Linnéa Smeds, and Hans Ellegren. Psmc analysis of effective population sizes in molecular ecology and its application to black-and-white ficedula flycatchers. *Molecular Ecology*, 25(5) :1058–1072, 2016. doi : https://doi.org/10.1111/mec.13540. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.13540.

Nobukazu Nawa and Fumio Tajima. Simple method for analyzing the pattern of dna polymorphism and its application to snp data of human. *Genes & Genetic Systems*, 83(4) :353–360, 2008. ISSN 1880-5779. doi : 10.1266/ggs.83.353. URL http://dx.doi.org/10.1266/ggs.83.353.

R Nielsen. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics*, 154(2) :931–942, 02 2000. URL https://pubmed.ncbi.nlm.nih.gov/10655242.

R. Nielsen. Genomic scans for selective sweeps using snp data. *Genome Research*, 15(11) :1566–1575, Nov 2005. ISSN 1088-9051. doi : 10.1101/gr.4252305. URL http://dx.doi.org/10.1101/gr.4252305.

T Ohta and M Kimura. Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population. *Genetics*, 68(4) : 571–580, 08 1971. URL https://pubmed.ncbi.nlm.nih.gov/5120656.

TOMOKO OHTA. Slightly deleterious mutant substitutions in evolution. *Nature*, 246(5428) :96–98, Nov 1973. ISSN 1476-4687. doi : 10.1038/246096a0. URL http://dx.doi.org/10.1038/246096a0.

Tomoko Ohta. Molecular evolution : Nearly neutral theory. *eLS*, 2013. doi : https://doi.org/10.1002/9780470015902.a0001801.pub4. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470015902.a0001801.pub4.

Pier Francesco Palamara, Todd Lencz, Ariel Darvasi, and Itsik Pe'er. Length distributions of identity by descent reveal fine-scale demographic history. *Am J Hum Genet*, 91(5) :809–22, Nov 2012.

Etienne Patin, Katherine J. Siddle, Guillaume Laval, Hélène Quach, Christine Harmant, Noémie Becker, Alain Froment, Béatrice Régnault, Laure Lemée, Simon Gravel, and et al. The impact of agricultural emergence on the genetic history of african rainforest hunter-gatherers and agriculturalists. *Nature Communications*, 5(1), Feb 2014. ISSN 2041-1723.

Hugh P Possingham, Sandy J Andelman, Mark A Burgman, Rodrigo A Medellin, Larry L Master, and David A Keith. Limits to the use of threatened species lists. *Trends in Ecology and Evolution*, 17(11) :503–507, 2002. ISSN 0169-5347. doi : https://doi.org/10.1016/S0169-5347(02)02614-9. URL https://www.sciencedirect.com/science/article/pii/S0169534702026149.

Fanny Pouyet, Simon Aeschbacher, Alexandre Thiéry, and Laurent Excoffier. Background selection and biased gene conversion affect more than 95of the human genome and bias demographic inferences. *eLife*, 7, Aug 2018. ISSN 2050-084X. doi : 10.7554/elife.36317. URL http://dx.doi.org/10.7554/eLife.36317.

Javier Prado-Martinez, Peter H. Sudmant, Jeffrey M. Kidd, Heng Li, Joanna L. Kelley, Belen Lorente-Galdos, Krishna R. Veeramah, August E. Woerner, Timothy D. O'Connor, Gabriel Santpere, and et al. Great ape genetic diversity and population history. *Nature*, 499(7459) :471–475, Jul 2013. ISSN 1476-4687.

Ivan Prates, Alexander T. Xue, Jason L. Brown, Diego F. Alvarado-Serrano, Miguel T. Rodrigues, Michael J. Hickerson, and Ana C. Carnaval. Inferring responses to climate dynamics from historical demography in neotropical forest lizards. *PNAS ; Proceedings of the National Academy of Sciences*, 113(29) :7978–7985, 2016.

R. B Primack. *A Primer of Conservation Biology*, volume Fifth Edition. Sinauer Associates, Sunderland, MA., 2012.

Jonathan K. Pritchard and Molly Przeworski. Linkage disequilibrium in humans : Models and data. *The American Journal of Human Genetics*, 69(1) :1–14, 2001. ISSN 0002-9297. doi : https://doi.org/10.1086/321275. URL https://www.sciencedirect.com/science/article/pii/S0002929707614396.

Jonathan K. Pritchard and Noah A. Rosenberg. Use of unlinked genetic markers to detect population stratification in association studies. *The American Journal of Human Genetics*, 65(1) :220–228, 1999. ISSN 0002-9297. doi : https://doi.org/10.

1086/302449. URL https://www.sciencedirect.com/science/article/pii/S000292970763746X.

Molly Przeworski. Estimating the Time Since the Fixation of a Beneficial Allele. *Genetics*, 164(4) :1667–1676, 08 2003. ISSN 1943-2631. doi : 10.1093/genetics/164.4.1667. URL https://doi.org/10.1093/genetics/164.4.1667.

Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A.R. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I.W. de Bakker, Mark J. Daly, and Pak C. Sham. Plink : A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, 81(3) :559 – 575, 2007. ISSN 0002-9297.

Aaron P Ragsdale and Simon Gravel. Unbiased Estimation of Linkage Disequilibrium from Unphased Data. *Molecular Biology and Evolution*, 37(3) : 923–932, 11 2019. ISSN 0737-4038. doi : 10.1093/molbev/msz265. URL https://doi.org/10.1093/molbev/msz265.

Aaron P Ragsdale and Ryan N Gutenkunst. Inferring Demographic History Using Two-Locus Statistics. *Genetics*, 206(2) :1037–1048, 06 2017. ISSN 1943-2631. doi : 10.1534/genetics.117.201251. URL https://doi.org/10.1534/genetics.117.201251.

Matthew D. Rasmussen, Melissa J. Hubisz, Ilan Gronau, and Adam Siepel. Genome-wide inference of ancestral recombination graphs. *PLOS Genetics*, 10 (5) :1–27, 05 2014. doi : 10.1371/journal.pgen.1004342. URL https://doi.org/10.1371/journal.pgen.1004342.

Claire Régnier, Guillaume Achaz, Amaury Lambert, Robert H Cowie, Philippe Bouchet, and Benoît Fontaine. Mass extinction in poorly known taxa. *Proceedings of the National Academy of Sciences*, 112(25) :7761–6, Jun 2015.

Ana S L Rodrigues, John D Pilgrim, John F Lamoreux, Michael Hoffmann, and Thomas M Brooks. The value of the iucn red list for conservation. *Trends in Ecology and Evolution*, 21(2) :71–6, Feb 2006.

Alan R Rogers and Chad Huff. Linkage Disequilibrium Between Loci With Unknown Phase. *Genetics*, 182(3) :839–844, 07 2009. ISSN 1943-2631. doi : 10.1534/genetics.108.093153. URL https://doi.org/10.1534/genetics.108.093153.

Zvi Rosen, Anand Bhaskar, Sebastien Roch, and Yun S Song. Geometry of the sample frequency spectrum and the perils of demographic inference. *Genetics*, 210(2) :665–682, 10 2018.

Maurizio Rossetto, Jia-Yee Samantha Yap, Jedda Lemmon, David Bain, Jason Bragg, Patricia Hogbin, Rachael Gallagher, Susan Rutherford, Brett Summerell, and Trevor C. Wilson. A conservation genomics workflow to guide practical management actions. *Global Ecology and Conservation*, 26 :e01492, 2021. ISSN

2351-9894. doi : https://doi.org/10.1016/j.gecco.2021.e01492. URL https://www.sciencedirect.com/science/article/pii/S2351989421000421.

P. C. Sabeti, S. F. Schaffner, B. Fry, J. Lohmueller, P. Varilly, O. Shamovsky, A. Palma, T. S. Mikkelsen, D. Altshuler, and E. S. Lander. Positive natural selection in the human lineage. *Science*, 312(5780) :1614–1620, 2006. ISSN 0036-8075. doi : 10.1126/science.1124309. URL https://science.sciencemag.org/content/312/5780/1614.

Enrique Santiago, Irene Novo, Antonio F PardiÃ±as, MarÃ-a Saura, Jinliang Wang, and Armando Caballero. Recent Demographic History Inferred by High-Resolution Analysis of Linkage Disequilibrium. *Molecular Biology and Evolution*, 37(12) :3642–3653, 07 2020. ISSN 0737-4038. doi : 10.1093/molbev/msaa169. URL https://doi.org/10.1093/molbev/msaa169.

Stephan Schiffels and Richard Durbin. Inferring human population size and separation history from multiple genome sequences. *Nat Genet*, 46(8) :919–25, Aug 2014.

Claudia Schmegner, Josef Hoegel, Walther Vogel, and Günter Assum. Genetic variability in a genomic region with long-range linkage disequilibrium reveals traces of a bottleneck in the history of the european population. *Human Genetics*, 118(2) :276–286, Sep 2005. ISSN 1432-1203. doi : 10.1007/s00439-005-0056-2. URL http://dx.doi.org/10.1007/s00439-005-0056-2.

Jason Schweinsberg. Coalescent processes obtained from supercritical galton–watson processes. *Stochastic Processes and their Applications*, 106(1) :107–139, 2003. ISSN 0304-4149. doi : https://doi.org/10.1016/S0304-4149(03)00028-0. URL https://www.sciencedirect.com/science/article/pii/S0304414903000280.

Thibaut Paul Patrick Sellinger, Diala Abu Awad, Markus Moest, and Aurélien Tellier. Inference of past demography, dormancy and self-fertilization rates from whole genome sequence data. *PLOS Genetics*, 16(4) :e1008698, Apr 2020. ISSN 1553-7404. doi : 10.1371/journal.pgen.1008698. URL http://dx.doi.org/10.1371/journal.pgen.1008698.

Thibaut Paul Patrick Sellinger, Diala Abu-Awad, and Aurélien Tellier. Limits and convergence properties of the sequentially markovian coalescent. *Molecular Ecology Resources*, n/a(n/a), 2021. doi : https://doi.org/10.1111/1755-0998.13416. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.13416.

M Slatkin. Linkage disequilibrium in growing and stable populations. *Genetics*, 137(1) :331–336, 1994. ISSN 0016-6731. URL https://www.genetics.org/content/137/1/331.

Montgomery Slatkin. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, 9(6) :477–485, Jun 2008. ISSN 1471-0064. doi : 10.1038/nrg2361. URL http://dx.doi.org/10.1038/nrg2361.

John Maynard Smith and John Haigh. The hitch-hiking effect of a favourable gene. *Genetical Research*, 23(1) :23–35, 1974. doi : 10.1017/S0016672300014634.

P. Stam. The distribution of the fraction of the genome identical by descent in finite random mating populations. *Genetical Research*, 35(2) :131–155, Apr 1980. ISSN 1469-5073.

V T Stefanov. Distribution of genome shared identical by descent by two individuals in grandparent-type relationship. *Genetics*, 156(3) :1403–10, Nov 2000.

Megan A. Supple and Beth Shapiro. Conservation of biodiversity in the genomics era. *Genome Biology*, 19(1), Sep 2018. ISSN 1474-760X. doi : 10.1186/s13059-018-1520-3. URL http://dx.doi.org/10.1186/s13059-018-1520-3.

Walter S Sutton. On the morphology of the chromoso group in brachystola magna. *The Biological Bulletin*, 4(1) :24–39, 1902.

Walter S Sutton. The chromosomes in heredity. *The Biological Bulletin*, 4(5) : 231–250, 1903.

J.A. Sved. Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theoretical Population Biology*, 2(2) :125–141, 1971. ISSN 0040-5809. doi : https://doi.org/10.1016/0040-5809(71)90011-6. URL https://www.sciencedirect.com/science/article/pii/0040580971900116.

F Tajima. Statistical method for testing the neutral mutation hypothesis by dna polymorphism. *Genetics*, 123(3) :585–95, Nov 1989.

Naoyuki Takahata. Molecular Clock : An Anti-neo-Darwinian Legacy. *Genetics*, 176(1) :1–6, 05 2007. ISSN 1943-2631. doi : 10.1534/genetics.104.75135. URL https://doi.org/10.1534/genetics.104.75135.

Simon Tavaré, David J. Balding, R. C. Griffiths, and Peter Donnelly. Inferring coalescence times from dna sequence data. *Genetics*, 145(2) :505–518, 1997. ISSN 0016-6731. URL https://www.genetics.org/content/145/2/505.

Jonathan Terhorst, John A Kamm, and Yun S Song. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat Genet*, 49 (2) :303–309, Feb 2017.

David L. Wagner, Eliza M. Grames, Matthew L. Forister, May R. Berenbaum, and David Stopak. Insect decline in the anthropocene : Death by a thousand cuts. *PNAS ; Proceedings of the National Academy of Sciences*, 118(2), 2021.

Jeffrey D. Wall. Recombination and the power of statistical tests of neutrality. *Genetical Research*, 74(1) :65–79, 1999. doi : 10.1017/S0016672399003870.

G A Watterson. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol*, 7(2) :256–76, Apr 1975. doi : 10.1016/0040-5809(75)90020-9.

Wilhelm Weinberg. ber den nachweis der vererbung beim menschen. *Jahres. Wiertt. Ver. Vaterl. Natkd.*, 64 :369–382, 1908.

August Weismann. *Das Keimplasma : eine theorie der Vererbung*. G. Fischer, 1892.

C Wiuf and J Hein. Recombination as a point process along sequences. *Theor Popul Biol*, 55(3) :248–59, Jun 1999.

Sewall Wright. Coefficients of inbreeding and relationship. *The American Naturalist*, 56(645) :330–338, Jul 1922. ISSN 1537-5323. doi : 10.1086/279872. URL http://dx.doi.org/10.1086/279872.

Sewall Wright. Evolution in mendelian populations. *Genetics*, 16(2) :97–159, 1931.

Shuai Zhan, Wei Zhang, Kristjan Niitepõld, Jeremy Hsu, Juan Fernández Haeger, Myron P. Zalucki, Sonia Altizer, Jacobus C. de Roode, Steven M. Reppert, and Marcus R. Kronforst. The genetics of monarch butterfly migration and warning colouration. *Nature*, 514(7522) :317–321, Oct 2014. ISSN 1476-4687. doi : 10.1038/nature13812. URL http://dx.doi.org/10.1038/nature13812.

Weihua Zhang, Andrew Collins, Jane Gibson, William J. Tapper, Sarah Hunt, Panos Deloukas, David R. Bentley, and Newton E. Morton. Impact of population structure, effective bottleneck time, and allele frequency on linkage disequilibrium maps. *PNAS ; Proceedings of the National Academy of Sciences*, 101(52) : 18075–18080, 2004.

Emile Zuckerkandl and Linus Pauling. Evolutionary divergence and convergence in proteins. In Vernon Bryson and Henry J. Vogel, editors, *Evolving Genes and Proteins*, pages 97–166. Academic Press, 1965. ISBN 978-1-4832-2734-4. doi : https://doi.org/10.1016/B978-1-4832-2734-4.50017-6. URL https://www.sciencedirect.com/science/article/pii/B9781483227344500176.