Inférences sur l'histoire des populations à partir de leur diversité génétique : étude de séquences démographiques de type fondation-explosion

Claire Calmet

Thèse présentée par Claire Calmet

pour obtenir le grade de

Docteur de l'Université Paris VI

spécialité Biologie des Populations

sur le thème

Inférences sur l'histoire des populations à partir de leur diversité génétique : étude de séquences démographiques de type fondation-explosion

Soutenance prévue le 16 décembre 2002, à l'Université Pierre et Marie Curie, après avis de

M. Mark Beaumont	Rapporteur
M. Denis Couvet	Rapporteur

devant un jury composé de

Mme Marie-Catherine Boisselier	Directrice de thèse
M. Mark Beaumont	Rapporteur
M. Denis Couvet	Rapporteur
M. Dominique Higuet	Examinateur
M. Michel Pascal	Examinateur
Mme Sarah Samadi	Examinatrice

Coordonnées actuelles de l'auteur

Claire CALMET
Laboratoire Biométrie et Biologie Évolutive CNRS-UMR 5558
Université Claude-Bernard Lyon-1
43 boulevard du 11 novembre 1918
69622 Villeurbanne cedex, France

Coordonnées du laboratoire de Doctorat

Service de Systématique CNRS-FR 1541
Museum National d'Histoire Naturelle
43 rue Cuvier
75005 PARIS, France

Abstract

Studying demography in a historical perspective can help understand evolutionary processes. Through their genealogical and mutational history, population samples at genetic markers record —with loss of information— this demographic history. This potential and the increasing ease of genotyping have recently motivated development of new statistical tools aimed at extracting demographic information from raw molecular data [55].

In this thesis, the Bayesian inference method proposed in 1999 by M. Beaumont [6] is extended to more general demographic and mutational models. As the original method, the extended one (i) is based on Kingman's coalescent with variable population size [75], (ii) uses Metropolis-Hastings algorithm [62] to sample from the posterior distribution of parameters of interest, and (iii) allows to analyse genetic data at several unliked microsatellite loci. The demographic model underlying the extended version is that of a sudden size change, immediately followed by an exponential size change until sampling time —instead of a monotonic size change. The mutational model supposed is a two-phase model [31]. It is the first time a fully-probabilistic method allows microsatellite mutations with amplitude greater than one.

The demographic and mutational model is explored. Simulated data sets are used to compare the posterior distribution of the parameters, for several historical scenarios : *e.g.* a stable size history, an exponential increase for a time period and a founder-flush history. A typology is proposed for posterior distribution. Advice is given about the sampling and genotyping effort in empirical studies that aim at using the method : a unique microsatellite marker can lead to a strongly structured posterior distribution. However, with monolocus samples, highest posterior density domains always comprise scenarios of several kinds (*e.g.* not only founder-flush, but also exponential decline or increase). A sample of moderate size (50 haploid genomes typed at 5 microsatellite markers), were shown to strongly support a founder-flush history (99% of the posterior sample), with parameter values used to simulate the data in the 95% highest posterior density domain. Consequences of the violation of some hypothesis underlying the method are discussed : the shape of the demographic explosion is shown to be especially important. It is established that simplifying a TPM mutation process by a SMM model can lead to the detection of a false genetic disequilibrium. Interestingly, the modelisation of the TPM allows to erase this false signal.

The method is succinctly applied to the study of two founder-flush histories : the introduction of cat *Felis catus* on the Kerguelen archipelago, and the introduction of the brown rat *Rattus norvegicus* on Brittany islands. It is first shown that the frequentist method of Cornuet and Luikart [24] does not detect any significant departure from mutation-drift equilibrium, despite the strong founder events these populations experienced. This is probably due to cancelling effects of the founder and flush, on the summary statistics the method is based on. Probably for the same reason, the Bayesian method does not detect any disequilibrium signal if a step-like size change is supposed. The foundation and subsequent explosion become both detectable if they are parameterized. However, correlations between parameters make it impossible to infer a single parameter with precision less than several orders of magnitude. Prior information on some parameters (*e.g.* the time of the foundation) considerably constraint the possible values of others (*e.g.* the mutation rate). This confirms the potential of populations with documented history, to indirectly estimate the parameters of a mutational model for microsatellite markers.

Résumé

L'étude de la démographie dans une perspective historique participe à la compréhension des processus évolutifs. Les données de diversité génétique sont potentiellement informatives quant au passé démographique des populations : en effet, ce passé est enregistré avec perte d'information par les marqueurs moléculaires, par l'intermédiaire de leur histoire généalogique et mutationnelle. L'acquisition de données de diversité génétique est de plus en plus rapide et aisée, et concerne potentiellement n'importe quel organisme d'intérêt. D'où un effort dans la dernière décennie pour développer les outils statistiques permettant d'extraire l'information démographique des données de typage génétique [55].

La présente thèse propose une extension de la méthode d'inférence bayésienne développée en 1999 par M. Beaumont [6]. Comme la méthode originale, (i) elle est basée sur le coalescent de Kingman avec variations d'effectif [75], (ii) elle utilise l'algorithme de Metropolis-Hastings [62] pour échantillonner selon la loi *a posteriori* des paramètres d'intérêt et (iii) elle permet de traiter des données de typage à un ou plusieurs microsatellites indépendants. La version étendue généralise les modèles démographique et mutationnel supposés dans la méthode initiale : elle permet d'inférer les paramètres d'un modèle de fondation-explosion pour la population échantillonnée et d'un modèle mutationnel à deux phases [31], pour les marqueurs microsatellites typés. C'est la première fois qu'une méthode probabiliste exacte incorpore pour les microsatellites un modèle mutationnel autorisant des sauts.

Le modèle démographique et mutationnel est exploré. L'analyse de jeux de données simulés permet d'illustrer et de comparer la loi *a posteriori* des paramètres pour des scénarios historiques : par exemple une stabilité démographique, une croissance exponentielle et une fondation-explosion. Une typologie des lois *a posteriori* est proposée. Des recommandations sur l'effort de typage dans les études empiriques sont données : un unique marqueur microsatellite peut conduire à une loi *a posteriori* très structurée. Toutefois, les zones de forte densité *a posteriori* représentent des scénarios de différents types. 50 génomes haploïdes typés à 5 marqueurs microsatellites suffisent en revanche à détecter avec certitude (99% de la probabilité *a posteriori*) une histoire de fondation-explosion tranchée (figure page 2). Les conséquences de la violation des hypothèses du modèle démographique sont discutées, ainsi que les interactions entre processus et modèle mutationnel. En particulier, il est établi que le fait de supposer un processus mutationnel conforme au modèle SMM, alors que ce processus est de type TPM, peut générer un faux signal de déséquilibre génétique. La modélisation des sauts mutationnels permet de supprimer ce faux signal.

La méthode est succinctement appliquée à l'étude de deux histoires de fondation-explosion : l'introduction du chat *Felis catus* sur les îles Kerguelen et celle du surmulot *Rattus norvegicus* sur les îles du large de la Bretagne. Il est d'abord montré que la méthode fréquentiste développée par Cornuet et Luikart (1996) ne permet pas de détecter les fondations récentes et drastiques qu'ont connu ces populations. Cela est vraisemblablement dû à des effets contraires de la fondation et de l'explosion, sur les statistiques utilisées dans cette méthode. La méthode bayésienne ne détecte pas non plus la fondation si l'on force une histoire démographique en marche d'escalier, pour la même raison. La fondation et l'explosion deviennent détectables si le modèle démographique les autorise. Toutefois, les dépendances entre les paramètres du modèle empêchent de les inférer marginalement avec précision. Toute information *a priori* sur un paramètre contraint fortement les valeurs des autres paramètres. Ce constat confirme le potentiel de populations d'histoire documentée pour l'estimation indirecte des paramètres d'un modèle de mutation des marqueurs.

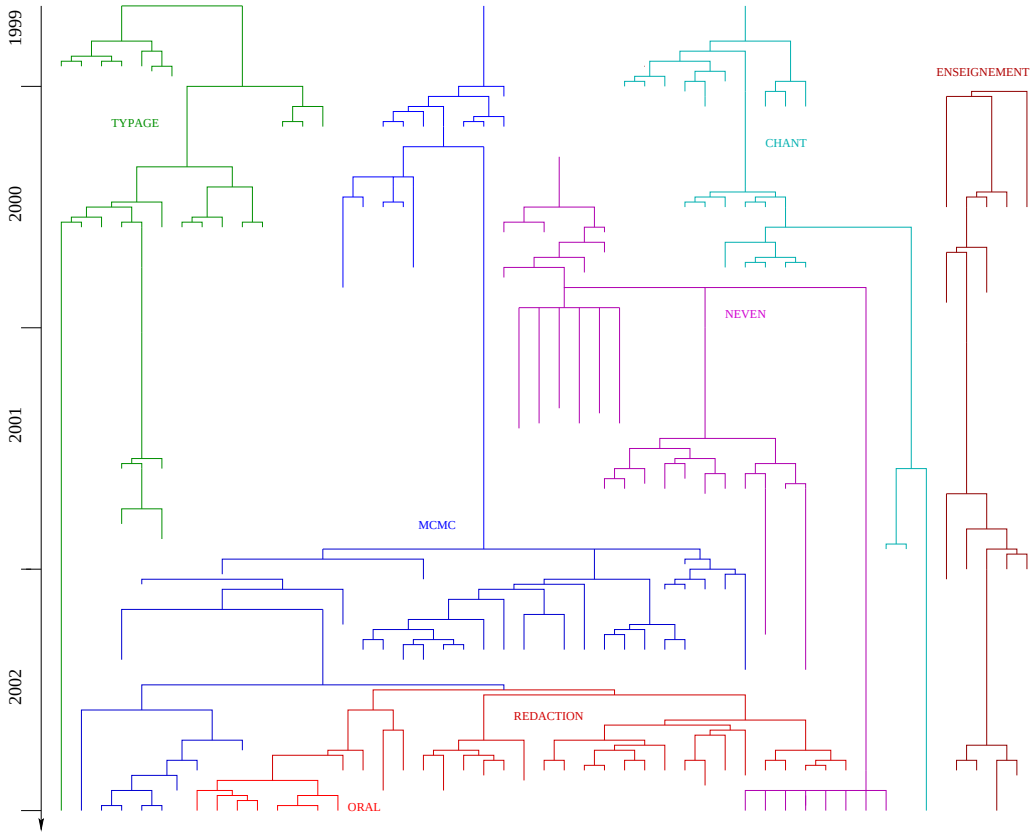


FIG. *1: Histoire généalogique de mes préoccupations non éteintes, tronquée au 01/09/2002. Toutes ont contribué à l'avancement et à l'achèvement de cette thèse, soit directement (TYPAGE, MCMC, REDACTION, ORAL), soit indirectement (ENSEIGNEMENT), soit de très loin (CHANT, NEVEN).*

FIG. *2: Ensemble $\mathcal{P}$ des personnes et organismes remerciés, et ensemble $\mathcal{Q}$ des remerciements exprimés. Exercice : compléter la schématisation de la relation de $\mathcal{P}$ vers $\mathcal{Q}$ à l'aide de flèches. Indices : la relation est une application, ni surjective, ni injective. Solution au verso.*

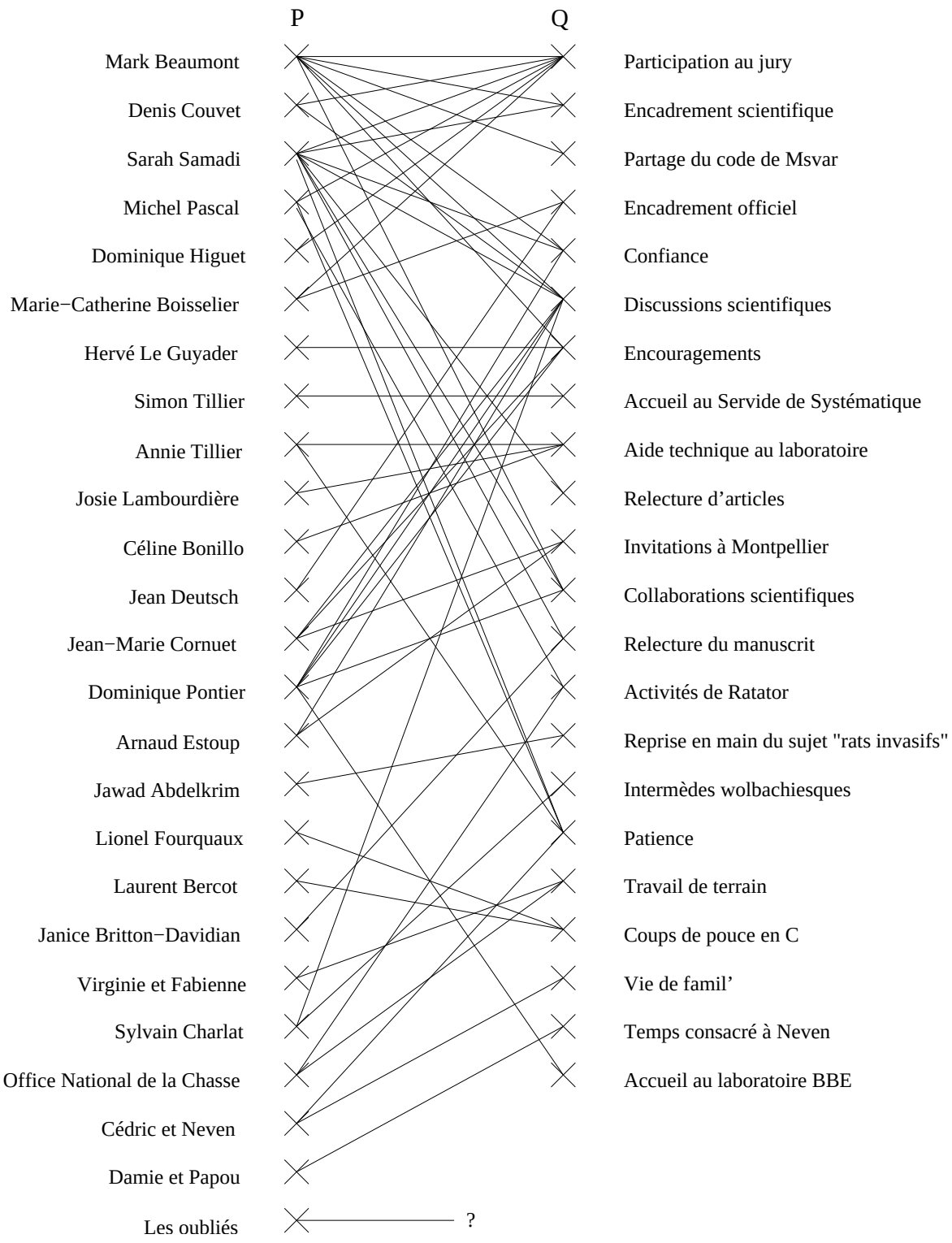


FIG. *3: Ensemble $\mathcal{P}$ des personnes et organismes remerciés, et ensemble $\mathcal{Q}$ des remerciements exprimés.*

Table des matières

Introduction

L'histoire du vivant est une succession de spéciations, de colonisations, d'invasions, de remplacements d'espèces par d'autres, d'extinctions [111]. La démographie des espèces, c'est à dire les variations de leurs effectifs, est au coeur de cette histoire, et de son interprétation en termes d'évolution. La théorie darwinienne de l'évolution repose en effet sur le fait que le vivant a une tendance intrinsèque à la croissance démographique alors que les ressources sont limitées, avec pour conséquence directe la sélection naturelle [27, 158]. Dans la théorie synthétique de l'évolution, les effectifs déterminent l'intensité de la force de tri neutre qu'est la dérive [61].

La compréhension des mécanismes de l'évolution suppose de considérer les espèces au sein des écosystèmes, et en particulier au sein des communautés d'espèces. La mise en place des communautés et leurs changements de composition, spatiaux et temporels, ont en effet des conséquences évolutives importantes [1]. La question de la dynamique des communautés est de plus motivée par des préoccupations de gestion de la diversité biologique. En effet, les écologues sont fortement sollicités pour prédire les conséquences d'introductions ou de disparition d'espèces dans les écosystèmes [1], dans un contexte de forte pression humaine sur l'environnement. Des études théoriques permettent d'explorer les conséquences possibles des changements de composition des communautés. Des études empiriques permettent ensuite d'identifier, parmi les effets possibles, ceux qui effectivement se produisent dans les populations naturelles. Les questions posées dans ces études empiriques sont de nature historique : comment se sont mises en place les relations entre espèces ? Quelles ont été les conséquences de la disparition de certains maillons des réseaux trophiques ? Les processus coévolutifs ont-il joué un rôle majeur dans l'établissement des communautés ? La démographie est de nouveau un élément de réponse clef : les effectifs des espèces en présence sont en effet des facteurs qui déterminent les pressions sélectives au sein des écosystèmes. Les conséquences de l'ajout ou du retrait d'une espèce sur les autres espèces d'une communauté sont de nature démographique (modification de leurs effectifs relatifs et absolus). Enfin, les effectifs sont le matériau sur lequel agissent les forces évolutives qui modèlent les communautés (mutation, dérive, sélection, migration).

L'étude de la démographie dans une perspective historique participe donc à la compréhension des processus évolutifs. Les milieux insulaires sont un laboratoire de l'évolution tout désigné pour la partie empirique de cette étude [3, 44, 11, 12]. Plusieurs caractéristiques qui leur sont propres peuvent en effet être mises à profit : (i) Les îles sont des espaces bien délimités, parfois de petite taille et d'habitat relativement homogène. (ii) Du fait de leur faible surface et de leur isolement, elles abritent des communautés simplifiées, plus faciles à modéliser. (iii) Les milieux insulaires connaissent des introductions volontaires ou fortuites d'espèces, qui souvent peuvent être datées et localisées. (iv) Ces introductions se soldent parfois par des événements d'invasion explosifs, et une évolution rapide des communautés. (v) Enfin, la richesse des communautés insulaires en espèces endémiques motive des tentatives de contrôle voire d'éradication des espèces

invasives, qui constituent des expériences grandeur nature.

Trois espèces du genre *Rattus* ont été introduites par l'homme dans plus de 80% des îles du monde, avec des conséquences souvent importantes sur les écosystèmes insulaires [3]. Paradoxalement, on sait peu de choses sur la dynamique de ces invasions, et sur les modalités de perturbation des écosystèmes. Pourtant, la diversité des types d'écosystèmes colonisés par ces rats et la présence dans le genre *Rattus* de plusieurs espèces invasives et non invasives, devrait permettre par une approche comparative d'identifier des facteurs qui favorisent l'invasion.

Lors de mon DEA, je me suis intéressée aux populations introduites du surmulot *Rattus norvegicus* dans des îles bretonnes, parmi lesquelles des réserves naturelles. Les questions posées étaient les suivantes. Depuis quand *R. norvegicus* est-il présent sur ces îles ? Quelles ont été les modalités d'introduction ? Les introductions ont-elles coincidé avec le déclin d'espèces autochtones ? Répondre à ces questions nécessite de croiser les résultats d'approches variées : le fonctionnement actuel des communautés doit être étudié ; les rares témoignages doivent être entendus ; les signatures des événements passés doivent être recherchées : restes fossiles éventuels, perturbations de la faune parasitaire, signatures génétiques des variations d'effectif passées. Un travail préliminaire a consisté à caractériser génétiquement les populations. La différenciation d'une île à l'autre à des marqueurs microsatellites s'est avérée extrêmement forte, et la diversité sur la plupart des îles très réduite. De ces informations génétiques, il a été possible de déduire que les populations insulaires ont vraisemblablement été fondées indépendamment par un petit nombre d'individus, sans migrations ultérieures efficaces. L'éradication du surmulot programmée sur certaines îles avait donc de bonnes chances d'être durable, en l'absence de recolonisation depuis les îles voisines.

Toutefois, l'étude de structuration génétique ne permettait pas d'exclure que la faible diversité et la différenciation observées soit dues à une dérive génétique post-fondation. Seule une modélisation de l'histoire généalogique des gènes pouvait permettre de trancher et de quantifier les effets relatifs sur la diversité génétique (i) de la diversité dans les populations sources, (ii) du nombre de fondateurs et (iii) de la forme de la reprise démographique. La variabilité génétique observée dans une population est potentiellement très informative car elle enregistre —avec perte d'information— l'histoire démographique de la population (variations d'effectif passées, migrations, événements sélectifs), son fonctionnement (systèmes de reproduction, cycle de vie de l'espèce) et l'histoire mutationnelle des marqueurs moléculaires.

On peut reconstituer plus précisément une histoire démographique si l'on dispose d'informations sur le processus mutationnel. À l'inverse, les populations dont l'histoire démographique est bien connue peuvent nous informer sur le processus mutationnel. En ce qui concerne les populations insulaires bretonnes du surmulot, les données démographiques non génétiques sont plus ou moins complètes. La date de l'introduction et l'effectif précis au moment de l'échantillonnage sont tous deux connus pour l'île Trielen. Cela suggère la possibilité d'inférer les paramètres d'un modèle mutationnel d'après la population de rats de Trielen, puis d'utiliser cette information pour mener des inférences démographiques sur les autres îles, avec des motivations de biologie de la conservation.

Pour cela, une méthode statistique autorisant l'incorporation des informations *a priori*, et des incertitudes sur ces informations, est nécessaire. Les méthodes bayésiennes, dont le principe fondateur est ancien (Bayes, 1701-1761 ; Laplace 1749-1827) sont toutes désignées pour cela. Au début de ma thèse, ces méthodes commençaient tout juste à faire leur apparition dans le domaine de l'inférence démographique. J'avais été particulièrement marquée pendant mon

DEA par un article de 1998 de I. Wilson et D. Balding [154], qui traitait de populations stables, de microsatellites liés, mais me semblait faire des choix intéressants du point de vue de la méthodologie statistique. Un an plus tard, un article de M. Beaumont [6] présentait une méthode proche, mais permettant le traitement de loci microsatellites indépendants, avec un modèle démographique autorisant une variation d'effectif. Cette méthode, implémentée dans le programme `msvar0.4.2b.c`, était inutilisable en l'état sur mes données, car elle supposait un modèle de variation d'effectif monotone, alors que les populations insulaires du rat ont manifestement connu un fort effet fondateur avant l'invasion.

Le problème de l'estimation des paramètres d'une fondation-explosion n'est pas spécifique aux cas discutés ci-dessus. Il se pose fréquemment dans la littérature d'écologie évolutive [24, 83, 86, 122, 146], en particulier en biologie de la conservation. L'objectif de cette thèse est de contribuer à l'étude de ce problème général, en proposant une méthode pour l'inférence de variations d'effectif de type fondation-explosion dans les populations, à partir de données de typage à des marqueurs très utilisés, les microsatellites. Cette méthode est une extension de celle proposée en 1999 par M. Beaumont [6]. Les conséquences des fondation-explosion sur la diversité des microsatellites y sont étudiées, et les limites liées aux jeux de données et aux modèles démographique et mutationnel y sont envisagées. La méthode y est appliquée à l'étude de deux histoires de fondation-explosion documentées : l'introduction du chat sur l'archipel des Kerguelen et celle du rat dans les îles du plateau continental breton.

Première partie

Motivations et outils pour l'étude des fondation-explosion

Une fondation-explosion consiste en une diminution d'effectif (ou fondation, ou crash), suivie d'une explosion démographique (ou flush). Une fondation-explosion se produit chaque fois que quelques particules infectieuses colonisent un hôte et s'y multiplient, quand une nouvelle aire géographique est colonisée avec succès par un petit groupe d'individus, ou encore lorsque les conditions extérieures ramener temporairement les effectifs d'une population à quasiment zéro. L'effondrement peut ainsi être lié à des contraintes environnementales, à des interactions entre espèces ou à des mouvements migratoires. La reprise démographique est une tendance intrinsèque des espèces placées dans un milieu favorable. Selon les espèces, les fondation-explosion peuvent être des accidents isolés, ou au contraire être la règle dans un cycle vital (voir encarts page 25). Dans le chapitre 1, nous développerons les motivations pour l'étude de la génétique des fondation-explosion : ces événements sont fréquents dans l'histoire du vivant, et peuvent avoir des conséquences évolutives importantes.

L'étude de la génétique des fondation-explosion peut emprunter deux grandes voies. L'une est *exploratoire*, visant à comprendre les effets des variations d'effectif sur la diversité génétique et sa distribution ; l'autre est *inférentielle*, visant à déterminer d'après la diversité quelle a été l'histoire des variations d'effectif des populations (figure 4).

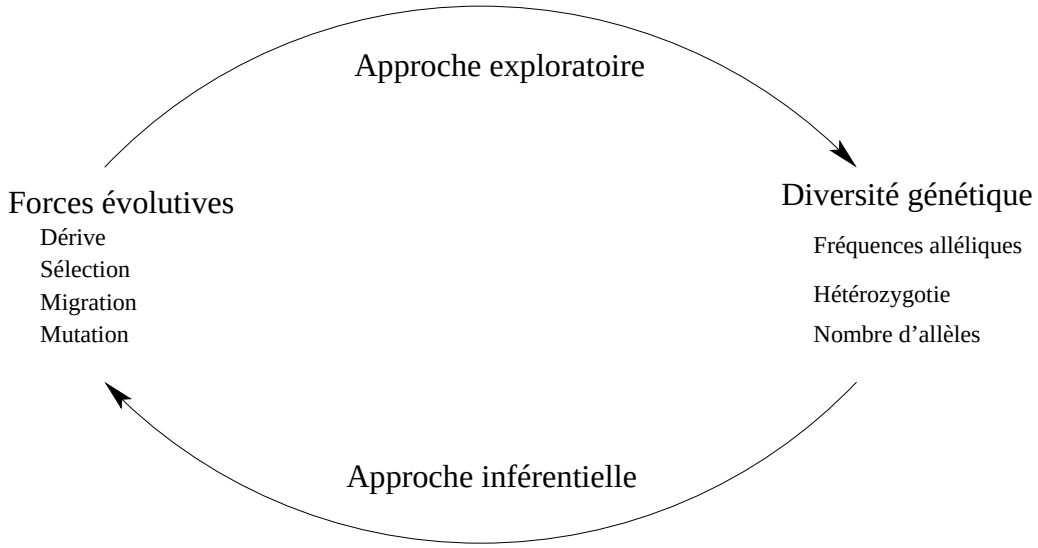


FIG. *4: Approches exploratoire et inférentielle pour étudier les relations entre forces évolutives et diversité biologique. Dans l'approche exploratoire, on quantifie les conséquences génétiques de différents régimes de forces évolutives. Dans l'approche inférentielle, on cherche à déduire les forces évolutives de l'observation de la diversité.*

Les études exploratoires (voir [83] et références incluses) ont montré que des caractéristiques importantes de la fondation sont la diversité dans la population d'origine des fondateurs, le nombre d'individus fondateurs et la forme de la reprise démographique, dont dépend *in fine* la constitution génétique de la population fondée. Lorsqu'il s'agit d'inférer les paramètres d'une fondation-explosion à partir de données de diversité génétique, des problèmes se posent. Premièrement, dans quelle mesure peut-on distinguer les effets génétiques d'une fondation-explosion, de ceux d'autres événements réducteurs de diversité comme les événements sélectifs ? Le choix de

plusieurs marqueurs moléculaires neutres, bien répartis dans le génome, permet d'augmenter la probabilité que les événements détectés soient bien de nature démographique [46]. Deuxièment, une fondation-explosion est la concaténation de variations d'effectif en sens contraire ; dans quelle mesure les signatures génétiques des deux événements interfèrent-elles ? Un jeu de données de taille moyenne contient-il assez d'information pour permettre des inférences précises sur les différentes phases d'une fondation-explosion ? Avant d'aborder ces questions dans la partie II, nous rappellerons dans le chapitre 2 le lien généalogique entre variations d'effectif et diversité génétique, lien qui fonde les méthodes d'inférence démogénétique. Nous présenterons au chapitre 3 les outils et méthodes actuellement disponibles, pour l'étude du lien entre diversité génétique et variations d'effectif.

Introductions en série chez le crapaud *Bufo marinus*

Le crapaud géant *Bufo marinus* est autochtone en Amérique équatoriale. Pour des motivations de lutte biologique, il a été introduit volontairement dans des îles tropicales, dès le début du $19^{\text{ième}}$ siècle. Il est actuellement invasif sur le continent australien, où il est parvenu au terme d'une histoire d'introductions en série remarquablement bien documentée [33]. Des individus originaires des Guyanes française et anglaise ont été introduits successivement aux îles Barbades (1833), à la Jamaïque (1844), puis à Puerto Rico (1923), sur les îles Hawaï (1932), et enfin au nord de l'Australie en 1935.
À chaque événement d'introduction, quelques dizaines d'individus sont concernés (*e.g.* 40 à Puerto-Rico, 150 sur Hawaï). Les quelques 10000 oeufs pondus annuellement par chaque femelle alimentent les premières générations, les plus nombreuses. Puis la population se stabilise à un effectif moindre. Une introduction comporte donc une fondation, une explosion extrêmement rapide, puis une stabilisation de la population à un effectif intermédiaire.

Les translocations d'espèces dues aux activités humaines sont une cause majeure de perturbation des écosystèmes [153, 123]. Il est donc important de comprendre la dynamique des invasions, et pour cela, de disposer de méthodes d'inférence sur le processus d'invasion (lire [39] pour une application au cas spécifique de *Bufo marinus*). Les paramètres intéressants à inférer peuvent être la date de l'introduction, le nombre d'individus fondateurs ou encore la dynamique de la reprise démographique [123].

Fondation-explosion et passage de l'hiver chez le puceron *Pemphigus bursarius*

Cette espèce de puceron des régions tempérées, responsable de la galle du peuplier, connaît un cycle annuel synchronisé sur les saisons, avec des variations d'effectif importantes. Au début du printemps, l'espèce n'existe que sous la forme d'oeufs qui ont passé l'hiver sous l'écorce des peupliers. Cette forme de résistance éclôt en autant de femelles fondatrices. Chacune injecte dans une feuille de peuplier un facteur de croissance qui entraîne la formation d'une galle. La fondatrice y met au monde par parthénogenèse de l'ordre de 200 filles ailées. Une fois adultes, celles-ci s'envolent, colonisent des champs de laitue, et continuent la reproduction parthénogénétique en produisant des femelles aptères et souterraines. En quelques générations, la population souterraine atteint des effectifs très importants. À la fin de l'été, la dernière génération sort de terre et développe des ailes, ce qui lui permet de rejoindre les peupliers. Elle produit parthénogénétiquement une génération d'individus mâles et femelles. C'est le seul stade du cycle annuel où les mâles apparaissent. Après fécondation, chaque femelle dépose un unique oeuf sous l'écorce d'un peuplier, ce qui permettra le passage de la mauvaise saison.

Cette espèce de puceron connaît donc annuellement des expansions clonales par parthénogenèse pendant la période de végétation, puis des goulots d'étranglement démographique pendant l'hiver. La compréhension des traits d'histoire de vie des espèces (*e.g.* alternance de la reproduction sexuée et asexuée dans ce cas) en relation avec les conditions de l'environnement nécessite une connaissance de leurs variations d'effectif.

Chapitre 1

Importance biologique des fondation-explosion

Les fondation-explosion sont intéressantes d'un point de vue théorique, car il s'agit événements démographiques aux conséquences génétiques plus complexes que des changements monotones d'effectif. Comprendre les conséquences génétiques des fondation-explosion a de plus des applications dans le domaine de la théorie de la spéciation et en biologie de la conservation. Enfin les populations ayant connu une fondation-explosion documentée fournissent l'opportunité d'estimer les paramètres de modèles mutationnels, à partir de données génétiques populationnelles.

1.1 Fondation-explosion et cladogenèse

On distingue classiquement deux types de processus évolutifs : l'anagenèse, qui correspond à l'évolution le long des lignées phylétiques, et la cladogenèse, ou bifurcation d'une seule lignée phylétique en plusieurs (typiquement deux). La cladogenèse, ou spéciation, donne par définition naissance à de nouvelles espèces. De nombreux modèles de spéciation, non exclusifs, ont été proposés (figure 1.1). Très schématiquement, on distingue [92, 100] des modèles qui mettent en jeu la disjonction entre les aires de répartition des futures espèces (modèles de spéciation allopatrique) et des modèles qui envisagent la cladogenèse sans séparation géographique (modèles de spéciation sympatrique). Dans tous les cas, les flux de gènes efficaces entre deux groupes doivent être interrompus pour donner des lignées définitivement divergentes. L'importance relative des deux grandes familles de spéciations dans la genèse de la biodiversité est difficilement évaluable, mais le mode allopatrique est plus facile à appréhender.

En allopatrie, le temps joue en faveur de la spéciation : par le jeu de la mutation, de la dérive et de la sélection, les groupes d'individus séparés divergent génétiquement sur l'ensemble de leurs caractères, et en particulier sur ceux qui déterminent la compatibilité reproductrice (comportement, morphologie, organisation du génome). L'isolement reproducteur est donc un sous-produit de la divergence. Les variations d'effectif déterminent la vitesse de la dérive, et le potentiel d'action de la sélection naturelle. C'est sur cette base qu'ont été proposées des précisions du modèle de spéciation allopatrique. Certaines supposent que les populations en cours de spéciation sont toutes deux de grande taille, à tous les instants de leur séparation jusqu'à la spéciation accomplie (modèles en "haltère", terme qui traduit l'équivalence entre les deux nouvelles espèces). La validité des modèles de spéciation allopatrique en haltère fait

l'unanimité, et le mécanisme en jeu est principalement la divergence adaptative [100]. Les données empiriques abondent sur des spéciations de ce type —les spéciations écologiques—, suite à la formation de barrières géographiques. Des modèles plus controversés supposent que l'une au moins des deux populations a subi une réduction d'effectif, transitoire mais drastique (modèles de spéciation par fondation-explosion). Des mécanismes de spéciation rapide faisant intervenir une ou plusieurs fondations ont été proposés successivement par E. Mayr [91, 93], H. Carson [19, 20] ou encore A. Templeton (*e.g.* [147]). Dans les modèles envisagés, un petit nombre d'individus fondateurs sont géographiquement isolés de leur population d'origine, de grande taille et d'évolution lente. Le groupe fondateur connaît une explosion démographique dans son aire géographique d'accueil. Effet d'échantillonnage initial et intensité (supposée) de la dérive génétique dans le groupe fondateur conduisent à une forte différenciation entre le groupe source et le groupe fondé. Des effets sélectifs peuvent provoquer une transition de pic adaptatif dans la population fondée, avec pour résultat l'isolement reproducteur. Un lien a été proposé entre la spéciation par fondation-explosion (à partir d'isolats périphériques d'une grande population) et la théorie des équilibres ponctués ([34, 54], mais []). Les périodes de stase et les ponctuations observées dans certains registres fossiles pourraient en effet correspondre au remplacement périodique d'une espèce de large effectif par une nouvelle espèce, fondée en périphérie par un petit groupe issu de l'espèce initiale, puis invasive. Barton [5] a pointé des faiblesses des modèles de spéciation par fondation-explosion tels que formulés par Mayr, Carson et Templeton, qui selon lui rendent ces modèles peu plausibles. Dans ces modèles, l'accent est le plus souvent mis sur les conséquences génétiques de la fondation-explosion : en particulier, les changements de pressions sélectives sont vus par E. Mayr comme résultant de l'augmentation d'homozygotie, par H. Carson comme dus à l'explosion démographique, et par A. Templeton comme des conséquences de la consanguinité. Le possible rôle de la sélection divergente par l'environnement a été négligé par les différents auteurs [21].

La question de savoir si les fondation-explosion, par leurs effets génétiques ou écologiques, favorisent la spéciation est une question ouverte. Pragmatiquement, déterminer la pertinence des modèles de fondation-explosion pour la formation de nouvelles espèces nécessite (i) une analyse théorique, (ii) des études expérimentales en laboratoire et (iii) des études empiriques dans des populations naturelles en cours de spéciation.

(i) Les modèles de spéciation par fondation-explosion cités précédemment manquent de rigueur [5], mais en 1996, M. Slatkin [133] a défendu dans un article théorique le principe selon lequel les fondation-explosion peuvent accélérer la spéciation. Pour cela, il a formulé dans des termes de génétique des populations classique les processus génétiques dans une population en croissance rapide, initialement fondée par un petit nombre d'individus. Il a montré qu'une fondation-explosion crée des conditions favorables à une évolution rapide de la population fondée, et donc éventuellement à la spéciation. La base de son raisonnement est la suivante. À la fondation, la dérive génétique est la force évolutive majeure, et rend invisibles les différences sélectives entre génotypes. Dans la phase de croissance, la dérive est au contraire relativement faible, même si l'effectif initial de la population est petit. En conséquence, la sélection sera fortement opérante pendant la croissance démographique. Des allèles ou des combinaisons d'allèles en faible fréquence initiale peuvent être fixés par sélection plus efficacement et rapidement qu'ils ne le seraient dans une population de taille constante (quelle que soit cette taille d'ailleurs, petite ou grande). Pour peu que la population post-fondation soit dans le domaine d'attraction d'un pic adaptatif différent de celui de la population source (par effet d'échantillonnage), la croissance démographique augmentera la probabilité de rejoindre effectivement ce pic adaptatif.

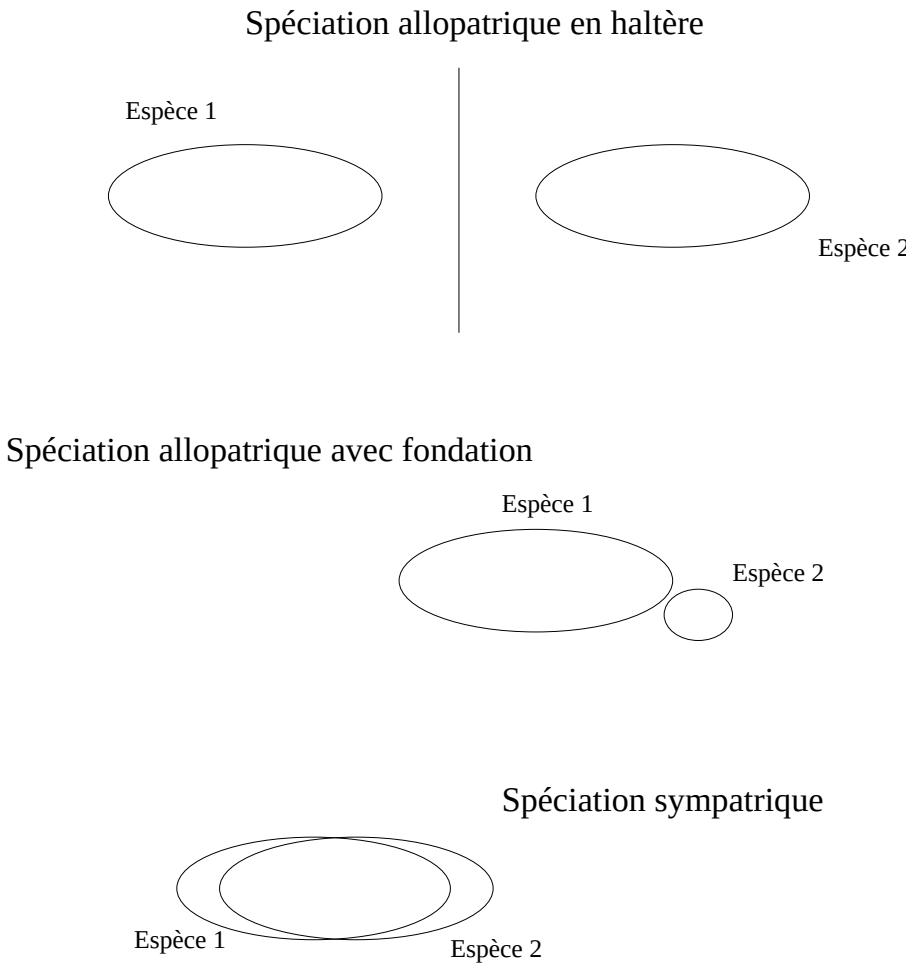


FIG. *1.1: Principaux modèles de spéciation. Dans la spéciation allopatrique en haltère, l'aire de répartition d'une espèce se trouve disjointe par une barrière géographique. Au moment de la séparation, les deux groupes ont des compositions génétiques proches. La dérive et la divergence adaptative conduisent avec le temps à une différenciation des deux futures espèces. Dans les modèles de spéciation allopatrique avec effet fondateur, un groupe de faible effectif est séparé du reste de l'espèce, et reste éventuellement en périphérie de son aire de répartition (spéciation péripatrique). Le groupe fondé est d'emblée génétiquement différencié de l'espèce d'origine. La sélection et la dérive accentuent cette différenciation. Une fois la spéciation accomplie, l'espèce dont l'effectif est demeuré important a peu divergé de l'espèce ancêtre, au contraire de l'espèce qui a connu l'effet fondateur. Dans les modèles de spéciation sympatrique, l'isolement reproducteur s'installe sans séparation géographique entre les deux futures espèces. Cet isolement peut être pré-zygotique (dates de floraison différentes, modifications du comportement de cour) ou post-zygotique (remaniements chromosomiques, changements de ploïdie).*

(ii) Les résultats expérimentaux (listés dans [121]), réalisés essentiellement sur la drosophile ne permettent pas de conclure clairement sur l'existence d'une relation entre fondation-explosion et spéciation.

(iii) La dernière section de l'article de M. Slatkin [133] donne l'idée d'un test pour la détection de séquences démographiques de type fondation-explosion. La méthode proposée utilise le fait que la fondation crée un fort déséquilibre de liaison entre positions du génome ayant des allèles rares dans la population source. La phase de croissance préserve ce déséquilibre, d'autant plus longtemps que la recombinaison entre sites est faible. Il est donc théoriquement possible de tester une histoire de fondation-explosion, par calcul du déséquilibre de liaison à des sites liés (idéalement les sites variables le long d'une même séquence). Toutefois, une comparaison avec la population source est nécessaire pour déterminer si les positions en déséquilibre y portent effectivement des allèles rares, ce qui rend le test inutilisable dans la plupart des cas. Comme annoncé, la partie II du présent document propose une méthode de détection des fondation-explosion qui exploite des données de comptage allélique à des marqueurs microsatellites non liés. Cette méthode pourrait servir à tester si des espèces en formation ont connu une fondation-explosion récente.

Au delà de l'étude des fondation-explosion, les progrès méthodologiques dans le domaine de l'inférence démographique permettent d'aborder des modèles de plus en plus proches de modèles de spéciation (notamment des modèles de fission de population, *e.g.* [98]). L'application récente de ces méthodes à des données empiriques a permis de faire le lien entre les relations phylogénétiques —à l'échelle interspécifique— et les relations généalogiques entre copies physiques d'un marqueur moléculaire —à toutes les échelles de l'intrapopulationnel à l'interspécifique [18, 79].

1.2 Fondation-explosion et biologie de la conservation

Les patrons de variation à des marqueurs moléculaires sont influencés par des facteurs démographiques importants pour le gestionnaire et le généticien de la conservation [82, 94], comme l'effectif efficace, les variations d'effectif et les flux migratoires. Les marqueurs génétiques sont donc potentiellement informatifs sur l'histoire et le fonctionnement des populations, intégrés sur le moyen et le long terme. Exploiter cette information nécessite le développement de méthodes d'inférence adaptées aux questions de biologie de la conservation. S'agissant par exemple des déclins de populations, les plus importants à détecter pour un généticien de la conservation sont les plus drastiques et les plus récents. En effet, ils sont les plus susceptibles de poser des problèmes génétiques aux populations concernées, notamment par mise à l'état homozygote d'allèles délétères récessifs. Si de tels déclins sont détectés à temps. des mesures d'amélioration de l'habitat ou l'apport d'immigrants peuvent encore être efficaces, pour éviter une trop forte perte d'hétérozygotie et de variation quantitative.

J.-M. Cornuet et G. Luikart [24] ont développé des tests pour l'identification de populations ayant connu un déclin. Ces tests sont implémentés dans le programme BOTTLENECK [105]. Ils utilisent le fait que suite à un déclin d'effectif, des allèles rares sont perdus, plus rapidement que les allèles fréquents [88, 89], ce qui conduit à un excès transitoire d'hétérozygotie par rapport à ce qu'impliquerait le nombre d'allèles dans une population à l'équilibre démographique et génétique. Ces tests ont été évalués sur des jeux de données simulés [24] et sur des données provenant de populations naturelles dont le déclin est documenté [86]. Leur puissance est bonne, et même très bonne pour les déclins intenses et relativement récents. Une méthode graphique, d'usage

élémentaire, a été proposée pour le même objectif [85]. Cette méthode se base sur la forme de la distribution de fréquences alléliques : dans une population à l'équilibre mutation-dérive, la majorité des allèles sont en faible fréquence. Dans une population ayant connu un déclin marqué, le mode de la distribution est déplacé vers une classe de fréquence intermédiaire. On dispose donc d'outils relativement efficaces et simples d'usage pour la détection de déclins. Cette efficacité dépend toutefois fortement du processus mutationnel des marqueurs moléculaires utilisés, et de la forme de la courbe démographique historique. Par exemple, lorsque un bottleneck est transitoire, la phase d'explosion qui suit le déclin risque d'effacer les signatures de ce déclin (*e.g.* [83] et résultats non-publiés de J.-F. Cosson). Or le conservateur peut souhaiter savoir si une population passe par des goulots d'effectif pendant lesquels le risque d'extinction pour des raisons démographiques est fort. La méthode présentée dans la partie II permet de détecter des bottlenecks même suivis de reprise démographique, et d'en inférer les caractéristiques (voir aussi [39]).

1.3 Fondation-explosion et information mutationnelle

L'étude des conséquences génétiques des variations d'effectif (ou plus généralement, des forces de tri de l'évolution), ne peut être séparée de celle de la mutation, source de la variation sur laquelle s'effectue le tri. La pertinence de bien des méthodes utilisées en génétique des populations dépend de la façon dont est modélisé le processus mutationnel. Malheureusement, celui-ci est difficile d'accès. On peut chercher à observer directement les mutations, par exemple par typage de paires mère-enfant ou par typage dans des lignées maintenues selon des modalités de croisement contrôlées [26, 130, 148]. Ces méthodes sont extrêmement lourdes. En effet, pour un marqueur de taux de mutation 10^{-5}, le typage de 10^6 paires mère-enfant conduira à l'observation de seulement 10 événements mutationnels en espérance, bénéfice bien maigre lorsqu'il s'agit de se faire une idée de la distribution des types de mutations [36, 157, 149, 63]. Les potentialités d'approches indirectes, populationnelles, pour s'informer sur le processus mutationnel des marqueurs, ont été tôt remarquées [66]. Elles sont particulièrement prometteuses dans les cas où l'histoire démographique des populations est connue, tout simplement parce que la diversité neutre dépend de la démographie et de la mutation, et que la connaissance de l'un réduit la variance des estimateurs de l'autre.

Les fondation-explosion sont doublement intéressantes dans ce contexte. Premièrement, elles font partie des histoires démographiques pour lesquelles on dispose de cas documentés (espèces introduites et invasives en milieu insulaire, réintroductions d'espèces protégées remarquablement bien suivies du point de vue démographique, avec parfois des prélèvements réguliers d'échantillons). Deuxièmement, leur démographie est beaucoup plus contraignante qu'une histoire d'effectif stable, du fait de la réduction de diversité au moment de la fondation. Une question se pose toutefois : dans quelle mesure les incertitudes sur la constitution de la population ancestrale, et sur les modalités de reprise démographique suite à la fondation minent-elles cet espoir de tirer de données populationnelles des informations sur les modalités de mutation des marqueurs ?

Chapitre 2

Approche démogénétique pour l'étude des fondation-explosion

2.1 Cadre conceptuel de la démogénétique

Les pionniers de la génétique des populations se sont d'emblée interessés au lien entre démographie et diversité génétique [88, 89]. Les conséquences génétiques des modes de reproduction, des variations d'effectif, ou de la structuration des populations ont été explorées et continuent de l'être. Toutefois, du simple fait de la complexité des études hors équilibre [57], l'essentiel des résultats de génétique des populations classique repose sur l'hypothèse d'un équilibre dynamique entre forces évolutives.

La construction par J.F.C. Kingman d'un objet mathématique qui décrit les généalogies de gènes —le coalescent standard, ou n-coalescent—, a considérablement simplifié l'étude des déséquilibres génétiques, d'un point de vue analytique aussi bien que numérique. L'approche par coalescence, qui considère l'histoire d'un échantillon de gènes en remontant dans le temps, a permis la réinterprétation simple de résultats obtenus à grand peine par des approches classiques (l'exemple d'usage est la réinterprétation de la formule d'échantillonnage de Ewens [61]). Le coalescent a relancé l'étude des relations entre le passé démographique des populations et leur diversité génétique, et a fondé la démogénétique moderne.

Jusqu'à présent, en démogénétique, la modélisation démographique est toujours restée simple (des démographes diraient même simpliste) : on n'inclut généralement pas explicitement des caractéristiques telles que la structuration en classes d'âge ou le sexe. La raison de cette simplification est double. (i) Dans ses débuts, une discipline comme la démogénétique traite préférentiellement de situations simples. Leur bonne compréhension fournit une base de travail et de comparaison pour des modèles plus réalistes. (ii) L'intégration des subtilités démographiques autrement que par l'usage de paramètres synthétiques est techniquement délicate. Elle est possible dans le domaine exploratoire (lire par exemple les thèses de F. Austerlitz [4] et de A. Sibert [132]), mais reste hors de portée dans les approches inférentielles qui nous concernent, pour des raisons —liées au calcul de vraisemblance— qui seront explicitées plus loin.

2.1.1 L'approche par coalescence, un changement de perspective

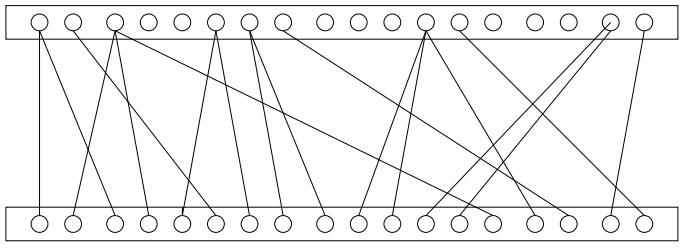
La génétique des populations, traditionnellement prospective

La génétique des populations classique pose des questions de type prospectif : quelle diversité peut-on attendre dans une population sous un régime de reproduction donné, pour un modèle de mutation, de migration, de sélection donné ? On s'est d'abord intéressé à la genèse de la variabilité génétique dans les populations en considérant la transmission du génome de parents à descendants. Cette approche respecte l'écoulement du temps dans le sens usuel, et permet de modéliser plus ou moins explicitement la reproduction. Le plus célèbre et le plus répandu des modèles démographiques utilisés pour modéliser l'évolution des populations est le modèle de Wright-Fisher (encart page 32). Ce modèle fait partie de la grande famille des modèles échangeables [76], pour lesquels tous les individus sont équivalents. L'approximation du comportement de ces modèles, notamment par l'usage de la théorie de la diffusion, a donné des résultats remarquables (travaux de l'école de Kimura ; voir aussi le livre de référence de Ewens [40]). Il faut distinguer les résultats qui concernent les populations dans leur ensemble, et les résultats qui concernent des échantillons tirés de ces populations. Ces derniers sont généralement plus délicats à obtenir, mais plus directement utilisables pour comprendre la diversité telle qu'on l'observe dans des échantillons de populations.

Le modèle de Wright-Fisher

La génétique des populations est une théorie mathématique de l'évolution, qui permet de quantifier les effets de forces évolutives sur la diversité des populations naturelles. Elle procède par la modélisation de ces forces évolutives, dans des populations théoriques nécessairement simplifiées, et par la comparaison des prédictions de ces modèles avec des données empiriques.

Le modèle de base de la génétique des populations, appelé modèle de Wright-Fisher, considère une population idéale aux caractéristiques suivantes : elle comporte un nombre N constant d'individus diploïdes (ce qui correspond à $2N$ copies de chaque gènes). Les générations ne sont pas chevauchantes, et chacune est générée selon des croisements pangamiques. Cela revient à ce que le gène parent de chaque gène de la génération $(n+1)$ soit tiré uniformément avec remise parmi les gènes de la population n. La population est close (elle ne recoit pas de migrants) et tous les individus ont la même espérance de reproduction (il n'y a pas de sélection naturelle).



Le modèle de Wright-Fisher fournit une base de comparaison pour des modèles plus réalistes, incluant une spatialisation par exemple. Ce modèle standard peut-être approché, lorsque N tend vers $+\infty$, par des modèles en temps continu : modèles de diffusion [156, 72] et coalescent de Kingman [77]. Dans ces deux limites du modèle de Wright-Fisher, l'unité de temps naturelle est $2N$ générations.

Le processus de coalescence, réinterprétation des modèles prospectifs discrets

Les travaux de Kingman au début des années 80 ont déclenché un véritable changement de point de vue de la génétique des populations, de prospectif à rétrospectif (voir [78, 144] pour un commentaire historique). Nous illustrons ici ce changement de perspective en réinterprétant le modèle prospectif de Wright-Fisher dans la logique coalescente.

Considérons une population de $2N$ gènes transmis de génération en génération conformément au modèle de Wright-Fisher. Chaque génération est formée par échantillonnage uniforme avec remise de $2N$ gènes parents à la génération précédente (encart page 32). Le nombre de gènes qui proviennent d'un gène particulier de la génération ancêtre est donc distribué selon une loi binomiale à $2N$ répétitions d'un événement de probabilité $1/2N$, ce qui mathématiquement équivaut à dire que la loi jointe du nombre de descendants produits par chacun des $2N$ gènes ancêtres est une loi multinomiale symétrique ([67], cité par [132]). Considérons maintenant les relations généalogiques obtenues par la répétition de ce processus sur plusieurs générations. De façon prospective, on observe la bifurcation des lignées lorsqu'un gène a strictement plus d'un descendant à la génération suivante, et l'extinction des lignées lorsqu'un gène ne laisse aucun descendant. Au bout d'un nombre suffisant de générations, seul un gène de la génération considérée au départ conserve des descendants. Cela correspond au processus de dérive. De façon rétrospective, la dérive revient à une fusion (ou coalescence) des lignées ancestrales de deux ou plusieurs gènes lorsqu'ils ont le même gène ancêtre (figure 2.1 page 34). Le nombre de lignées ancestrales d'un échantillon de n gènes ne peut donc que diminuer lorsqu'on remonte dans le temps, jusqu'à finalement atteindre 1, à partir de la génération de l'ancêtre commun le plus récent (MRCA) de l'échantillon.

Ajoutons à cette vision rétrospective du modèle de Wright-Fisher deux remarques et leurs conséquences méthodologiques. (i) Des variants neutres à un locus ont, par définition, le même succès reproducteur (en espérance) que les autres allèles à ce locus. On peut donc manipuler les arbres généalogiques à un locus dont les variants sont neutres sans se préoccuper des états alléliques le long des généalogies. L'affectation d'états alléliques peut se faire dans un second temps, par surimposition sur des généalogies vierges, selon un modèle mutationnel pertinent pour le type de marqueurs considéré. (ii) Seuls comptent, dans l'histoire de la population, les gènes qui ont effectivement eu des descendants dans l'échantillon que l'on considère. Les autres n'interviennent qu'en constituant les effectifs passés de la population, effectifs dont dépend tout de même la probabilité pour deux gènes d'avoir un ancêtre commun à la génération précédente. Ces deux remarques sont particulièrement importantes si l'on s'intéresse à la simulation de jeux de données. Une simulation selon le modèle de Wright-Fisher prospectif nécessite à chaque génération le tirage aléatoire de la totalité des individus à la génération suivante. Les individus héritent de la classe allélique de leurs ancêtres, éventuellement avec mutation, et on ne peut guère conserver les relations de parenté —sauf pour de petites populations—, pour des questions de limitation de mémoire. Une modélisation à rebours ne nécessite que le tirage des ancêtres de l'échantillon. Plus la taille de l'échantillon est faible, plus la taille de la population est grande, et meilleur est le gain. Toutefois, il est nécessaire de mémoriser les relations généalogiques lorsque l'on simule le processus de coalescence. En effet, les classes alléliques dans l'échantillon dépendent de la classe allélique de l'ancêtre, et des mutations éventuelles le long des généalogies. On a donc besoin des relations de parenté jusqu'au MRCA pour déterminer les classes alléliques dans un échantillon simulé.

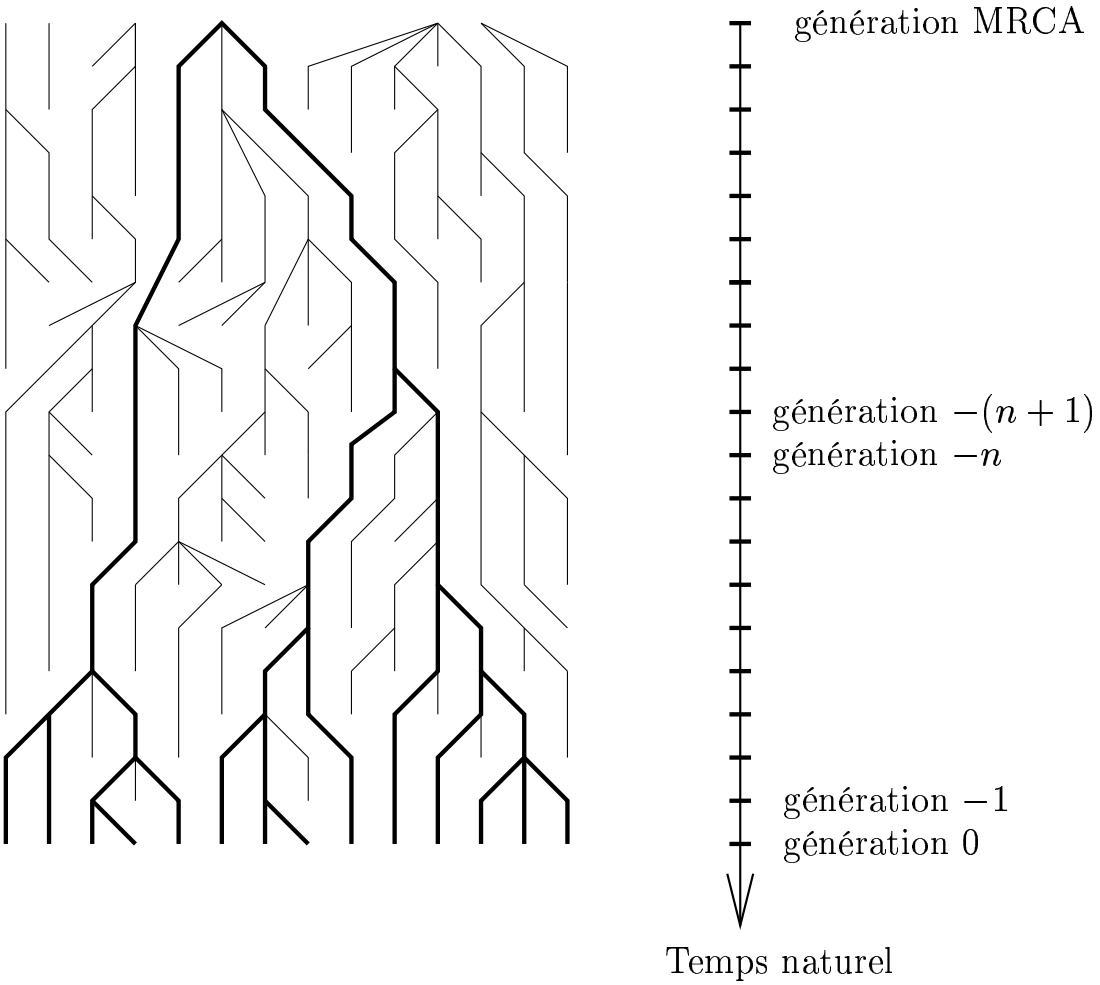


FIG. *2.1: Le modèle de Wright-Fisher décrit la transmission des gènes neutres de génération en génération dans une population idéale. Dans le sens naturel de l'écoulement du temps, le phénomène de dérive conduit inéluctablement à l'extinction des lignées descendantes d'un gène donné. Considérée à rebours du temps naturel, la dérive correspond au fait que toute la population de gènes partage un ancêtre commun (MRCA). Les lignées descendantes des gènes contemporains de l'ancêtre commun se sont éteintes successivement.*

2.1.2 Le coalescent standard, une approximation en temps continu

Le processus de coalescence tel que nous l'avons présenté à partir du modèle de Wright-Fisher est un processus discret, dans lequel des coalescences simultanées et multiples sont tout à fait possibles (figure 2.1). Ce n'est généralement pas sous cette forme que l'on décrit le processus de coalescence, mais sous une forme en temps continu, qui s'obtient comme limite de la forme discrète lorsque la taille de population tend vers $+\infty$: le coalescent standard ou n-coalescent. Une infime partie des propriétés connues du coalescent standard sera utilisée dans la méthode inférentielle présentée dans la partie II, et nous introduisons seulement les notions indispensables.

La généalogie d'un échantillon de gènes a deux composantes : sa topologie, c'est à dire la position relative des branches de la généalogies, et la date des événements de coalescence successifs. La topologie est simple à modéliser dans le cas neutre qui nous intéresse (figure 2.2). En effet, si la loi du nombre de descendants est la même pour tous les gènes, alors toutes les lignées sont équivalentes et ont la même probabilité de coalescer.

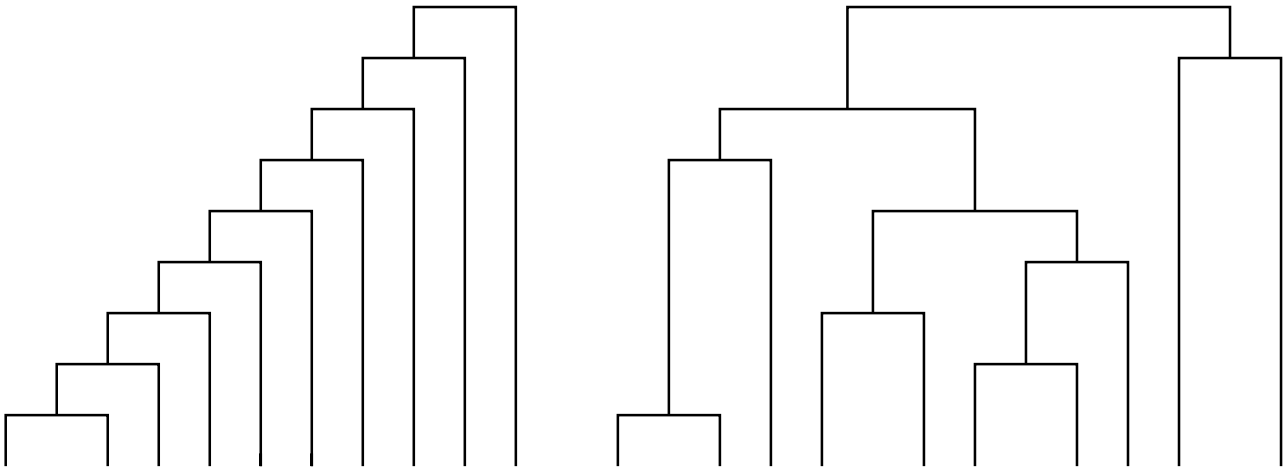


FIG. *2.2: Deux topologies possibles de l'arbre généalogique d'un échantillon de taille* 10 *gènes : à gauche, une topologie déséquilibrée —sous-arbres descendants en poupées russes—, à droite, une topologie mieux équilibrée. Pour un échantillon de n gènes, le nombre de topologies possibles est donné par* $\prod_{i=2}^{n} C_n^2$. *L'application numérique pour* $n = 10$ *donne* 2 571 912 000 *topologies. Dans le coalescent standard, toutes les lignées non encore coalescées ont equiprobabilité de fusionner à chaque événement de coalescence, si bien que toutes les topologies sont equiprobables.*

La loi du temps d'attente entre deux coalescences successives dans le coalescent standard s'obtient par une approximation en temps continu de la loi du temps d'attente dans la version discrète. Considérons deux gènes dans une population de taille $2N$ gènes, se reproduisant selon le modèle de Wright-Fisher. La probabilité pour que les lignées ancestrales de ces deux gènes coalescent à la génération précédente est $1/(2N)$, et la probabilité pour qu'elles coalescent exactement $k > 0$ générations dans le passé (après $k-1$ générations de non coalescence), est

$$\frac{1}{2N}\left(1-\frac{1}{2N}\right)^{k-1}.$$

Le nombre de générations avant coalescence des deux lignées considérées est donc une variable aléatoire qui suit une loi géométrique de paramètre $1/2N$, de moyenne $2N$. Si $2N$ est grand, cette

loi géométrique peut, après renormalisation de l'échelle des temps en unités de $2N$ générations, être approchée par une loi exponentielle (encarts pages 38 et suivantes) de moyenne 1, appelée $T(2)$. La loi approchée du temps d'attente $T(k)$ avant la coalescence de 2 parmi k lignées peut être déduite de cette loi exponentielle comme la loi du plus petit parmi C_k^2 temps d'attente tirés indépendamment selon $T(2)$. On montre que cette loi du temps $T(k)$ pendant lequel on a k lignées ancestrales (avec $0 < k \leq n$) est la loi exponentielle de paramètre $k(k-1)/2$, et la moyenne du temps d'attente entre deux coalescences successives est $2/[k(k-1)]$. Partant d'un échantillon de taille n, on devra attendre de plus en plus longtemps en espérance l'occurrence des coalescences successives. L'allure des coalescents standards à effectif constant est illustrée à la figure 2.5 page 50 (voir aussi [99]). On gardera à l'esprit que le temps d'attente $T(2)$ avant la dernière coalescence est en moyenne plus long que la somme de tous les temps d'attente précédents, et que la variance de la durée totale du coalescent est du même ordre de grandeur que cette durée, et principalement due à la variance de $T(2)$.

Pour résumer, *le coalescent standard donne la loi de la généalogie d'un échantillon de n gènes pris à un locus particulier, c'est à dire d'une part la loi discrète de la topologie de la généalogie, et d'autre part (indépendemment de la topologie) la loi continue des temps d'attente* $T(n), T(n-1) \ldots T(2)$ *entre coalescences successives, le tout préalablement à l'observation des états alléliques dans l'échantillon.* Remarquons que les coalescences multiples et les coalescences simultanées sont négligeables dans l'approximation continue, si bien que le nombre de lignées ancestrales de l'échantillon est décrémenté de 1 à des dates relatives qui sont des variables aléatoires exponentiellement distribuées, mutuellement indépendantes.

L'une des forces du n-coalescent réside dans le fait que, pour une large classe de modèles ayant en commun l'équivalence entre individus et une taille constante de la population de gènes, la structure stochastique de la généalogie d'un échantillon ne dépend pas des détails du mécanisme de reproduction [75, 76, 77]. Le modèle de Wright-Fisher n'est que l'un des modèles populationnels dont le processus de coalescence est approché par le coalescent standard. Bien des subtilités démographiques (l'existence de générations chevauchantes, de sexes séparés) semblent pouvoir être incorporées indirectement dans les approches 'par coalescence', au moyen d'un changement de l'échelle des temps dans le coalescent [99]. Le coalescent standard est en ce sens une approximation robuste du processus de coalescence discret, dans les cas neutres. L'avantage de cette robustesse est que les approches par coalescence peuvent être utilisées pour l'étude de données concernant des organismes variés, quelle que soit l'organisation de leur cycle de vie. Le revers de la médaille est que l'approche par coalescence ne donne pas d'informations sur les paramètres directement mesurables dans les populations (les paramètres naturels), mais sur des combinaisons de ces paramètres. Le retour aux paramètres naturels nécessite l'incorporation d'informations extérieures (par exemple par l'adoption d'une approche bayésienne [145, 139]).

2.1.3 Le coalescent avec variations d'effectif

Les variations d'effectif déterministes dans les populations sont l'un des phénomènes biologiques dont l'influence sur le processus de coalescence peut être exprimée par un changement d'échelle temporelle dans le coalescent standard [32]. Le processus discret peut être traité comme précédemment, à ceci près que N est fonction de la génération t considérée. Plus le nombre de parents possibles à une génération est élevé, moindre est la probabilité de coalescence de deux lignées à cette génération. L'approximation continue d'un tel processus de coalescence est un coalescent dont l'échelle temporelle est déformée, et reflète le taux instantané de coalescence.

L'approximation est valable à la limite lorsque la taille de population $N(t = 0)$ à la date d'échantillonnage tend vers $+\infty$ et pour peu que tous les rapports $N(t)/N(0)$ aient une limite finie. Dans le coalescent standard, t générations dans le passé correspondent à $t/(2N)$ unités de temps de coalescence, et τ unités de temps de coalescence correspondent à $\lfloor 2N\tau \rfloor$ générations, où $\lfloor x \rfloor$ est la partie entière de x. Quand la taille de population change [55], t générations dans le passé correspondent cette fois à $g(t)$ unités de temps de coalescence et τ unités de temps de coalescence correpondent à $\lfloor g^{-1}(\tau) \rfloor$ générations, avec g^{-1} la fonction inverse de g et

$$g(t) = \sum_{i=1}^{t} \frac{1}{2N(i)}.$$

On peut remarquer que c'est la moyenne harmonique de la taille de population qui est déterminante pour le temps de coalescence.

Un tirage selon le coalescent avec des variations d'effectif particulières peut être très simplement obtenu par application de la fonction g^{-1} aux temps de coalescence d'un tirage selon le coalescent standard. L'expression précise de $g(t)$ et de $g^{-1}(\tau)$ dépend du modèle de variations d'effectif considéré. Une représentation visuelle de coalescents avec variations d'effectif sera donnée au paragraphe 2.3.2.

Temps d'attente d'un événement : loi géométrique et loi exponentielle

Un événement e peut survenir inopinément et se répéter fortuitement. La désintégration d'un composé radioactif est l'exemple d'usage. On s'intéresse au temps d'attente, aléatoire, avant la prochaine manifestation de e.

Heuristique du problème et modélisation continue

Hypothèse 1 : On considère que le phénomène est homogène dans le temps, c'est à dire qu'il n'a pas plus de chances de se produire dans un petit intervalle de temps que dans un autre de même durée.

Hypothèse 2 : On suppose que le phénomène est sans mémoire, c'est à dire que sa probabilité d'occurrence est indépendante de ce qui s'est produit avant.

Hypothèse 3 : On impose que e soit un événement rare à l'échelle de temps considérée, *i.e.* qu'il ne se produise pas deux fois presque en même temps.

En supposant que les dates d'occurrence possibles sont dans $]0;+\infty[$, cela s'exprime :

hypothèses 1 et 2 : la probabilité d'occurrence de e dans l'intervalle $]t;t+\delta t]$ ne dépend que de δt. On note cette probabilité $p(\delta t)$.

hypothèse 3 : $p(\delta t)$ tend vers $\lambda \delta t$ lorsque δt tend vers 0, avec $\lambda > 0$.

hypothèse 2 : si $]t_1;t_2]$ et $]t_3;t_4]$ sont disjoints, les événements "e se produit dans $]t_1;t_2]$" et "e se produit dans $]t_3;t_4]$" sont indépendants.

Soit E la variable aléatoire "temps d'attente de la prochaine occurrence de e". Les hypothèses ci-dessus suffisent à résoudre la question de la loi de E. Nous allons toutefois commencer par montrer que cette loi s'obtient aussi, naturellement, comme limite de la loi obtenue pour un modèle discret.

Modélisation discrète du temps d'attente et loi géométrique

On fait une observation par unité de temps δt pour voir si e s'est produit. Les instants d'observation sont notés $1, 2, \ldots, k-1, k, \ldots, T$. Pour tout $i > 0$, la probabilité de l'événement "e se produit entre k et $k+1$" vaut $p(\delta t) = p$ (hypothèse 1), et l'événement contraire a pour probabilité $1-p$. Soit G la variable aléatoire "temps d'attente discrétisé de e". La probabilité $\mathbb{P}(G = k)$ de l'événement "e se produit pour la première fois entre $k-1$ et k" s'exprime

$$\mathbb{P}(G = k) = p(1-p)^{k-1}$$

car l'occurrence de e sur chacun des pas de temps est indépendante de l'occurrence sur les autres pas de temps, et que $k-1$ pas de temps sans observation de e ont précédé l'observation de e au $k^{\text{ième}}$ pas de temps. La loi du temps d'attente discret G est appelée loi géométrique de paramètre p.
On vérifie facilement que cette loi à valeurs entières strictement positives est une probabilité (sommation à 1 des probabilités élémentaires) et on calcule l'espérance $1/p$ de la variable aléatoire G, qui est le nombre moyens de pas de temps qu'il faut attendre jusqu'à l'observation de la première occurrence de e.

Nous sommes directement préoccupés par l'occurrence de mutations et de coalescences dans une généalogie de gènes. Le temps d'attente des mutations se conforme bien aux modèles discret et continus proposés. Le temps d'attente des coalescences est plus subtil en ce sens que le paramètre de sa loi change à chaque coalescence (en population de taille constante) voire entre deux coalescences (cas de populations d'effectif variable).

Convergence de la loi géométrique vers la loi exponentielle

On rapproche les temps d'observation, effectués à des instants multiples de la fraction $1/n$ d'unité de temps. La probabilité d'occurrence de e dans un intervalle de temps de durée $\delta t = 1/n$ vaut $p(1/n) = \lambda/n$ (hypothèse 3 et formalisation correspondante).
Soit E la variable aléatoire continue du temps d'attente de e, et F_E sa fonction de répartition, définie par $F_E(t) = \mathbb{P}(E \leq t)$. Selon le théorème des accroissements finis, si F_E est continue et dérivable de dérivée f_E, on a :

$$p\left(t < E \leq t + \frac{1}{n}\right) = F_E\left(t + \frac{1}{n}\right) - F_E(t) = \frac{1}{n} \times F'_E(\tau) = \frac{1}{n} \times f_E(\tau).$$

avec $t < \tau < t + \dfrac{1}{n}$.

D'après le paragraphe *modélisation discrète*, la loi de la variable aléatoire G_n "nombre de fractions d'unité de temps à attendre la première occurrence de e" est géométrique de paramètre $p(1/n) = p$. Supposons que e se produise pour la première fois entre les instants $(k_n - 1)/n$ et k_n/n.

Si $G_n = k_n$, alors $\dfrac{k_n - 1}{n} < E \leq \dfrac{k_n}{n}$
et $\mathbb{P}\left(\dfrac{k_n - 1}{n} < E \leq \dfrac{k_n}{n}\right) = \mathbb{P}(G_n = k_n) = \dfrac{1}{n} \times f_E(\tau_n)$ où $\dfrac{k_n - 1}{n} < \tau_n < \dfrac{k_n}{n}$.
On obtient donc $f_E(\tau_n) = n \times p\left(\dfrac{1}{n}\right)\left(1 - p\left(\dfrac{1}{n}\right)\right)^{k_n - 1}$.

Lorsque n tend vers $+\infty$, τ_n et $\dfrac{k_n}{n}$ tendent vers t , d'où $f_E(\tau_n) \to f_E(t)$.
Dans le second membre, $n \times p\left(\dfrac{1}{n}\right)$ tend vers λ et comme $n \times \log\left(1 - p\left(\dfrac{1}{n}\right)\right) \sim n\left(-p\left(\dfrac{1}{n}\right)\right)$ tend vers $-\lambda$, on obtient

$$\left(1 - p\left(\frac{1}{n}\right)\right)^{k_n - 1} = \left(1 - p\left(\frac{1}{n}\right)\right)^{n \times (k_n - 1)/n} \to e^{-\lambda t}$$

La loi de E est donc $f_E(t) = \lambda e^{-\lambda t}$, appelée loi exponentielle de paramètre λ. Son espérance est $1/\lambda$ et sa fonction de répartition s'exprime $F_E(t) = 1 - e^{-\lambda t}$.

Obtention directe de la loi exponentielle à partir du modèle continu

Reprenons la variable E "temps d'attente continu". On cherche E telle que sa fonction de répartition F_E soit continue, dérivable, et vérifie

$$F_E(t) = \mathbb{P}(E \leq t) \quad , \text{E}{>}0.$$

La fonction $t \longmapsto g(t) = 1 - F_E(t) = \mathbb{P}(E > t)$ est une fonction décroissante telle que $g(0) = 1$. $g(t+s)$ est la probabilité pour que e ne se produise ni avant t, ni entre t et $t+s$. Ces deux événements sont indépendants puisque $]0;t]$ et $]t;t+s]$ sont disjoints.

L'événement "e ne se produit pas entre t et $t+s$" a pour probabilité

$$1 - \mathbb{P}(t < E < t+s) = 1 - p(s) = g(s)$$

(d'après l'hypothèse 1 d'homogénéité).

On obtient donc, pour tous t et s positifs,

$$g(t+s) = g(t) \times g(s), \tag{2.1}$$

relation fonctionnelle qui permet de déterminer la fonction g. Comme on a supposé F_E continue dérivable, g l'est aussi. En dérivant l'égalité 2.1 par rapport à s (à t constant), on peut écrire que pour tout $t > 0$ et $s > 0$,

$$g'(t+s) = g(t) \times g'(s).$$

Lorsque $s \to 0$, g étant supposée continue, $g'(t) = g(t) \times g'(0)$. En posant $g'(0) = -\lambda > 0$ et sachant que $g(0) = 1$, il vient $g(t) = e^{-\lambda t}$. On retrouve donc la fonction de répartition de la loi de E, $F_E(t) = 1 - e^{-\lambda t}$ pout $t > 0$. Sa densité $f_E(t) = \lambda e^{-\lambda t}$ s'obtient par dérivation de F_E.

Nombre d'occurrences dans un intervalle de temps et loi de Poisson

De la loi exponentielle du temps d'attente d'un événement, on peut déduire (c'est loin d'être immédiat) la loi du nombre N d'évéments qui se produisent entre les instants 0 et T donnés :

$$p(N=k) = \frac{(\lambda T)^k}{k!} \times e^{-\lambda T}.$$

C'est la loi de Poisson de paramètre λT, notée $\mathcal{P}(\lambda T)$. Son espérance et sa variance sont toutes deux égales à λT.

Logiquement, le modèle discret doit lui aussi permettre de retrouver que la loi du nombre d'occurrences dans un intervalle de temps donné est une loi de Poisson. On part des mêmes hypothèses que précédemment, et on étudie le phénomène e sur l'intervalle de temps $[0;\tau]$ fixé. On cherche la loi du nombre N d'occurrences de e dans cet intervalle.

On discrétise le problème en découpant $[0;\tau]$ en n intervalles de durée $\delta t = \tau/n$. La probabilité pour que e apparaisse dans l'un quelconque de ces intervalles est

$$p(\delta t) = \lambda \delta t = \frac{\lambda \tau}{n}$$

Pour chaque intervalle, l'apparition de e réalise une épreuve de Bernoulli de paramètre p. Les épreuves correspondant aux différents intervalles de temps sont indépendantes (hypothèse 3), et l'événement $N=k$ est réalisé si e apparaît dans exactement k parmi les n intervalles de temps possibles. Autrement dit, N suit une loi binomiale à n répétitions d'un événement de probabilité p, notée $\mathcal{B}(n,p)$:

$$p(N=k) = C_n^k p^k (1-p)^{n-k} = \frac{1 \times \left(1-\frac{1}{n}\right) \ldots \left(1-\frac{k-1}{n}\right)}{k!} \times n^k p^k (1-p)^{n-k}.$$

L'entier k étant fixé, lorsque $n \to +\infty$, le numérateur tend vers 1 et $np \to \lambda\tau$, et comme

$$\log\left[(1-p)^{n-k}\right] \sim (n-k) \times (-p) \sim (n-k) \times \frac{-\lambda\tau}{n} \underset{n\to+\infty}{\longrightarrow} -\lambda\tau,$$

il vient

$$p(N=k) = \frac{(\lambda\tau)^k}{k!} \times e^{\lambda\tau},$$

c'est à dire que N suit une loi de Poisson $\mathcal{P}(\lambda T)$, comme espéré.

Une application directe nous dit que si μ est le nombre moyen de mutations par génération et par copie physique du gène considéré, le nombre de mutations sur une lignée de durée δt est une variable aléatoire de loi de Poisson $\mathcal{P}(\mu\delta t)$. Comme évoqué page 38, la loi du temps d'attente change entre coalescences successives, si bien que le nombre de coalescences sur un intervalle de temps donné ne peut s'exprimer simplement par une loi de Poisson.

2.2 Microsatellites et inférence démographique

Les données génétiques à interpréter en termes d'histoire démographique concernent des échantillons de population, caractérisés à un ou plusieurs marqueurs moléculaires. La mise en oeuvre de l'approche démogénétique nécessite une modélisation du processus mutationnel de ces marqueurs. Un marqueur idéal pour l'inférence démographique enregistrerait l'histoire mutationnelle sans perte d'information (c'est le cas pour les séquences qui se conforment au modèle à nombre infini de sites), son typage serait suffisamment simple pour que plusieurs loci indépendants puissent être étudiés (c'est le cas des microsatellites et des SNPs), et son processus mutationnel serait bien connu. Dans le contexte de la détection d'événements qui réduisent la diversité, un fort taux de mutation est de plus souhaitable. Les microsatellites constituent un bon compromis entre ces quatre conditions.

2.2.1 Processus et modèles mutationnels pour les microsatellites

Les microsatellites consistent en la répétition en tandem d'un court motif d'ADN (généralement de moins de 5 paires de bases), comme par exemple $(AC)_n$ ou $(TATC)_n$ où n est le nombre de répétitions du motif (ouvrage de référence [51]). Ils ont été détectés dans les génomes eucaryotes dans les années 70, puis se sont révélés truffer ces génomes [143]. Avec la généralisation de l'emploi de la PCR [109, 124], les microsatellites sont devenus des marqueurs mendéliens de choix, alliant codominance, neutralité, et facilité de typage [66]. Ils sont couramment utilisés en médecine légale, en cartographie génomique, et en biologie des populations pour des études de parenté ou de structuration des populations.

Le polymorphisme des microsatellites consiste essentiellement en des variations du nombre de répétitions du motif constitutif. C'est en tout cas la part du polymorphisme que la grande majorité des études empiriques considère (mais lire *e.g.* [38]). Les données expérimentales directes [127] suggèrent que la grande majorité des mutations sont des pertes ou des gains d'un seul motif (mutations d'amplitude 1), et que les rares mutations sauts sont le plus souvent d'amplitude 2 [26, 36, 63]. Ces observations ont été prises en compte dans la proposition de modèles mutationnels (définis dans l'encart page 44), ensuite incorporés dans les outils logiciels adaptés au traitement de données microsatellites [87].

Des études théoriques (de type exploratoire) ont permis de prédire la distribution des fréquences alléliques, d'abord pour le modèle SMM (encart page 44) [131, 150]. Conçu initialement pour modéliser les mutations des isoenzymes, ce modèle s'est révélé une bonne approximation du processus mutationnel des microsatellites, et le couple IAM/SMM encadre les modèles plus réalistes. Les études directes et populationnelles ont peu à peu conduit à complexifier le modèle SMM (par exemple sous la forme d'un modèle à deux phases, ou TPM [31]), et ont mis en évidence par exemple des biais directionnels ([120, 23, 28, 149, 157], mais [35, 9]) ou des contraintes de taille [47]. Des études théoriques [95, 43, 159] ont étudié les conséquences des simplifications des modèles. Il est apparu particulièrement important [30, 2, 53] de tenir compte de l'hétérogénéité du processus mutationnel entre loci microsatellites mise en évidence par les études directes [36, 130, 128]. Un compromis entre simplicité et pouvoir descriptif est de considérer un modèle de type TPM (ou GSM) à taux de mutation variable d'un locus à un autre [107, 39].

2.2.2 Intérêt des microsatellites pour l'inférence démographique

Neutralité et indépendance entre généalogie et mutation

Les microsatellites sont, en première approximation, considérés comme des marqueurs dont les variants sont neutres [51]. Cette neutralité permet de modéliser les généalogies de microsatellites par le coalescent de Kingman, dans lequel le taux de coalescence et le taux de mutation sont indépendants [104]. La diversité allélique dans un échantillon est la mémorisation —avec perte d'information— de la généalogie de cet échantillon, laquelle est une réalisation d'un processus stochastique qui ne dépend que de l'histoire démographique de la population.

Certains loci trinucléotides très instables sont associés à des maladies génétiques graves (maladie du X fragile, chorée de Huntington) et ont des allèles fortement délétères [51]. Ces loci à l'évidence non neutres ne sont pas utilisés comme marqueurs moléculaires. Les dinucléotides et tétranucléotides sont probablement bien plus proches de la neutralité, mais ils peuvent être soumis à des effets d'auto-stop ou de balayage sélectif, qui indirectement modifient leurs généalogies [129]. Plus un génome est compact et moins il recombine, plus les écarts à la neutralité des microsatellites risquent de se manifester.

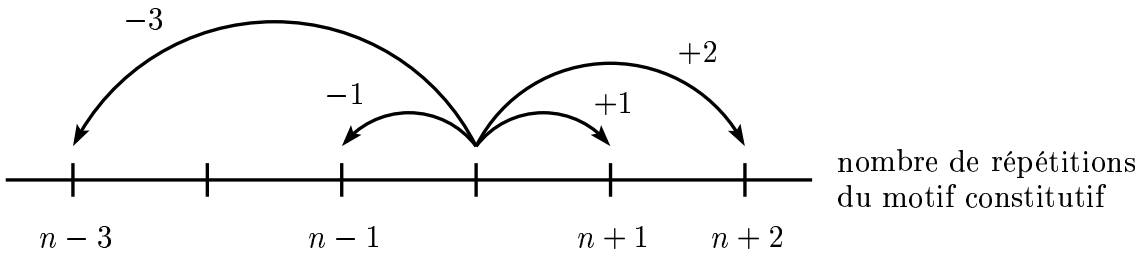
Jeux de données multilocus et qualité des inférences

Comme nous venons de le signaler, un locus marqueur peut être fortement lié à un locus soumis à sélection. Une généalogie courte pour un marqueur microsatellite supposé neutre peut ainsi être due à un effet sélectif, et non pas à l'histoire démographique, et une méthode inférentielle risque d'être abusée. La bonne répartition des microsatellites le long des différents chromosomes d'un génome, et leur facilité de typage, permettent de travailler sur des jeux de données de plusieurs loci indépendants. L'effort de typage est alors réparti entre plusieurs loci dont la probabilité individuelle de présence dans la zone d'auto-stop d'un locus récemment soumis à un effet sélectif est faible [46]. Idéalement, les méthodes inférentielles développées doivent être robustes à la contamination d'un jeu de données par une petite proportion de loci sélectionnés. Même en l'absence de loci qui dévient des hypothèses de travail, le fait de disposer de plusieurs loci indépendants est plus informatif sur les processus démographique et mutationnel. En effet, même pour une histoire démographique très contraignante, la diversité des généalogies et des histoires mutationnelles possibles reste considérable [99], et plusieurs réalisations du processus stochastique à l'origine des données permettent de se faire une idée bien plus précise de ce processus stochastique. Une analogie avec un processus stochastique élémentaire permettra de s'en convaincre (encart page 45).

La majorité des méthodes d'inférence démographique disponibles traitent les données d'une unique séquence ou d'un jeu de microsatellites liés [154, 155]. Dans les deux cas, il y a risque de sélection directe ou indirecte (en particulier pour des séquences à transmission monoparentale, non recombinantes), et on a un problème d'inférence à partir d'une seule réalisation d'un processus stochastique complexe. La méthode publiée en 1999 par M. Beaumont, implémentée dans les versions successives du programme MSVAR, traite de données à des microsatellites non liés, et évite donc ces deux écueils (nous verrons à l'usage qu'elle pose des problèmes plus terre à terre).

Modèles mutationnels pour les microsatellites

Les microsatellites sont des répétitions en tandem d'un court motif d'ADN, présents dans les génomes eucaryotes. Les microsatellites purs sont entièrement caractérisés par le nombre n de répétitions du motif, par exemple $(AC)_n$. Ils sont les plus utilisés en cartographie du génome. Des microsatellites interrompus comme $(AC)_nC(AC)_m$ ou composés comme $(AC)_n(GT)_m$ sont également fréquents dans les génomes, et sont utilisés comme marqueurs, en mélange avec des microsatellites purs dans beaucoup d'études populationnelles.
Les modèles mutationnels ignorent généralement ces subtilités, et considèrent que le polymorphisme des microsatellites est réductible à un nombre de répétitions du motif de base, directement déductible de l'observation de la longueur de produits PCR incluant les marqueurs. Les mutations d'un microsatellite peuvent alors être modélisées comme des déplacements sur l'axe des entiers naturels, la position sur cet axe représentant le nombre de répétitions du motif. On appelle amplitude d'une mutation le nombre d'unités de répétitions perdues ou ajoutées au marqueur, et on précise le sens de déplacement par l'ajout d'un signe + ou −.



Dans ce contexte, un modèle mutationnel doit spécifier la probabilité μ de mutation du marqueur (par copie physique du marqueur et par génération, *cf.* encarts page 38 et suivantes), et la loi de l'amplitude des mutations. Les principaux modèles rencontrés dans la littérature sont décrits ci-dessous [37]. Leur réalisme et leur facilité de traitement sont rapidement discutés dans le texte principal :

- Le modèle IAM (Infinite Allele Model) considère que toute mutation atterrit à une position jusque là inoccupée sur l'axe des entiers, indépendante de la position du gène parent.
- Le modèle KAM (K Allele Model) restreint le modèle IAM à un nombre fini K d'allèles, atteints uniformément et indépendamment de l'allèle parent.
- Le modèle SMM (Stepwise Mutation Model) donne équiprobabilité 1/2 aux mutations +1 et −1.
- Le modèle AMM (Admixture Mutation Model) considère une probabilité p de mutations d'amplitude 1 et une probabilité $1-p$ de mutations d'amplitude $a > 1$, a fixé []. Il est donc spécifié par deux paramètres.
- Le modèle GSM (Generalised Stepwise Model) donne à l'amplitude des mutations une distribution arbitraire, le plus souvent une loi géométrique, spécifiée par un seul paramètre (*e.g.* sa variance ou sa moyenne).
- Le modèle TPM (Two-Phase Model) donne une probabilité p de suivre le SMM et une probabilité $1-p$ de suivre le GSM de type géométrique.

Des subtilités (biologiquement fondées) peuvent être incorporées à ces modèles. Par exemple, les modèles dissymétriques affectent une probabilité différente aux mutations de même amplitude mais de directions opposées. Les modèles avec taille-dépendance donnent une loi de l'amplitude qui varie avec le nombre de répétitions du gène parent.

Inférences sur un processus stochastique

On souhaite inférer le nombre n de bus desservant une ville, d'après la liste des numéros des bus que l'on voit passer. Les bus sont numerotés de 1 à n, et on suppose que lorsqu'un bus passe, il peut s'agir de n'importe lequel de façon équiprobable.

Si la liste des bus observés ne comporte qu'un numéro (on n'a croisé qu'un bus), mettons 37, on ne pourra que faire l'inférence suivante : au moins 37 bus desservent la ville. Si la liste des données s'allonge, le plus grand des nombres obtenus donne la borne inférieure du nombre de bus, lequel se rapproche peu à peu de n. Par exemple, si les numéros successivement observés sont $\{37, 12, 4, 39, 4, 12, 45, 1, 23, 32, 44, 51\}$, on pourra dire à l'issue des 4 premiers tirages qu'il y a au moins 39 bus, à l'issue du $7^{\text{ième}}$, qu'il y en a au moins 45 bus, et on révisera notre position à l'issue du $12^{\text{ième}}$ tirage, pour dire que la ville est desservie par au moins 51 bus.

Quand on commence à avoir une distribution qui ressemble à quelque chose d'uniforme sur $1 \ldots n$, avec un grand nombre d'observations de chaque bus, on est convaincu de connaître le véritable nombre n de bus (et on a sans doute raison).

Moralité : plus on observe de réalisations d'un processus stochastique, mieux il est cerné. Le processus stochastique qui nous intéresse présentement, et qui génère les données de diversité génétique, est autrement plus complexe, avec sa superposition de topologies, de dates de coalescence, de nombres, positions et types de mutations. L'information que l'on a sur lui est très indirecte, et ne croît que très lentement lorsque la taille des jeux de données augmente. Le temps de calcul, lui, a la mauvaise idée d'exploser avec la quantité de données.

2.3 Variations d'effectif : modélisation et signatures

Lorsque les fluctuations d'effectif dans une population sont régulières et de faible ampleur, le modèle de Wright-Fisher peut rester adéquat pour décrire la diversité génétique. L'effectif pertinent est alors l'effectif efficace, c'est à dire l'effectif que devrait avoir une population idéale théorique pour que la dérive génétique y soit de même intensité que dans la population d'intérêt. Différentes mesures de l'intensité de la dérive coexistent [156, 73, 61], plus ou moins adaptées pour quantifier différents types d'écarts à la situation idéale (variations d'effectif, sex-ratio, chevauchement entre générations). Lorsque les variations d'effectif correspondent à des événements démographiques ponctuels de forte ampleur, on est toutefois amené à les modéliser explicitement de façon schématique. Dans la pratique, on incorpore les fluctuations d'effectif et les caractéristiques du système de reproduction sous la forme d'un effectif efficace instantané, pour se concentrer sur les variations d'effectif paramétrées dans le modèle. Dans la partie II, tous les effectifs seront des effectifs efficaces instantanés.

2.3.1 Modèles démographiques avec variations d'effectif

Nous ne présenterons ici que les modèles les plus couramment utilisés en génétique des populations, pour traduire les variations d'effectif dans une population unique, close et panmictique.

Nous déciderons à cette occasion d'une terminologie parfois en désaccord avec celle de la littérature, mais qui évitera toute confusion. Nous décrirons les variations d'effectif en prenant $t = 0$ pour la date la plus récente (typiquement la date de l'échantillonnage d'une population) et $t < 0$ pour les dates antérieures. Autrement dit, le sens naturel de l'écoulement du temps sera conservé.

Une première façon de modéliser des variations d'effectif —qui a l'avantage de la simplicité— est de supposer que la taille de la population est constante par morceaux (figure 2.3). On obtient ainsi une famille de modèles dont les plus simples sont l'*expansion en marche d'escalier* [70, 74, 114] ou le *déclin en marche d'escalier* [89], pour lesquels on n'a qu'une discontinuité au temps m dans le passé. Dans les deux cas, $N(t) = N_1$ pour t dans $]-\infty; -m[$ et $N(t) = N_0$ pour t dans $[-m; 0]$. L'ordre de N_0 et N_1 détermine si l'on a affaire à une expansion ($N_1 < N_0$) ou à un déclin ($N_1 > N_0$). Remarquons que le déclin en marche d'escalier correspond à ce que l'on désigne usuellement sous le terme de *bottleneck* [85]. Toute la famille peut être générée par combinaison de marches d'escalier, montantes ou descendantes, à des dates données. Le *bottleneck transitoire* [114] est ainsi obtenu par la concaténation d'un déclin et d'une expansion, à des dates respectives m et n, avec $m < n < 0$. Nous éviterons soigneusement d'utiliser le terme de bottleneck pour désigner un bottleneck transitoire (cette acception est malheureusement présente dans la littérature [88, 41]). Notons que le déclin en marche d'escalier n'est pas toujours modélisé explicitement, mais parfois en supposant une variabilité faible ou nulle au début de l'intervalle de temps considéré [71, 88].

Les variations d'effectif sur de courtes durées sont classiquement modélisées comme exponentielles [117]. La variation décrite peut être un déclin ou une expansion (figure 2.4). Dans la pratique, la phase exponentielle est tronquée en considérant une situation ancestrale stable, et un début de variation à une date t_v dans le passé [6, 70, 74, 152], par exemple selon $N(t) = N_0 r^{-t/t_v}$. Si l'effectif ancestral stable N_2 est différent de l'effectif N_1 au début de la phase exponentielle, on peut décrire une variété de courbes démographiques, en fonction des valeurs relatives de N_0, N_1 et N_2. En particulier, si $N_2 > N_1 < N_0$, on a affaire à une *fondation-explosion*, ou *crash-flush* tel que modélisé dans la partie II [71].

Il a parfois été considéré des modèles logistiques de croissance de population [96], dans lesquels le taux de croissance dépend de l'effectif. Ces modèles sont plus réalistes que les précédents pour des questions d'inférence démographique, en ce sens qu'une croissance exponentielle de population ne peut être soutenue sur le long terme. Malheureusement, la loi des temps de coalescence pour un modèle logistique ne peut être calculée analytiquement (on ne peut pas inverser $g(t) = \sum_{i=1}^{t} 1 \,/\, 2N(i)$ dans le cas logistique, voir page 37). Pour les approches par coalescence, on est donc amené à approcher la variation logistique par une exponentielle tronquée. Plus généralement, n'importe quelle forme de variation d'effectif peut être remplacée par des morceaux exponentiels, linéaires ou par des marches d'escalier. Chaque jonction pose toutefois des problèmes analytiques, et on doit chercher à saisir l'essence des variations d'effectif avec un nombre minimal de morceaux continus, dérivables et —pour les applications utilisant le coalescent— intégrables.

Pour conclure, la modélisation des variations d'effectif et l'échelle de temps considérée doivent être adaptées à la question posée. Par exemple, si l'on s'intéresse aux effets d'échantillonnage lors d'une fondation, on peut choisir un modèle de déclin en marche d'escalier, et ignorer la croissance ultérieure et les mutations qui se produisent pendant cette phase [89]. Si l'on s'intéresse à la façon dont la diversité retrouve son niveau d'équilibre à la suite d'un événement réducteur de diversité, il est indispensable de modéliser la phase de croissance, alors que la perte de diversité

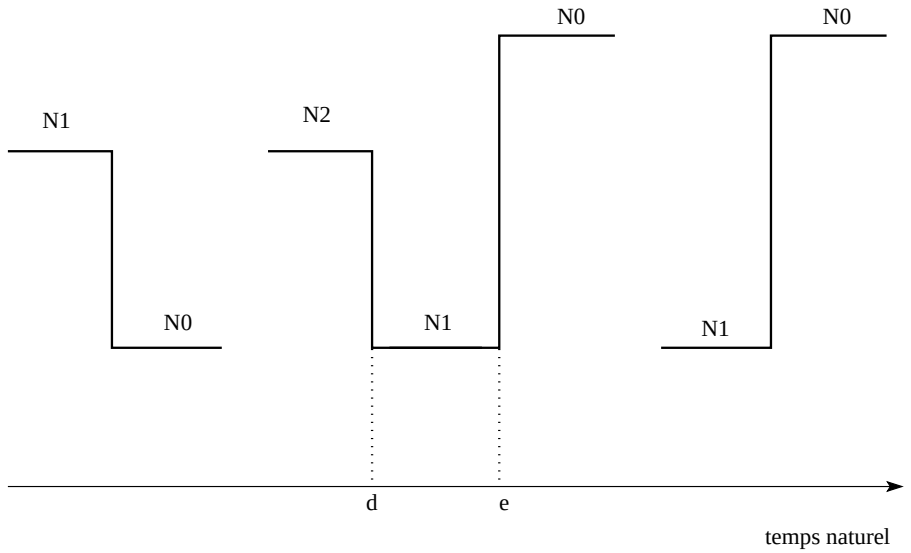


FIG. *2.3: Modèles de variations d'effectif constants par morceaux. Les plus simples sont le déclin en marche d'escalier (à gauche) et l'expansion en marche d'escalier (à droite) dont la concaténation dans cet ordre donne le goulot démographique transitoire (au centre). Ces modèles sont spécifiés par les dates des événements et par les effectifs avant et après (N_0, N_1, N_2). Le bottleneck transitoire est ainsi caractérisé par* 5 *paramètres. On peut en abandonner un premier en supposant que la population revient à son effectif initial après le goulot transitoire ($N_0 = N_2$), et un second en remplaçant N_0/N_1 et $(e-d)$ par un unique paramètre de sévérité , pour peu que la mutation puisse être ignorée pendant le bottleneck transitoire [46]. Dans le cas de marqueurs microsatellites, on ne peut ignorer la mutation que sur quelques générations [13].*

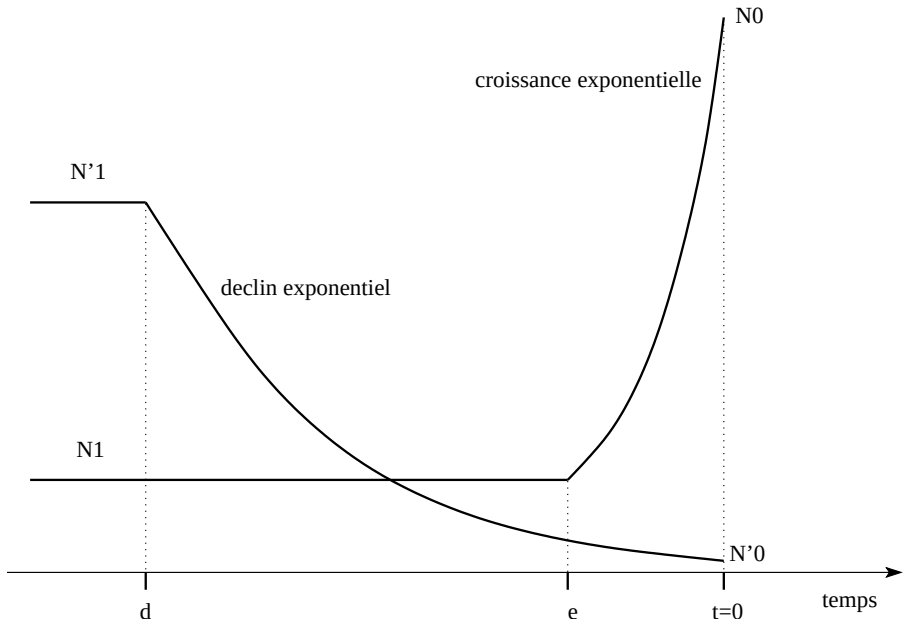


FIG. *2.4: Modèles de variation d'effectif exponentielle. La période de croissance ou de décroissance exponentielle est généralement tronquée à une date passée, et raccordée à une phase d'effectif stable. Il faut alors* 3 *paramètres pour spécifier le modèle, l'effectif de la population ancestrale (N_1 resp. N_1'), la date du début de variation exponentielle (d resp. e), et l'effectif à $t = 0$ (N_0 resp. N_0').*

peut être supposée sans modélisation explicite [88]. Dans tous les cas, les modèles cherchent à saisir l'essence de complexes variations d'effectif dans les populations naturelles [83, 90]. Ils peuvent être utilisés pour étudier les conséquences génétiques des variations d'effectif, et pour construire des méthodes de détection et d'inférence à partir de données empiriques.

2.3.2 Signatures génétiques des variations d'effectif

Les différences entre deux copies d'un marqueur moléculaire qui divergent depuis une date T résultent de l'accumulation de mutations sur les deux branches de longueur T qui les séparent de la copie ancestrale. Si la mutation est un événement rare à l'échelle T, le nombre de mutations accumulées est distribué selon une loi de Poisson de moyenne $2\mu T$ où μ est le taux de mutation instantané. La moyenne du nombre de mutations qui se sont produites est donc proportionnelle à la date de divergence. Si l'on considère non pas deux copies, mais un échantillon de n copies d'un marqueur, on peut généraliser ce résultat en considérant la généalogie de l'échantillon : les n copies du marqueur sont reliées par une généalogie, caractérisée par sa topologie datée. Les effectifs des différentes classes alléliques dans l'échantillon dépendent de la topologie datée par l'intermédiaire des mutations qui s'accumulent en nombre d'espérance proportionnelle à la longueur des branches.

L'étude des signatures génétiques des variations d'effectif se fera donc au travers de celle des variations de forme et de longueur des généalogies. Deux questions indépendantes dans le cas de marqueurs neutres sont de savoir d'une part comment les variations d'effectif influent sur la forme et la longueur des généalogies et d'autre part comment la forme et la longueur des généalogies influe sur la diversité allélique.

Variations d'effectif et forme des généalogies

On illustre la forme des généalogies en population stable, ce qui correspond à la référence du coalescent standard. On compare à cette référence les généalogies obtenues pour des scénarios à la fois classiques et utilisés dans la partie II : une croissance exponentielle de la population —dans le sens naturel de l'écoulement du temps— et pour une histoire démographique de fondation-explosion.

Dans une population d'*effectif stable*, la densité de probabilité de coalescence pour une paire de lignées est une constante. À chaque coalescence, le nombre de lignées ancestrales de l'échantillon est décrémenté, et avec lui le nombre de paires de lignées pouvant coalescer. Pour un échantillon de taille 20 par exemple, extrait d'une population de taille constante N_0 génomes haploïdes, la loi du temps d'attente de la première coalescence est exponentielle de moyenne $1/C_{20}^2$, en unité de N_0 générations. La loi du temps d'attente de la seconde coalescence est exponentielle de moyenne légèrement supérieure $1/C_{19}^2$, et ainsi de suite jusqu'à la loi du temps d'attente de la coalescence des deux dernières lignées, exponentielle de moyenne $1/C_2^2 = 1$. La figure 2.5 schématise trois réalisations du coalescent pour un échantillon de taille 20 en population stable (méthode de simulation décrite au paragraphe 5.1). La moyenne et la variance théoriques de la date de la dernière coalescence sont respectivement $E(t_{MRCA}) = 2(1 - 1/n)$ soit 1.9 pour un échantillon de taille $n = 20$, et $V(t_{MRCA}) = 1.16$.

Dans une population en *croissance exponentielle*, le taux de coalescence entre paires de gènes décroit exponentiellement dans le sens naturel de l'écoulement du temps. Par rapport à une population de taille constante N_0 gènes, une population de même taille à $t = 0$, mais en croissance exponentielle, a des temps de coalescence plus courts entre paires de gènes. Les

généalogies d'échantillons s'en trouvent distordues (figure 2.7), et au lieu d'être de plus en plus longs, les temps d'attente entre coalescences successives en remontant dans le passé sont de plus en plus courts. La date du MRCA est fortement réduite par rapport au cas stable, et sa variance l'est plus encore (figure 2.7). À l'extrême, on peut obtenir des généalogies en étoile, dans lesquelles toutes les coalescences sont concentrées dans un intervalle de temps négligeable au regard du temps de coalescence global pour l'échantillon.

Lorsqu'une population connait un *déclin en marche d'escalier*, le taux de coalescence est diminué brutalement à la date du déclin (en remontant dans le passé). De ce fait, les lignées n'ayant pas coalescé à cette date peuvent coalescer à une date très ancienne (figure 2.9).

Enfin, lorsqu'un déclin et une explosion démographique sont concaténés, les signatures des deux événements se superposent, et les généalogies sont de type exponentiel jusqu'à une certaine date dans le passé, date au delà de laquelle les dernières lignées peuvent survivre fort longtemps avant de coalescer (figure 2.9). Dans ces deux derniers cas, la durée totale des généalogies correspond principalement à la durée de leur partie antérieure au déclin (et de même pour la variance de t_{MRCA}).

Forme des généalogies et distribution de fréquences alléliques

La forme des généalogies détermine de façon probabiliste le nombre et la position des mutations. On a bien sûr en espérance moins de mutations sur une généalogie courte que sur une généalogie longue, et donc une moindre diversité allélique dans le premier cas. À somme des longueurs de branches égale, on s'attend à un même nombre de mutations quelle que soit la longueur relative des branches. Mais le nombre de gènes de l'échantillon touchés par ces mutations dépend de la profondeur des branches affectées : les mutations sur les branches terminales des généalogies ne concernent qu'un gène, alors que les mutations sur les branches profondes affectent tout un sous-arbre descendant. En *population stable*, les branches les plus longues sont les deux plus profondes, et on s'attend donc à ce que beaucoup de mutations les affectent. Ces mutations donnent des haplotypes fréquents dans l'échantillon. En *croissance exponentielle*, les branches les plus longues —les plus touchées par les mutations— sont au contraire les branches peu profondes, dont les sous-arbres descendants sont petits, si bien que beaucoup de mutations ne concernent qu'une petite part de l'échantillon. Les *fondation-explosion* donnent des situations hybrides, dans lesquelles les branches profondes sont longues, à condition que le MRCA date d'avant la fondation. Dans le cas de marqueurs microsatellites mutant selon un modèle SMM, ces différences se traduisent par des changements du nombre et de l'espacement des modes de distribution des fréquences alléliques (figures 2.6, 2.8 et 2.10). Les distributions sont le plus souvent bimodales ou trimodales en population stable, et tous les allèles intermédiaires sont en général présents. Lorsque les généalogies dépendent d'une phase de croissance exponentielle, les distributions sont plus fréquemment monomodales. Enfin, suite à une fondation-explosion, et lorsque les dernières coalescences sont plus anciennes que la fondation, on peut obtenir une distribution avec deux modes distants, séparés par des classes non représentées.

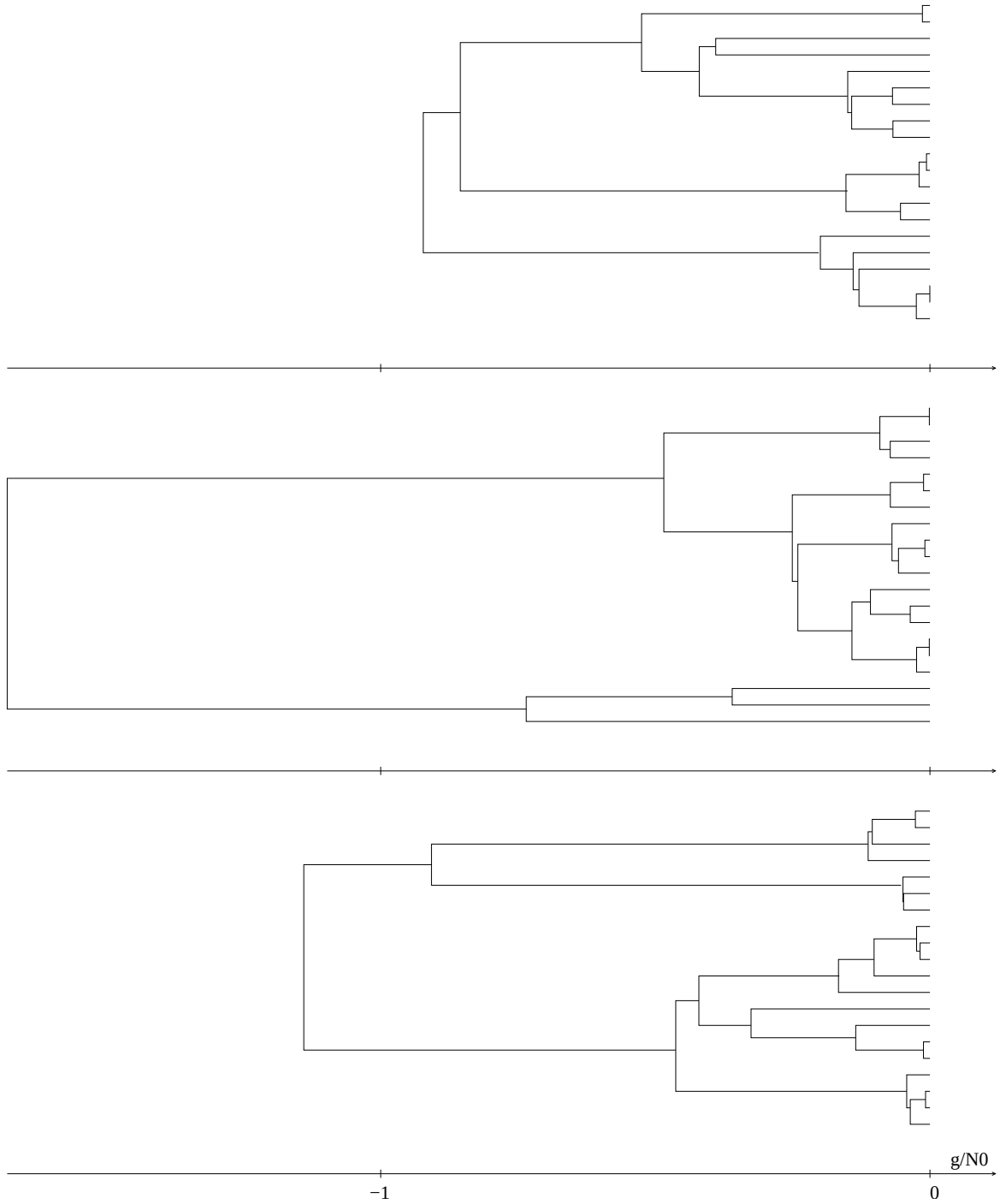


FIG. *2.5: Généalogies de gènes en population stable. Le temps est mesuré en unités de N_0 générations (g/N_0), où N_0 est l'effectif de la population de gènes au temps $t = 0$. Trois réalisations indépendantes du coalescent standard pour un échantillon de taille* 20 *gènes sont représentées. Le temps moyen de coalescence de l'échantillon est $E(t_{MRCA}) = 1.899$ et la variance de ce temps est $V(t_{MRCA}) = 1.156$ (valeurs empiriques calculées sur* 10000 *coalescents, proches des valeurs théoriquesde* 1.9 *et* 1.16*). En population stable, l'écart-type de la date du MRCA vaut plus de la moitié de sa moyenne. La forte variance sur t_{MRCA} résulte principalement de celle du temps d'attente des coalescences les plus profondes (*i.e. *les plus proches du MRCA).*

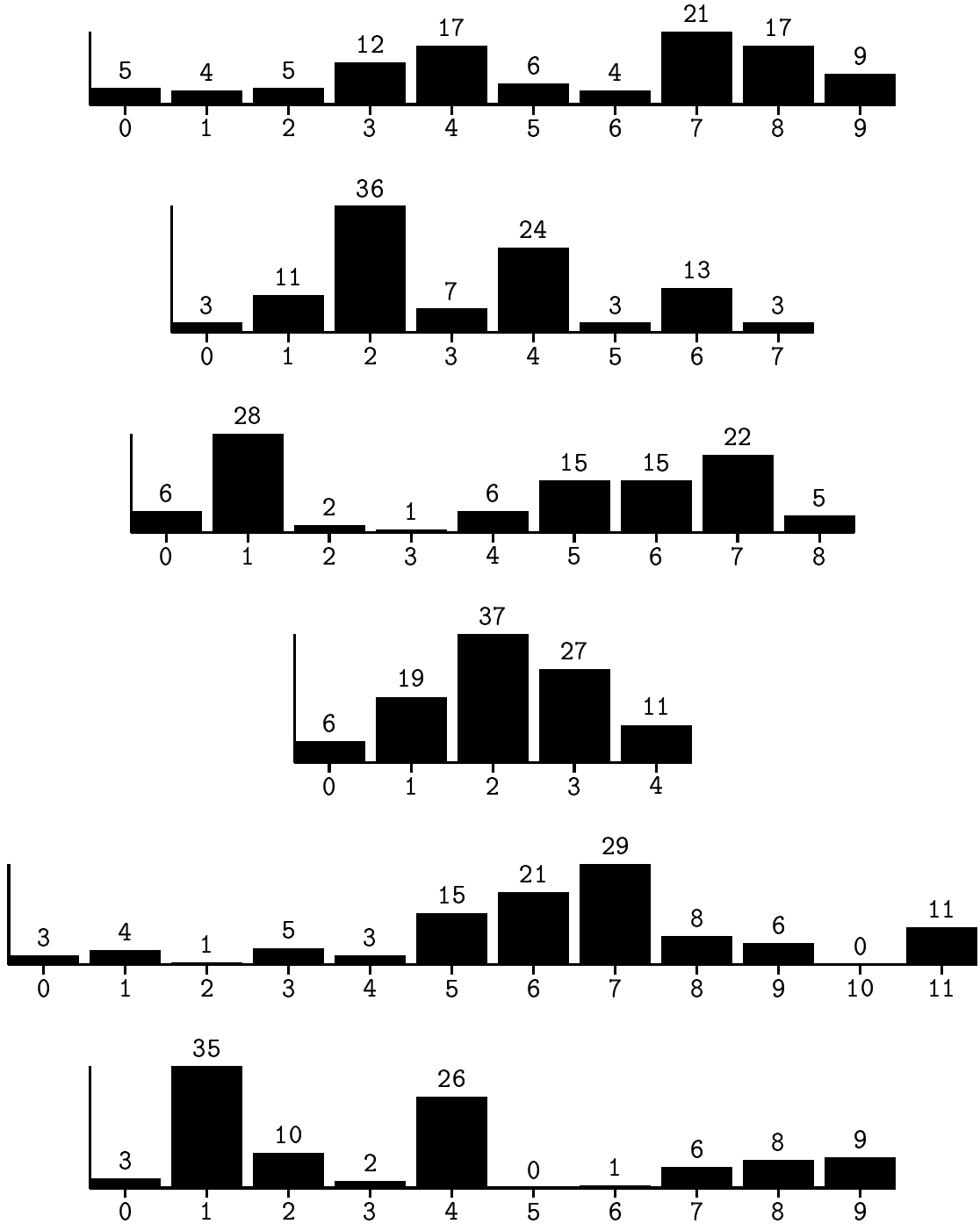


FIG. *2.6: Distribution de comptages alléliques dans des échantillons de taille* 100 *simulés en supposant un effectif stable, à des marqueurs microsatellites suivant un processus mutationnel SMM de taux* $\theta = 2N_0\mu = 10$. *Pour cette valeur de* θ, *on a typiquement une dizaine d'allèles avec un, deux ou trois modes de fréquence.*

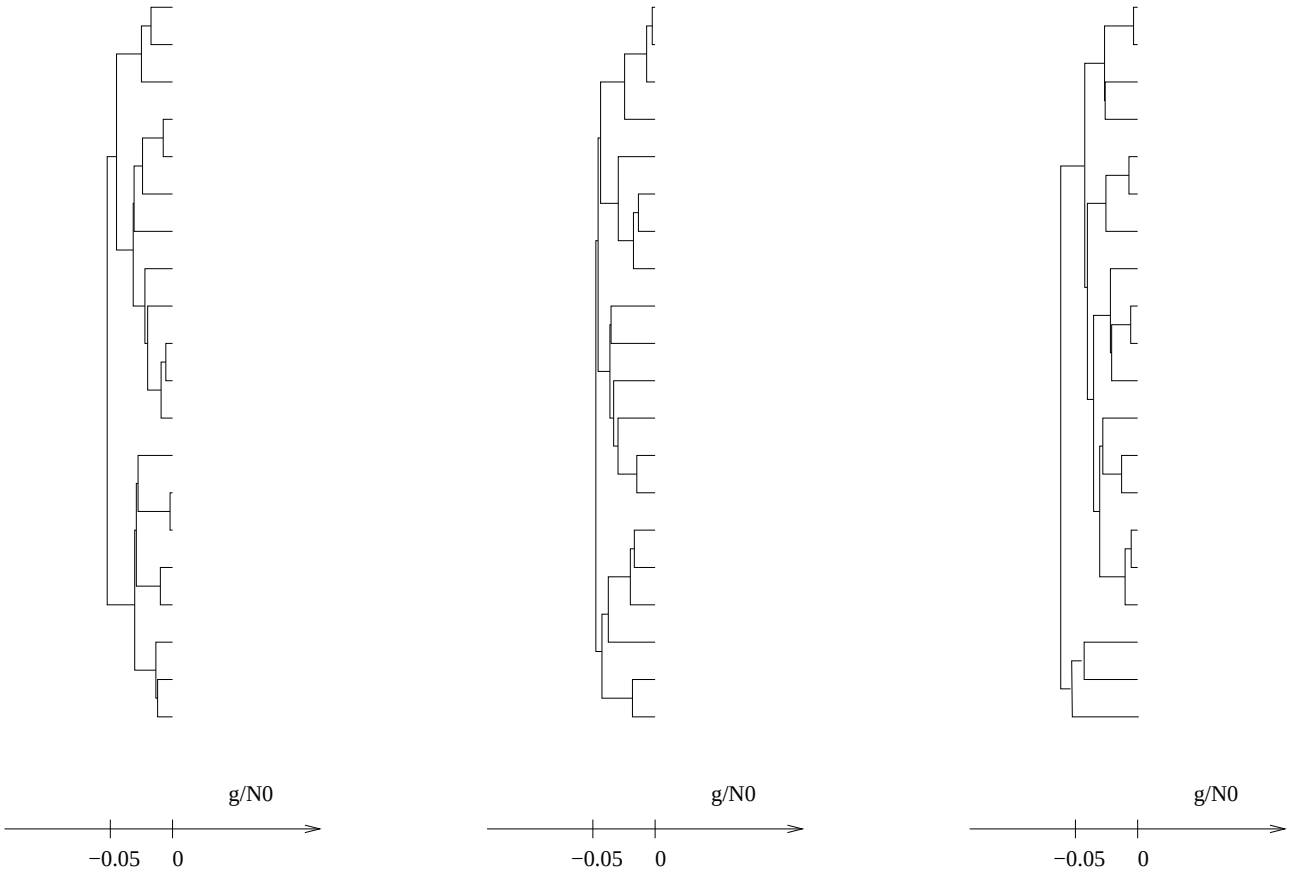


FIG. *2.7: Généalogies de gènes dans une population en croissance exponentielle. Le temps est mesuré en unité de N_0 générations, où N_0 est l'effectif de la population de gènes au temps $t = 0$. Trois réalisations indépendantes du coalescent pour un échantillon de taille $n = 20$ sont représentées pour un profil démographique tel que la taille de population soit constante et 100 fois inférieure à N_0 pour $t < -0.05$, et varie exponentiellement entre $t = -0.05$ et $t = 0$ pour atteindre N_0 à cet instant. Les généalogies sont courtes, mais structurées (il ne s'agit pas de généalogies en étoile) : bien que rapide, la croissance exponentielle s'accompagne d'une certaine dérive génétique. Dans les deux cas, les rares lignées qui survivent au delà de $t = -0.05$ coalescent rapidement dans la population ancestrale de faible effectif. Le temps moyen de coalescence $E(t_{MRCA}) = 0.058$ de l'échantillon est considérablement plus court qu'en population stable de même effectif N_0 et la variance est réduite à $V(t_{MRCA}) = 0.00011$ (valeur empiriques calculées sur 10000 coalescents).*

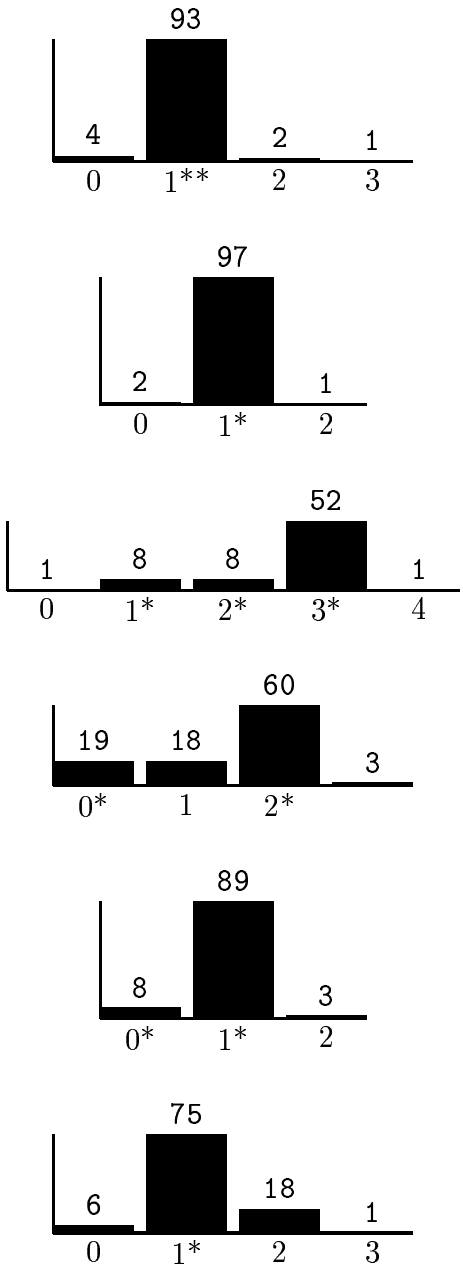


FIG. *2.8: Distributions de comptages alléliques dans des échantillons de taille* 100 *simulés en supposant une croissance exponentielle avec* $r = a = 100$*, depuis une date* $t_f = -0.05$*, à des marqueurs microsatellites suivant un processus mutationnel SMM de taux* $\theta = 2N_0\mu = 10$*. Pour cette valeur de* θ*, on a typiquement* 3 *ou* 4 *allèles de tailles consécutives. Les signes * indiquent la classe allélique des lignées non coalescées à* $t_f = -0.05$*. On constate que la dérive peut distordre assez fortement la configuration à* t_f *: les allèles fondateurs ne correspondent pas nécessairement aux allèles fréquents dans l'échantillon.*

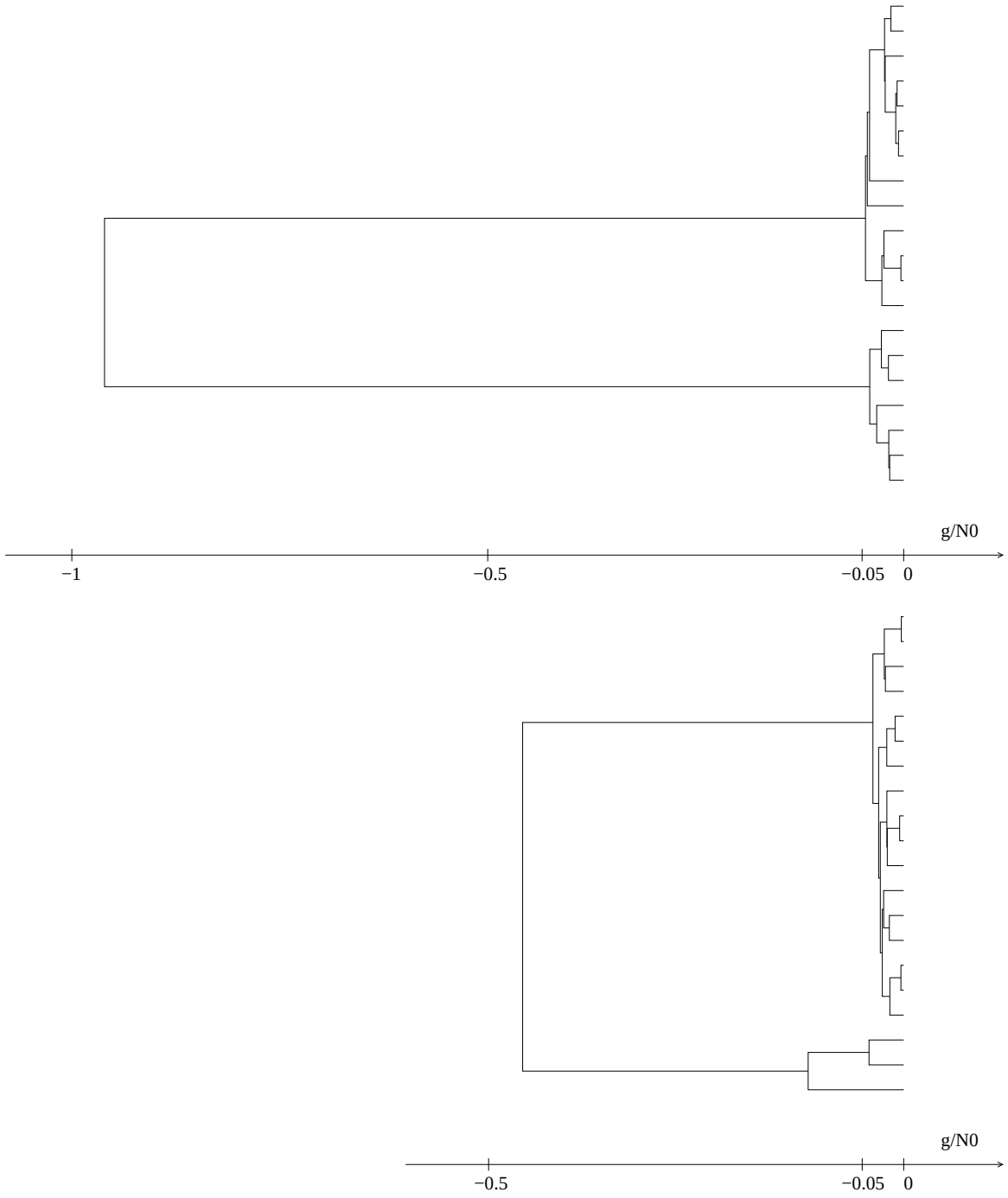


FIG. *2.9: Généalogies de gènes dans une population ayant connu une fondation-explosion. Le temps est mesuré en unité de N_0 générations comme précédemment. Deux réalisations indépendantes du coalescent sont représentées pour un profil démographique tel que la taille de population soit constante et égale à N_0 pour $t < -0.05$. À cette date, la taille de population est instantanément divisée par 100, puis elle varie exponentiellement à partir de $t = -0.05$ pour atteindre N_0 à $t = 0$. Les lignées qui survivent au delà de $t = -0.05$ mettent un certain temps à coalescer dans la population ancestrale d'effectif important, si bien que le temps moyen de coalescence de l'échantillon est élevé $E(t_{MRCA}) = 0.99$ ainsi que l'écart type $\sigma(t_{MRCA}) = 1.09$ (valeurs empiriques calculées sur 10000 coalescents).*

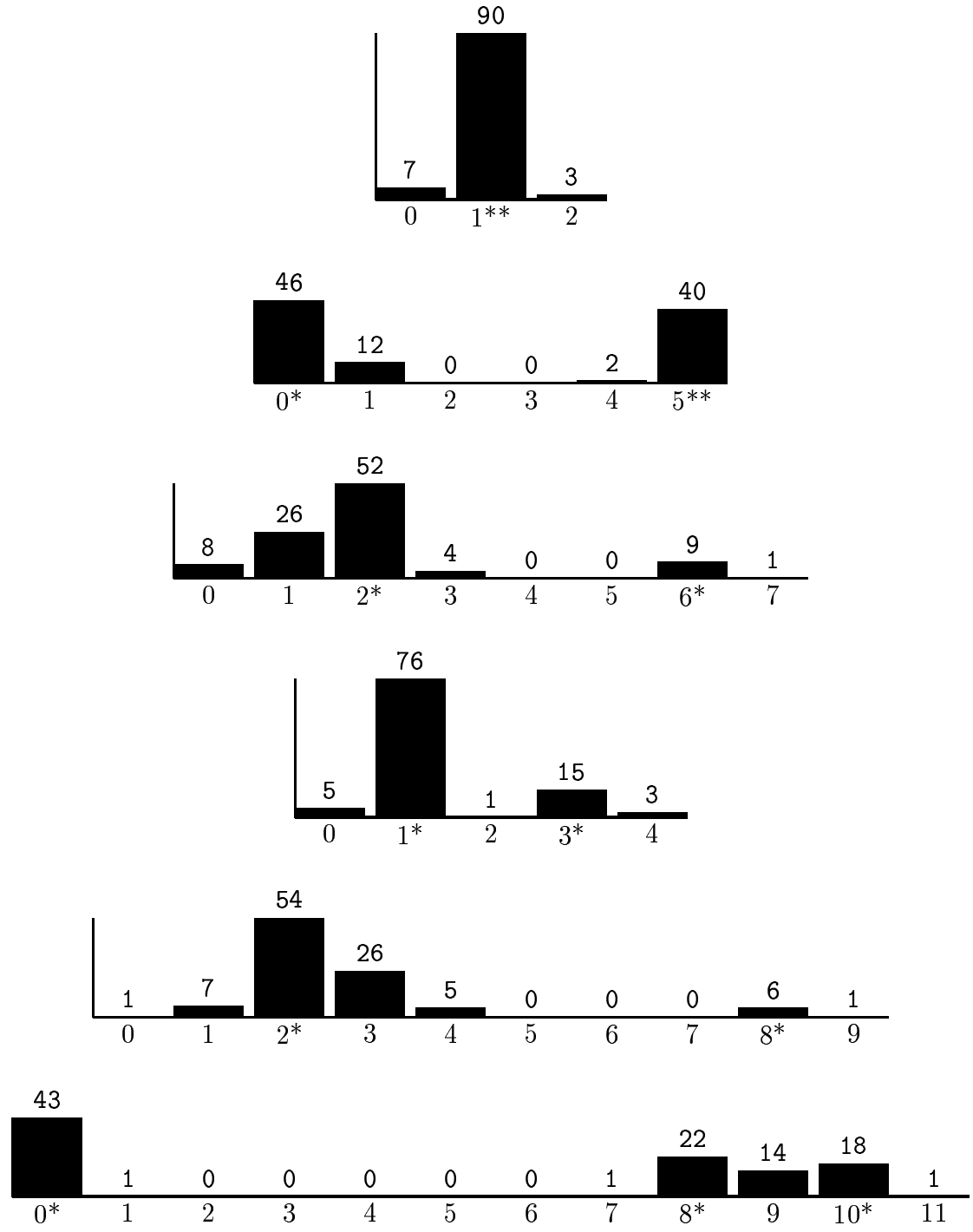


FIG. *2.10: Distribution de comptages alléliques dans des échantillons de taille* 100 *simulés en supposant une histoire de fondation-explosion avec* $r = 100$ *et* $a = 1$ *à une date* $t_f = -0.05$*, à des marqueurs microsatellites suivant un processus mutationnel SMM de taux* $\theta = 2N_0\mu = 10$*. Pour cette valeur de* θ*, on a typiquement* 5 *ou* 6 *allèles de tailles parfois fortement disjointes. Les signes * indiquent la classe allélique des lignées non coalescées à* $t_f = -0.05$ *—c'est à dire des gènes fondateurs. Comme pour le scénario exponentiel, du fait de la dérive, les classes alléliques des gènes fondateurs ne correspondent pas toujours aux classes les plus féquentes dans l'échantillon.*

Pour un jeu de valeurs des paramètres démographiques et mutationnels, on peut obtenir des configurations de types très différents, par la superposition de plusieurs processus aléatoires modélisés. La topologie de la généalogie est aléatoire, et peut être plus ou moins déséquilibrée : pour les généalogies neutres, toutes les répartitions de l'échantillon en sous-arbres sont équiprobables. Les dates relatives des coalescences sont des variables aléatoires dont la somme a un écart-type du même ordre de grandeur que sa valeur. Le nombre, la position et le type des mutations, qui déterminent *in fine* la configuration allélique de l'échantillon ajoutent aux aléas, même s'ils dépendent de la durée des différentes branches de la généalogie. La détection des variations d'effectif repose sur la distorsion des distributions de fréquences alléliques par rapport au cas stable, avec des méthodologies et des philosophies illustrées au chapitre suivant.

Chapitre 3

Méthodologie pour la détection des variations d'effectif

Les signatures génétiques des variations d'effectif peuvent potentiellement permettre d'inférer ces variations, d'après des données de typage dans des populations. Cette perspective est rendue particulièrement intéressante par les progrès dans l'acquisition de données populationnelles (notamment grâce à l'utilisation de la méthode d'amplification de l'ADN par PCR). Mais extraire l'information démographique des données de typage est techniquement délicat, du fait que la diversité génétique dans un échantillon est modélisée comme résultant d'une histoire généalogique et mutationnelle de nature stochastique. Les données empiriques dont l'information démographique peut être extraite concernent des marqueurs dont le processus mutationnel est modélisable, c'est à dire principalement des données de séquence, ou des données de typage à des marqueurs microsatellites. On ignorera dans la suite l'information génotypique et haplotypique, pour ne considérer que l'information sur les comptages alléliques, c'est à dire l'effectif de chaque classe allélique dans l'échantillon.

Après avoir donné le principe des méthodes statistiques classiques, nous montrerons que les méthodes probabilistes constituent une alternative généralement coûteuse en temps (de calcul et de développement), mais bien plus puissante.

3.1 Méthodes statistiques basées sur le calcul de moments

L'information de comptage allélique peut être résumée par le calcul de statistiques sommaires. Des statistiques communément choisies sont le nombre de classes alléliques dans l'échantillon et l'homozygotie calculée (*i.e.* la somme des carrés des fréquences alléliques). Pour des données de séquences on utilise aussi le nombre moyen de différences entre paires de séquences [140] ou encore le nombre de sites variables dans l'échantillon de séquences [151]. Pour des données microsatellites, on peut caractériser la distribution de taille allélique par des moments de cette distribution (variance, skewness, kurtosis, dans *e.g.* [52, 114]). Prises isolément, ces statistiques laissent de côté une grande partie de l'information contenue dans les données brutes. Par exemple, la statistique du nombre d'allèles peut prendre la même valeur pour des configurations pourtant très différentes de l'échantillon : dans un échantillon microsatellite comportant 4 classes alléliques, les 4 classes peuvent avoir des effectifs comparables ou déséquilibrés, les nombres de répétitions peuvent être consécutifs ou dispersés. Plusieurs statistiques choisies pour saisir des caractéristiques différentes du jeu de données permettent de mieux le caractériser [24, 114].

La façon *a priori* la plus simple de détecter des variations d'effectif consisterait à utiliser comme statistique une mesure de la diversité allélique (nombre d'allèles, hétérozygotie calculée), puisque cette diversité est influencée par des variations d'effectif (*cf.* figures 2.6, 2.8 et 2.10). Malheureusement, la valeur de statistiques exprimant la diversité n'est exploitable que par comparaison avec les valeurs observées pour les mêmes marqueurs dans la même espèce, à effectif stable, ce qui est rarement possible. En effet, à quel niveau de diversité s'attendre sans ce point de comparaison (ce qui nous manque est en fait la connaissance du taux de mutation) ?

Il est apparu plus informatif et plus robuste (car plus indépendant du niveau de diversité) de s'intéresser à l'effet différentiel des variations d'effectif sur *plusieurs* statistiques [141, 45]. Considérons par exemple [89, 24] le couple de statistiques (A, P) où A est le nombre d'allèles et P l'homozygotie calculée). Lorsqu'une population connait un déclin, des allèles sont d'emblée perdus. Ce sont majoritairement des allèles qui étaient en faible fréquence dans la population, et qui contribuaient peu à l'hétérozygotie [89]. Un déclin a donc pour conséquence un excès transitoire de l'hétérozygotie, comparée à sa distribution attendue à léquilibre mutation-dérive conditionnellement à la valeur du nombre d'allèles. Exactement de la même façon, lors d'un accroissement de taille, les mutations apportent des allèles neufs qui comptent pour un, alors qu'ils ne participent que très peu à l'hétérozygotie, d'où un déficit d'hétérozygotie calculée [24]. Ces propriétés ont été exploitées par J.M. Cornuet et G. Luikart dans une série d'articles à l'origine du logiciel BOTTLENECK [24, 105]. L'encart page 59 définit une autre mesure de déséquilibre basée sur ce principe de rupture de relations entre statistiques, utilisable sur des données microsatellites (indice de déséquilibre β de Kimmel).

On peut construire des méthodes d'inférence basées sur le calcul de statistiques (souvent des moments de distributions, d'où l'appellation collective de méthodes des moments). Les plus récentes de ces méthodes tirent parti des avantages de la simulation du processus de coalescence, et consistent à chercher les valeurs des paramètres d'un modèle qui minimisent la distance entre les valeurs espérées des statistiques et leurs valeurs observées.

Pour des données de séquence, la distribution la plus couramment utilisée est celle du nombre de différences entre paires de séquences ('mismatch distribution'). On trouvera une présentation de méthodes des moments utilisant cette distribution, appliquée à l'étude des populations humaines, dans des articles de A. Rogers [117, 119, 60, 118]. Pour les microsatellites, des approches équivalentes ont été basées sur la distribution du nombre de répétitions du motif constitutif. D. Goldstein *et al.* utilisent par exemple la variance et la kurtosis de cette distribution [52]. A. Renwick *et al.* se servent de trois statistiques, l'hétérozygotie calculée, la variance et une mesure apparentée à la skewness [114]. Même ainsi, l'analyse d'un jeu de données humaines gigantesque (5800 loci microsatellites, 20 individus minimum par locus) ne permet pas de choisir entre une histoire d'effectif stable, d'expansion en marche d'escalier ou de bottleneck transitoire [114]. Pour peu que l'on considère une variabilité du taux de mutation entre loci et un modèle mutationnel avec sauts, une bonne adéquation avec les données peut être obtenue pour n'importe lequel des trois scénarios démographiques envisagés [114]. Les données et/ou la méthode ne sont donc pas très discriminants quant à l'histoire des variations d'effectif.

Index de déséquilibre de Kimmel

Dans un article de 1998, Kimmel *et al.* [71] introduisent un index $\widehat{\beta}$ pour la détection de déséquilibres génétiques d'après des données microsatellites. Cet index mesure le rapport entre deux estimateurs à l'équilibre génétique du taux de mutation renormalisé θ. Le premier est l'estimateur $\widehat{\theta}_v$ de la variance de taille. Si l'on note X_i le nombre de répétitions du motif pour le $i^{\text{ème}}$ gène d'un échantillon de taille n,

$$\widehat{\theta}_v = \widehat{V} = \frac{1}{n(n-1)} \sum_{i \neq j} (X_i - X_j)^2 = \sum_{i=n}^{n} \frac{2}{n-1} (X_i - E(X_i))^2.$$

Le second est l'estimateur de l'homozygotie, qui s'exprime

$$\widehat{\theta}_p = \frac{1/\widehat{P}^2 - 1}{2} \quad \text{avec} \quad \widehat{P} = \frac{n \sum_{k=1}^{n} p_k^2 - 1}{n-1} \quad \text{où les } p_k \text{ sont les fréquences des allèles.}$$

Dans une population à l'équilibre mutation-dérive, sous le modèle SMM strict, $\widehat{\beta} = \widehat{V}/\widehat{P} = 1$ en espérance, pour chaque locus. Les deux estimateurs de θ réagissent différemment à des variations d'effectif, d'où un écart potentiellement détectable de $\widehat{\beta}$ par rapport à 1. Un tel écart est plus généralement le signe d'un déséquilibre génétique, ou d'une violation du modèle de mutation SMM. Pratiquement, β peut être estimé de plusieurs façons. Si L est le nombre de loci (indexés par i) dans l'échantillon, un premier estimateur est le rapport des moyennes,

$$\widehat{\beta}_r = \frac{\sum_{i=1}^{L} \widehat{\theta}_v^i}{\sum_{i=1}^{L} \widehat{\theta}_p^i},$$

un autre est la moyenne des rapports,

$$\widehat{\beta}_m = \frac{1}{L} \sum_{i=1}^{L} \left(\frac{\widehat{\theta}_v^i}{\widehat{\theta}_p^i} \right).$$

À l'équilibre génétique, les deux estimateurs ont des distributions en cloche, mais comme le remarquent les auteurs, alors que $\widehat{\beta}_r$ ne dépend que des distributions empiriques marginales de $\widehat{\theta}_v$ et de $\widehat{\theta}_p$, $\widehat{\beta}_m$ dépend de leur distribution jointe. Pour des jeux de données de taille importante (60 loci, 40 individus dans [74]), la puissance des tests basés sur l'estimation de β est bonne pour détecter des croissances exponentielles [74]. Avec $\widehat{\beta}_r$ la puissance s'effondre toutefois lorsque le taux de mutation diffère fortement entre loci. La puissance est meilleure, pour tous les cas illustrés dans [74], que celle de tests basés sur la forme de la distribution de taille allélique [113], ou sur la variance entre loci de l'un des estimateurs de θ [113].

3.2 Approches probabilistes basées sur la vraisemblance

Des avancées de la génétique théorique, de la méthodologie statistique et les progrès des matériels informatiques ont rendu possible, depuis une petite dizaine d'années [56, 55], le développement et la mise en pratique de méthodes probabilistes pour le traitement des données génétiques. Très généralement, les méthodes d'inférence probabilistes dont nous allons parler sont basées sur la construction d'un modèle d'histoire et de fonctionnement des populations, spécifié par des paramètres. Le vecteur ϕ de paramètres, à valeurs dans un espace Φ, est ce que l'on souhaite inférer au vu d'un jeu de données D. Pour cela, on traite les données comme étant une réalisation d'un processus aléatoire défini par le modèle, et on calcule la probabilité $p(D|\phi)$ d'obtention d'une réalisation identique à l'observation, pour une valeur fixée de ϕ. Vue comme une fonction de ϕ, cette probabilité est appelée la vraisemblance $L(\phi) = p(D|\phi)$. Cette approche rend possible l'utilisation des données sous leur forme brute et complète : les données que l'on considère sont tout simplement les comptages alléliques à des marqueurs moléculaires.

Avant de montrer comment calculer la vraisemblance, nous allons présenter deux logiques différentes pour en tirer des inférences : la méthode du maximum de vraisemblance et les approches bayésiennes. Les facteurs qui déterminent le choix de l'approche à utiliser sont multiples : présence ou absence d'informations auxiliaires, préférences méthodologiques, faisabilité des calculs, taille de l'échantillon, risque accepté ...

3.2.1 Maximum de vraisemblance versus approche bayésienne

Dans l'approche par *maximum de vraisemblance*, on estime un paramètre α par la valeur $\widehat{\alpha}$ de plus grande vraisemblance, c'est à dire par la valeur qui donne la plus forte probabilité à l'événement de configuration identique à celle du jeu de données. L'incertitude sur $\widehat{\alpha}$ est typiquement représentée par un intervalle de confiance d'interprétation fréquentiste [22, 59, 137]. L'approche par maximum de vraisemblance est extrêmement efficace lorsque la distribution de l'estimateur d'un paramètre converge (quand la quantité de données tend vers $+\infty$) vers une distribution normale de moyenne la vraie valeur de α. Dans ce cas, la théorie asymptotique s'applique, et la distribution de la statistique $-2\log[L(\alpha_v)/L(\widehat{\alpha})]$ converge vers une distribution du χ^2, avec α_v la vraie valeur du paramètre. Cela fournit un moyen simple de déterminer les bornes d'un intervalle de confiance au seuil 95% pour α : ces bornes sont données par les valeurs de vraisemblance 100 fois plus petite que le maximum de vraisemblance. L'application de la théorie asymptotique en génétique des populations est rarement valide [137, 97]. En effet, les observations à un locus ne sont pas, comme le demande la théorie asymptotique, indépendantes, mais corrélées par leur généalogie commune. Une fois $n > 30$ individus typés à un locus, l'effort qui consiste à en typer d'autres est pratiquement inutile. La plupart de l'information potentiellement disponible est déjà présente dans le premier sous-échantillon. Les modèles biologiquement réalistes sont fréquemment spécifiés par plus d'un paramètre, ce qui rend les choses plus délicates encore. Des solutions peuvent être obtenues numériquement [97, 98], par exemple par des méthodes de reéchantillonnage. La demande en temps de calcul et l'usage délicat de ces solutions diminuent considérablement l'attrait de l'approche par maximum de vraisemblance, dans les applications de génétique des populations, et notamment lorsque les modèles sont spécifiés par un grand nombre de paramètres.

Les méthodes d'*inférence bayésienne* utilisent elles aussi la surface de vraisemblance (éventuellement multiparamétrée), mais l'exploitent d'une façon différente. En effet, ces méthodes

expriment les inférences en termes de loi de probabilité sur l'espace Φ des paramètres (ou loi *a posteriori*). Cela nécessite de formaliser l'ensemble du modèle dans un langage probabiliste, et en particulier de se donner une loi de probabilité *a priori* (avant la révélation des données) sur Φ. L'effet du jeu de données est alors de permettre une mise à jour de ces *a priori* (les *a priori* sont modifiés par l'observation des données). De façon équivalente, la vraisemblance est pondérée par la loi *a priori* pour donner la loi *a posteriori*. Sur le principe, l'approche bayésienne est tout à fait simple (voire triviale). Elle repose sur la formule d'inversion de Bayes, qui relie la probabilité jointe de deux événements à leurs probabilités conditionnelles respectives. Si les deux variables aléatoires sont les données D et les paramètres ϕ, on a $p(D|\phi)p(\phi) = p(D,\phi) = p(\phi|D)p(D)$, et donc

$$p(\phi|D) = \frac{p(D|\phi)p(\phi)}{p(D)} \tag{3.1}$$

L'encart page 62 applique la formule d'inversion de Bayes dans un cas élémentaire tiré de la génétique des pedigree. On retiendra l'idée que la loi *a posteriori* $p(\phi|D)$ a forte densité pour des valeurs à la fois vraisemblables (compatibles avec les données) et compatibles avec l'information *a priori*, extérieure aux données. La loi *a posteriori* en tant que telle représente la loi des paramètres conditionnellement aux observations, et exprime nos incertitudes sur les valeurs des paramètres. À aucun moment la définition des intervalles de plus forte densité *a posteriori* —qui sont l'analogue bayésien des intervalles de confiance— ne repose sur des hypothèses concernant la forme de la distribution *a posteriori* : peu importe dans les approches bayésiennes que la théorie asymptotique ne s'applique pas.

Des critiques sont souvent émises, concernant la subjectivité des approches bayésiennes. Il est vrai que des *a priori* différents vont conduire à des inférences différentes, pour un même modèle démographique et mutationnel, et pour un même jeu de données. La loi *a priori* sur l'espace des paramètres n'est toutefois que l'une des manifestations du statisticien, avec la spécification du modèle stochastique à l'origine des données. Si l'on englobe le choix des lois *a priori* sous le terme de modèle, il devient tout a fait normal que des modèles différents conduisent, pour un même jeu de données, à des inférences différentes. L'accusation de subjectivité qui est faite parfois aux approches bayésiennes vaut tout autant pour les autres approches, qui nécessitent elles aussi la spécification subjective d'un modèle. Une vraie question, pertinente, est celle de savoir si un jeu de données contient suffisamment d'information pour une question d'inférence. Par une analyse de sensibilité des résultats aux lois *a priori*, l'approche bayésienne permet d'aborder cette question de façon rigoureuse. Une autre question est la définition de lois *a priori* utilisables lorsqu'on ne dispose pas d'informations *a priori*.

Formule d'inversion de Bayes appliquée à la génétique des pedigree

L'hémophilie est une maladie récessive liée au chromosome X. Considérons une femme F et son frère M atteint par la maladie. Leur mère, non malade, est donc porteuse hétérozygote du gène esponsable de la maladie. Sachant que leur père n'est pas hémophile, F a une probabilité 0.5 d'être porteuse saine du gène de l'hémophilie. La quantité ϕ qui nous intéresse est l'état de F, qui peut être porteuse ($\phi = 1$) ou non ($\phi = 0$) de l'allèle déficient. La synthèse des informations familiales disponibles permet de proposer les *a priori* suivants pour ϕ : les probabilités des deux états sont égales $p(\phi = 0) = p(\phi = 1) = 0.5$.
Pour mettre à jour ces *a priori*, on utilise une nouvelle information : l'affection d_i des fils i de F par l'hémophilie. Supposons que F a deux fils (non jumeaux) qui ne sont pas hémophiles. Conditionnellement à ϕ, d_1 et d_2 sont indépendants. En négligeant la probabilité d'apparition de l'allèle responsable de l'hémophilie par mutation, la vraisemblance est donnée par

$$p(d_1 = d_2 = 0|\phi = 1) = 0.5 \times 0.5 = 0.25 \ ,$$

$$p(d_1 = d_2 = 0|\phi = 0) = 1 \times 1 = 1 \ .$$

La formule d'inversion de Bayes permet de combiner l'information *a priori* et les données, pour calculer la probabilité *a posteriori* que F soit porteuse :

$$\begin{aligned} p(\phi = 1|d_1 = d_2 = 0) &= \frac{p(d_1 = d_2 = 0|\phi = 1)p(\phi = 1)}{p(d_1 = d_2 = 0)} \\ &= \frac{p(d_1 = d_2 = 0|\phi = 1)p(\phi = 1)}{p(d_1 = d_2 = 0|\phi = 1)p(\phi = 1) + p(d_1 = d_2 = 0|\phi = 0)p(\phi = 0)} \\ &= \frac{0.25 \times 0.5}{0.25 \times 0.5 + 1 \times 0.5} = \frac{0.125}{0.625} = 0.2 \end{aligned}$$

Intuitivement, le fait de savoir que les fils de F sont sains diminue la probabilité conditionnelle qu'elle soit porteuse, et la formule d'inversion de Bayes permet de calculer cette probabilité. Un aspect clé des approches bayésiennes est la facilité d'incorporation de nouvelles données. Par exemple, si F a un troisième fils, sain lui aussi, on peut utiliser comme information *a priori* la loi *a posteriori* obtenue pour deux fils sains :

$$p(\phi = 1|d_1 = d_2 = d_3 = 0) = \frac{0.5 \times 0.2}{0.5 \times 0.2 + 1 \times 0.8} = 0.111 \ .$$

$p(\phi = 1)$ diminue au fur et à mesure que le nombre de fils sains de F augmente, et le premier fils hémophile qui naît impose brutalement $p(\phi = 1) = 1$ (toujours si l'on ignore la mutation).

Cet exemple élémentaire (tiré de [48]) ne doit pas cacher la difficulté de mise en oeuvre des approches bayésiennes, lorsque les variables aléatoires d'intérêt sont multiples et continues. Mais il illustre le principe très simple de la formule d'inversion de Bayes, principe qui reste valable dans les cas complexes.

Inférence bayésienne et inférence par maximum de vraisemblance

On souhaite inférer le nombre n de bus desservant une ville, d'après la liste des numéros des bus que l'on voit passer (les hypothèses sont les mêmes qu'à l'encart de la page 45).

Supposons pour commencer que l'on ne voie passer qu'un bus numéroté k pendant le temps de l'expérience. Il est bien sûr impossible que k soit strictement plus grand que n. Pour $k \leq n$ en revanche, il y a probabilité $1/n$ d'observer le bus k. Pour chaque n, on a donc une loi de probabilité P_n sur l'espace $\{1, 2, \ldots\}$ des observations possibles : $P_n(k) = (1_{k \leq n})/n$.

L'approche par *maximum de vraisemblance* conduit à inférer pour n la valeur $\hat{n}$ qui maximise la probabilité $P_n(k)$ de l'observation k : en l'occurrence, l'estimateur du maximum de vraisemblance de n est $\widehat{n} = k$. Dans l'approche *bayésienne*, on se donne une loi a priori pour le paramètre n, soit $p(n)$, et on considère P_n comme une probabilité conditionnelle : $P_n(k) = P(k|n)$ est la probabilité d'observer le bus k sachant que le paramètre est n. On cherche alors a déterminer la loi conditionnelle de n, sachant que l'on a effectué l'observation k ; cette loi est dite loi *a posteriori*. Notons $P(\cdot|k)$ cette probabilité conditionnelle. Par application des règles élémentaires de calcul des probabilités, on trouve $P(n|k) = P(k|n)p(n)/P(k)$. Le terme $P(k)$ peut se calculer par décomposition selon un système complet d'événements pour n, $P(k) = \sum_n P(k|n)p(n)$. Un choix naturel de loi a priori consisterait à utiliser pour p la distribution du nombre de bus dans des villes de même taille pour lesquelles ce nombre est connu.

Les résultats donnés par les deux méthodes ne sont pas de même nature : l'estimateur de maximum de vraisemblance propose une valeur, tandis que l'approche bayésienne propose une loi (dans le cas présent, toutes les valeurs de n supérieures à k vont avoir une probabilité non nulle, pourvu qu'on ne les ait pas exclues a priori). Pour en déduire une valeur estimée, on peut calculer la valeur moyenne de n sous cette loi, ou bien la valeur de n qui maximise la loi (le mode).

Dans cet exemple, le maximum de vraisemblance $\hat{n}$ donne en moyenne un très mauvais résultat : si n est le véritable nombre de bus, la valeur moyenne observée est $n/2$, ce qui est aussi la valeur moyenne de $\hat{n}$. L'estimateur est donc fortement biaisé. Un autre exemple montrera comment un choix "absurde" de loi a priori mène à un résultat peu significatif par la méthode bayésienne. Partant d'une loi a priori de Poisson de paramètre λ, on trouve que le maximum de la loi *a posteriori* est obtenu pour $n = k$, ce qui rejoint l'estimateur du maximum de vraisemblance. La valeur moyenne de n conditionnellement à l'observation tend vers k quand λ tend vers 0, mais est asymptotiquement équivalente à λ quand λ tend vers l'infini : dans ce régime, c'est la loi a priori qui impose la moyenne. Il faut bien sûr prendre garde à cet écueil quand on emploie une méthode bayésienne. On note que la distribution est extrêmement étalée pour λ grand, puisque le mode est en k et la moyenne proche de λ : il s'agit là d'un indice du peu de précision de l'inférence.

La différence entre les approches par maximum de vraisemblance et bayésienne persiste-t-elle lorsque les données sont plus nombreuses, et que la liste des numéros des bus converge vers une distribution uniforme sur $[1; n]$? La réponse par maximum de vraisemblance demeure le plus grand numéro de bus observé, ce qui avec probabilité 1 converge vers la bonne réponse quand le nombre de bus observés tend vers $+\infty$. La réponse bayésienne tend également vers la bonne valeur de n, car les valeurs de n qui ont du poids *a priori* mais qui sont supérieures au plus grand numéro observé correspondent à des termes $P(k_1, ...k_i...|n)$ dont la probabilité tend vers 0. Les deux méthodes se réconcilient donc dans notre exemple, lorsque la quantité de données augmente.

3.2.2 Coalescent et calcul de vraisemblance

Pratiquement, comment calcule-t-on la vraisemblance $L(\phi) = p(D|\phi)$? Le jeu de données D est, comme nous l'avons dit, considéré comme une réalisation d'un processus stochastique qui détermine sa généalogie et son histoire mutationnelle. Le calcul de la vraisemblance est basé sur l'hypothèse que la loi de la généalogie est donnée par un coalescent. Cela peut être le coalescent standard pour un modèle de population close, d'effectif stable, se reproduisant selon le modèle de Wright-Fisher. Mais cela peut aussi être un coalescent avec variations d'effectif, un coalescent structuré, un coalescent avec sélection ou encore recombinaison, selon les complications du modèle supposé (voir revue sur les extensions du coalescent standard dans [99]).

La généalogie de l'échantillon, dénotée par $\mathcal{G}$, peut être considérée comme une donnée manquante. On peut donc réécrire $p(D|\phi)$ comme la somme des vraisemblances sur l'espace Γ de toutes les généalogies possibles :

$$p(D|\phi) = \int_{\Gamma} p(D|\phi, \mathcal{G}) p(\mathcal{G}|\phi) d\mathcal{G} \quad , \tag{3.2}$$

où $p(D|\phi, \mathcal{G})$ est plus ou moins facilement calculable et $p(\mathcal{G}|\phi)$ est donnée par la théorie de la coalescence (application pratique au paragraphe 5.1). Le calcul de $p(D|\phi, \mathcal{G})$ dépend de ce que l'on désigne précisément par $\mathcal{G}$: $\mathcal{G}$ peut selon les méthodes être l'histoire généalogique et mutationnelle toute entière, ou l'histoire généalogique seule. Dans le cas particulier où le choix de $\mathcal{G}$ détermine de façon unique la configuration en bouts de branches de la généalogie, $p(D|\phi, \mathcal{G})$ est tout simplement une variable indicatrice, c'est à dire qu'elle prend les valeurs *1/0* selon que cette configuration *est/n'est pas* celle du jeu de données.

Une caractéristique importante des coalescents neutres est que la vraisemblance d'une généalogie ne dépend pas des détails de l'histoire : peu importe quelles lignées fusionnent dans quel ordre, quelles branches sont affectées par les mutations. Seule compte la répartition des événements de coalescence et de mutation le long de l'axe des temps (figure 4.2), c'est à dire leurs types et leurs dates. Autrement dit, on peut projeter une généalogie sur l'axe des temps, et la vraisemblance est tout simplement le produit des densités d'occurrence des événements successifs aux dates successives. Plus précisément, si l'on appelle $\gamma(t)$ la densité d'occurrence des événements, la densité de probabilité conditionnelle $p(t_{i+1}|t_i)$ d'occurrence d'un quelconque événement à la date t_{i+1}, sachant que l'événement précédent a eu lieu à l'instant t_i, est le produit du taux instantané $\gamma(t_{i+1})$ et de la probabilité pour qu'aucun événement ne se soit produit entre t_i et t_{i+1} [55] :

$$p(t_{i+1}|t_i) = \gamma(t_{i+1}) \exp\left(- \int_{t_i}^{t_{i+1}} \gamma(t) dt \right).$$

Et si la généalogie considérée comporte $card(\mathcal{E})$ événements, sa densité de probabilité est proportionnelle au produit des probabilités $p(e, t_{i+1}|t_i)$ des événements particuliers :

$$\prod_{i=0}^{card(\mathcal{E})-1} p(\{e, t_{i+1}\}|t_i).$$

Le nombre des topologies non datées possibles croît très vite avec la taille de l'échantillon (tableau 3.1), et l'espace des possibles pour les autres composantes de $\mathcal{G}$ (nombre, position et type des mutations, dates des événements) n'est pas fini. L'espace Γ des généalogies de gènes est

n	2	3	4	5	6	7	8	9	10
t	1	3	18	180	2 700	56 700	1 587 600	57 153 600	2 571 912 000

TAB. *3.1: Relation entre la taille n d'un échantillon de gènes et le nombre t de topologies possibles pour la généalogie de l'échantillon.* $t = \prod_{k=2}^{n} C_k^2$ *augmente très rapidement avec n, ce qui rend impossible l'exploration de toutes les généalogies.*

donc de très grande dimension, partiellement discret (topologie des généalogies), partiellement continu (dates des événements).

On a donc à calculer une intégrale —une somme pour la composante discrète— en très grande dimension, ce qui rend inapplicables les méthodes numériques déterministes (méthodes de Riemann et de Lagrange *sl.*), dont le principe est de calculer la valeur de la fonction à intégrer sur un maillage de l'espace des paramètres, et de calculer l'intégrale par interpolation. Les méthodes de Monte Carlo —qui procèdent par tirages aléatoires dans cet espace— sont le seul moyen pratique d'approcher l'intégrale de l'équation 3.2, même si, comme A. Sokal le remarque en tête d'une série de cours sur l'usage des méthodes de Monte Carlo en mécanique statistique [135] :

"Monte Carlo is an extremely bad method. It should be used only when all alternative methods are worse.". En effet, les méthodes de Monte Carlo peuvent au mieux converger en $O(1/\sqrt{n})$ [58]. De plus, tout biais dans les simulations se reporte sur les estimations, point particulièrement délicat pour des modèles complexes —dont les problèmes généalogiques qui nous intéressent.

3.3 Échantillonnage de Monte Carlo

Toutes les méthodes de Monte Carlo consistent à générer des tirages selon une loi de probabilité f définie sur un espace $\mathcal{F}$ —dans le cas qui nous intéresse, f est la vraisemblance $L(\phi) = p(D|\phi)$ ou bien la loi $p(\phi|D)$, appelée loi *a posteriori*, proportionnelle à $p(D|\phi)p(\phi)$ pour D donné. Deux grands types peuvent être distingués, les méthodes de Monte Carlo statiques et dynamiques. Les méthodes *statiques* (Monte Carlo standard, échantillonnage pondéré) fournissent une séquence de tirages indépendants selon f. Les méthodes *dynamiques* (Monte Carlo par chaîne de Markov) sont basées sur l'invention d'une marche aléatoire qui permette d'explorer l'espace F, et qui ait pour unique distribution stationnaire la loi f. Les méthodes dynamiques génèrent des tirages asymptotiquement distribués selon f, mais corrélés.

3.3.1 Méthode de Monte Carlo standard

Une façon conceptuellement simple et de mise en oeuvre algorithmique rapide de calculer $L(\phi)$ est d'approcher 3.2 en faisant la moyenne d'un nombre suffisamment grand de termes correspondant chacun à un tirage selon $p(\mathcal{G}|\phi)$:

$$L(\phi) = p(D|\phi) = \int_{\Gamma} p(D|\phi,\mathcal{G})p(\mathcal{G}|\phi)\, d\mathcal{G} \approx \frac{1}{N_{\mathcal{G}}} \sum_{i=1}^{N_{\mathcal{G}_i}} p(D|\phi,\mathcal{G}) \;, \tag{3.3}$$

Tirer des généalogies selon $p(\mathcal{G}|\phi)$ est aisé dans le cadre du coalescent (nous le ferons au chapitre 5). Malheureusement, pour des jeux de données de dimension usuelle, la plupart des

généalogies échantillonnées donnent une probabilité nulle (ou très faible) d'obtenir la configuration de l'échantillon (encart page 66). De ce fait, l'estimateur de Monte Carlo de $L(\phi)$ a une très grande variance, et l'obtention de résultats fiables nécessite d'estimer $L(\phi)$ sur un très grand nombre $N_{\mathcal{G}}$ de généalogies (nombre inatteignable dans la pratique [138]). La méthode de Monte Carlo standard devient alors totalement inefficace.

Limites de la méthode de Monte Carlo directe

La possibilité de calculer $p(D|\phi)$ par échantillonnage de Monte Carlo d'histoires généalogiques et mutationnelles dépend de la probabilité de générer ainsi la configuration de D. Une série d'échantillons de tailles n croissantes illustre cette probabilité (les détails de la méthode de Monte Carlo utilisée seront donnés au chapitre 5.4). Ces échantillons ont tous été générés en population stable avec un modèle mutationnel SMM, et $\theta = 10$. On s'interesse par exemple à la vraisemblance en fonction de θ (les autres paramètres définissent une population stable et un modèle SMM). Cette vraisemblance peut être estimée par la méthode de Monte Carlo directe, pour chacun des échantillons D, comme la proportion $c = f/S$ de simulations où l'échantillon cible est atteint, avec f le nombre de fois où léchantillon cible est simulé et S le nombre total de simulations. Dans le tableau suivant, n est la taille des échantillons, D leurs configurations normalisées (encart page 78), et $c \pm \sigma(c)$ la proportion de fois où léchantillon cible D est atteint, et l'écart-type de cette proportion (le tout pour $S = 100\,000$).

n	D	f	$c \pm \sigma(c)$
2	$\{1,1\}$	28 131	$2.8\,10^{-1} \pm 1.4\,10^{-3}$
3	$\{3\}$	6 422	$6.4\,10^{-2} \pm 7.8\,10^{-4}$
4	$\{2,1,1\}$	4 042	$4.0\,10^{-2} \pm 6.2\,10^{-4}$
5	$\{2,1,2\}$	1 810	$1.8\,10^{-2} \pm 4.3\,10^{-4}$
6	$\{2,2,1,1\}$	937	$1.8\,10^{-2} \pm 4.3\,10^{-4}$
7	$\{2,0,2,3\}$	219	$9.4\,10^{-3} \pm 3.1\,10^{-4}$
10	$\{3,2,2,3\}$	107	$1.1\,10^{-3} \pm 1.0\,10^{-4}$
15	$\{1,9,5\}$	45	$4.5\,10^{-4} \pm 6.7\,10^{-5}$
20	$\{6,1,7,4,2\}$	3	$3.0\,10^{-5} \pm 1.7\,10^{-5}$

Pour des tailles d'échantillon de plus d'une dizaine de gènes, la méthode directe passe le plus clair de son temps à générer des généalogies incompatibles avec l'échantillon, et estime la vraisemblance du jeu de données avec une forte variance.

3.3.2 Échantillonnage d'importance

Une amélioration directe de la méthode de Monte Carlo standard —qui en conserve le principe— consiste à concentrer les tirages selon $p(\mathcal{G}|\phi)$ sur les termes pour lesquels $p(D|\phi,\mathcal{G})$ est le plus grand, afin de réduire la variance des estimateurs. Une introduction aux méthodes d'échantillonnage d'importance (IS), ou échantillonnage pondéré, peut être trouvée dans l'ouvrage de référence de Ripley [115]. L'application à l'inférence à partir de données de diversité

génétique est développée dans un article de fond de Stephens et Donnelly ([138], voir aussi [137, 136]).

En quelques mots, on se donne une distribution de probabilité $q(\mathcal{G}|\phi)$ selon laquelle on sait réaliser des tirages, et de densité non nulle pour les généalogies compatibles avec les données, c'est à dire les généalogies telles que $p(D, \mathcal{G}) > 0$. $L(\phi)$ peut alors être exprimé sous la forme

$$p(D|\phi) = \int_{\Gamma} p(D|\phi, \mathcal{G}) \frac{p(\mathcal{G}|\phi)}{q(\mathcal{G}|\phi)} \, q(\mathcal{G}|\phi) \, d\mathcal{G} \tag{3.4}$$

Le rapport $p(\mathcal{G}|\phi)/q(\mathcal{G}|\phi)$ est appelé rapport d'importance, et il est égal à la probabilité d'échantillonner la généalogie $\mathcal{G}$ pour le modèle stochastique (le coalescent) divisée par la probabilité d'échantillonner cette généalogie par tirage selon la fonction d'importance. L'estimateur de $p(D|\phi)$ correspondant, obtenu en moyennant sur un grand nombre de généalogies $\mathcal{G}_i$ indépendantes, tirées selon la distribution q, est :

$$p(D|\phi) \approx \frac{1}{N_{\mathcal{G}}} \sum_{i=1}^{N_{\mathcal{G}}} p(D|\phi, \mathcal{G}_i) \frac{p(\mathcal{G}_i|\phi)}{q(\mathcal{G}_i|\phi)} \tag{3.5}$$

En général, on fait en sorte que toutes les généalogies G_i échantillonnées soient compatibles avec léchantillon, c'est à dire que $p(D|\phi, \mathcal{G}_i) = 1$ [55, 138] . La distribution q idéale pour diminuer la variance de l'estimateur de la vraisemblance serait $q(\mathcal{G}|\phi) = p(\mathcal{G}|D, \phi)$. En effet, dans ce cas, chaque terme de la somme dans 3.5 vaut $p(\mathcal{G}_i|\phi)/p(\mathcal{G}_i|D, \phi)$, qui en remarquant que $p(\mathcal{G}_i, \phi) = p(\mathcal{G}_i, D, \phi)$ devient exactement la vraisemblance $p(D|\phi)$. La variance de l'estimateur est donc nulle. Plus raisonnablement, il faut trouver un compromis entre les propriétés souhaitables pour q, c'est à dire la facilité de simulation de tirages selon q, et la diminution de variance de l'estimateur [138].

Dans la pratique, il est extrêmement difficile de choisir une distribution q efficace pour une large gamme de valeurs des paramètres ([138, 137]. q est généralement construite à l'aide d'une valeur pivot du paramètre, et n'échantillonne efficacement les généalogies importantes qu'à proximité de cette valeur pivot. Lorsque la distribution d'importance q utilisée est la même pour toutes les valeurs des paramètres, on sous-estime typiquement la vraisemblance pour les valeurs des paramètres distantes de la valeur pivot, du fait que q manque déchantillonner les généalogies qui contribuent à la vraisemblance. Cela risque de donner la fausse impression que la vraisemblance a un maximum proche de la valeur pivot, et de donner des intervalles de plus forte vraisemblance plus étroits qu'ils ne devraient être. Les méthodes dites de bridge-sampling [42] et les méthodes d'échantillonnage groupé [137] cherchent à résoudre ce problème.

De plus, l'approche par échantillonnage d'importance partage avec la méthode standard l'inconvénient de nécessiter une estimation de la vraisemblance sur un maillage de l'espace des paramètres démographiques et mutationnels. Cela reste gérable lorsque le nombre de paramètres d'intérêt est faible, mais explorer ce maillage devient une tâche démesurée pour des modèles spécifiés par plusieurs paramètres. Le modèle traité dans la partie II, qui comporte dans sa forme complète 4 paramètres de fondation-explosion et 3 paramètres pour spécifier le modèle de mutation des marqueurs, serait par exemple difficile à traiter par importance sampling [46], et les méthodes dynamiques semblent avoir l'avantage dans des systèmes avec beaucoup de degrés de liberté comme celui-là [138].

3.3.3 Échantillonnage de Monte Carlo par chaîne de Markov

Dans les méthodes d'échantillonnage de Monte Carlo par Chaîne de Markov (MCMC), on invente un processus stochastique qui définisse une marche aléatoire sur $\mathcal{F}$ ayant pour unique distribution stationnaire la loi f selon laquelle on souhaite échantillonner. La marche aléatoire assure l'exploration de l'espace $\mathcal{F}$, et sous des conditions assez générales, les positions successives de la marche aléatoire dans $\mathcal{F}$ seront distribuées selon f. Notons dès à présent que le processus stochastique est choisi de façon à explorer $\mathcal{F}$, et ne cherche en rien à mimer quelque phénomène biologique que ce soit. Dans la pratique, le processus stochastique est une chaîne de Markov, construite à partir d'un état initial arbitraire en suivant l'algorithme de Metropolis-Hastings. Nous allons donc donner quelques éléments sur les chaînes de Markov et sur l'algorithme de Metropolis-Hastings (on se reportera pour plus de détails aux références [50, 48]).

Une chaîne de Markov est, de façon simple, une séquence $X_0, X_1, \ldots, X_n, X_{n+1}, \ldots$ de variables aléatoires à valeurs dans $\mathcal{F}$, telles que X_{n+1} ne dépende que de X_n, de façon probabiliste. Une chaîne est définie par une distribution sur $\mathcal{F}$ pour la variable aléatoire initiale X_0, et par un noyau de transition qui donne la probabilité $p_{x\to y}$ pour que X_{n+1} prenne la valeur y sachant que X_n a pour valeur x. On dit qu'une chaîne de Markov est irréductible si il est possible d'aller de n'importe quel état x à n'importe quel état y (x, y dans $\mathcal{F}$), en un nombre fini de pas (voir [48, 142, 116] pour des définitions plus rigoureuses, tenant compte en particulier du caractère continu de l'espace d'état). L'irréductibilité est nécessaire à l'exploration de l'espace des paramètres à des fins d'échantillonnage. Une densité de probabilité f sur $\mathcal{F}$ est appelée distribution stationnaire (ou invariante) d'une chaîne de Markov si

$$f(y) = \sum_{x\in\mathcal{F}} f(x)p(x \to y) \ ,$$

pour tout y dans $\mathcal{F}$. Remarquons que cette somme est plutôt une intégrale, pour la composante continue de l'espace d'état.

Une chaîne de Markov irréductible n'admet pas nécessairement de distribution stationnaire, mais si elle en admet une, celle-ci est unique [116]. Si la chaîne est apériodique (*i.e.* si pour tout x, $p(x \to x) > 0$), la chaîne de Markov converge vers la distribution f, quelle que soit la distribution initiale pour X_0 (là aussi, les choses sont plus complexes pour un espace d'état continu). En théorie, pour tirer selon f par une méthode de Monte Carlo dynamique, il suffit de se donner un noyau de transition $p(x \to y)$ qui satisfait aux conditions d'irréductibilité, d'apériodicité et de stationarité pour f. On peut alors partir de n'importe quel état initial et on a la garantie que le système converge vers la distribution stationnaire. Reste à déterminer des noyaux de transition qui donnent une méthode de Monte Carlo non seulement correcte, mais de plus efficace. La difficulté clé est que les états successifs de la chaîne sont corrélés, si bien que la variance des estimations produites par Monte Carlo dynamique risque d'être plus grande que par Monte Carlo statique. Une méthode de Monte Carlo dynamique est d'autant plus efficace que la corrélation entre états successifs est faible.

L'algorithme de Metropolis-Hastings [62, 48, 80] fournit un moyen tout à fait général de construire une chaîne de Markov irréductible, apériodique, de distribution stationnaire f, lorsque le rapport $f(x)/f(y)$ des densités de probabilité à deux points x et y de l'espace $\mathcal{F}$ peut être calculé. Soit $p(x \to y)$ une noyau de proposition arbitraire défini sur $\mathcal{F} \times \mathcal{F}$ (les propositions pourront être rejetées, c'est pourquoi je parle de noyau de proposition et non pas de noyau de

transition). Si x est la valeur de la variable aléatoire X_n de la chaîne de Markov, l'algorithme de Metropolis-Hastings définit la valeur de X_{n+1} de la façon suivante : un état y est tiré selon le noyau de proposition. Avec probabilité a, où

$$a = min\left(1, \frac{f(y)p(y \to x)}{f(x)p(x \to y)}\right) \tag{3.6}$$

l'état y est accepté et $X_{n+1} = y$ (éventuellement, $y = x$). Avec probabilité $1 - a$, l'état y est rejeté, et la chaîne de Markov reste sur place avec $X_{n+1} = x$. À condition que la chaîne de Markov obtenue soit irréductible (ce qui dépend du choix de la matrice de proposition), elle a pour distribution stationnaire f (encart page 70).

Choix d'une matrice de proposition Deux écueils sont à éviter lors du choix d'une matrice de proposition. Le premier est de proposer des positions qui ont une faible probabilité d'être acceptées. La chaîne a alors tendance à rester sur place, au lieu d'explorer l'espace des paramètres en repassant plus fréquemment dans les régions de forte vraisemblance. Le second est de se déplacer avec lenteur dans l'espace. Même avec un fort taux d'acceptation, on n'explore pas non plus efficacement l'espace dans ces conditions. Malheureusement, le compromis entre ces deux défauts est difficile à trouver. En effet, si l'on propose des positions trop éloignées de la position initiale, il est difficile d'obtenir un fort taux d'acceptation. Et de toute façon, il faut rester capable de calculer les probabilités de proposition $q(x \to y)$ et de proposition inverse $q(y \to x)$ pour pouvoir implémenter l'algorithme de Metropolis-Hastings. Si l'on mise sur un fort taux d'acceptation, on ne peut guère se déplacer dans l'espace des paramètres que par sauts de puce. Pour les modèles généalogiques traités dans [6] ou dans la présente thèse, un compromis raisonnable doit être trouvé par essais-erreurs.

3.3.4 Méthode de Monte Carlo et choix du paradigme statistique

Si ϕ désigne les paramètres d'intérêt d'un modèle, la surface de vraisemblance $L(\phi)$ —calculée ou approchée par des méthodes de Monte Carlo— peut être utilisée pour estimer les paramètres du modèle, et pour exprimer l'incertitude relative à ces inférences. Les méthodes de Monte Carlo statiques et dynamiques donnent toutefois accès à la surface de vraisemblance par des chemins différents. Dans la pratique le choix de la méthode d'échantillonnage oriente vers le traitement par maximum de vraisemblance ou par les méthodes bayésiennes.

Plus précisément, les méthodes de type *Importance Sampling* fournissent la valeur de la vraisemblance en différents points de l'espace des paramètres. C'est donc un chemin direct si l'on souhaite baser l'inférence sur le calcul du maximum de vraisemblance. En revanche, cela rend difficile l'usage de vraisemblances intégrées (marginales) [10], et si l'on adopte une approche bayésienne, l'obtention de lois *a posteriori* demande un travail d'intégration supplémentaire. Les méthodes de *Monte Carlo dynamiques* fournissent quant à elles des échantillons tirés selon la distribution jointe *a posteriori* des paramètres. L'obtention de la surface de vraisemblance, ou de lois *a posteriori*, nécessite un travail d'estimation de densité. Mais il est aussi simple d'obtenir des lois marginales que des lois jointes : il suffit pour cela d'ignorer les paramètres par rapport auxquels on souhaite intégrer, au moment de l'estimation de densité. Pour résumer, l'échantillonnage d'importance est surtout orienté vers l'inférence par maximum de vraisemblance dans sa forme classique, alors que les méthodes qui utilisent l'échantillonnage MCMC permettent en plus l'usage de vraisemblances intégrées et dans une optique bayésienne, l'estimation de lois *a posteriori*.

Stationnarité d'une chaîne de Markov pour la loi f

Soit une chaîne de Markov définie par les variables aléatoires $(X_0, X_1, \dots X_n, X_{n+1} \dots)$. On veut ici montrer que si l'algorithme de Metropolis-Hastings est appliqué pour déterminer les états successifs, alors f est une distribution stationnaire pour la chaîne de Markov. C'est à dire que si la loi $\mathcal{L}(X_n)$ de l'état X_n est f, alors cela reste valable au rang suivant, $\mathcal{L}(X_{n+1}) = f$.
On note $p_{x\to y}$ la probabilité de transition d'un état $X_n = x$ à un état $X_{n+1} = y$, et on pose

$$D = \left\{(x,y); \quad \frac{p_{y\to x}}{p_{x\to y}} \times \frac{f(y)}{f(x)} \leq 1\right\},$$

l'ensemble des couples (x, y) qui satisfont au critère de Metropolis-Hastings.
Soit x un état donné. Evaluons la loi de X_{n+1} au point y, conditionnellement à $X_n = x$. La formule exprimant $p(X_{n+1} = y|X_n = x)$ comporte deux termes d'acceptation et un terme de rejet :

$$\begin{aligned} p(X_{n+1} = y|X_n = x) = p_{x\to y}\frac{p_{y\to x}}{p_{x\to y}}\frac{f(y)}{f(x)}\, \mathbb{I}_{(x,y)\in D} + \; p_{x\to y}\, \mathbb{I}_{(x,y)\notin D} \\ + \left[1 - p_{x\to y'}\frac{p_{y'\to x}}{p_{x\to y'}} \times \frac{f(y')}{f(x)}\, \mathbb{I}_{(x,y')\in D} dy' - \int p_{x\to y'}\, \mathbb{I}_{(x,y')\notin D} dy'\right] \delta_{x=y} \quad (3.7) \end{aligned}$$

où δ_x désigne la masse de Dirac au point x, i.e. la mesure de probabilité définie par $\delta_x[A] = 1$ si $x \in A$, et $\delta_x[A] = 0$. En outre on a noté $\mathbb{I}_e$ la fonction valant 1 si e est vérifié, et 0 sinon.

La loi de X_{n+1}, évaluée en un point y, s'obtient en intégrant l'expression (3.7) contre la mesure de probabilité $f(x)\,dx$: on trouve

$$\begin{aligned} \mathcal{L}(X_{n+1})(y) \;\; = \;\; & \int p(y \to x)\, f(y)\, \mathbb{I}_{(x,y)\in D} dx + \int p(x \to y)\, f(x)\, \mathbb{I}_{(x,y)\notin D} dx \\ & + f(y) - \int p(y' \to y)\, f(y')\, \mathbb{I}_{(y,y')\in D} dy' - \int p(y \to y')\, f(y)\, \mathbb{I}_{(y,y')\notin D} dy'. \end{aligned}$$

Si on renomme y' en x, on obtient après regroupement

$$\mathcal{L}(X_{n+1})(y) \;\; = \;\; f(y) + \int p_{y\to x} f(y)[\mathbb{I}_{(x,y)\in D} - \mathbb{I}_{(y,x)\notin D}]\, dx - \int p_{x\to y} f(x)[\mathbb{I}_{(y,x)\in D} - \mathbb{I}_{(x,y)\notin D}]\, dx.$$

Cependant, les deux derniers termes se compensent, puisque

$$\mathbb{I}_{(x,y)\in D} - \mathbb{I}_{(y,x)\notin D} = \mathbb{I}_{[p_{x\to y} f(x) = p_{y\to x} f(y)]} = \mathbb{I}_{(y,x)\in D} - \mathbb{I}_{(x,y)\notin D}.$$

On obtient donc

$$\mathcal{L}(X_{n+1})(y) = f(y).$$

Dans la pratique, nous avons besoin pour notre échantillonneur de Metropolis-Hastings, en plus de la stationnarité en f, d'une propriété de convergence (de quelque distribution que l'on parte pour X_0, $\mathcal{L}(X_n) \to f$ lorsque $n \to +\infty$) et d'une propriété d'ergodicité (on veut que les états successifs d'une même chaîne soient distribués selon f, et non pas les états X_n de chaînes parallèles).

3.4 Méthodes probabilistes exactes et approchées

Les méthodes probabilistes exactes dont nous avons parlé sont coûteuses en temps de développement et de calcul. On peut gagner sur ces deux tableaux en choisissant des méthodes probabilistes approchées [107, 137]. On manipule pour cela, au lieu des jeux de données bruts, des jeux de données transformés par le calcul de statistiques sommaires. Par exemple, les méthodes d'acceptation-rejet fonctionnent sur le même principe que la méthode de Monte Carlo standard, par simulation à répétition de coalescents indépendants. Mais le jeu de données cible est considéré comme atteint lorsque les données simulées leur ressemblent et non pas lorsqu'elles ont exactement la même configuration. La ressemblance est typiquement définie comme une fonction décroissante de la distance entre un vecteur de statistiques sommaires cible et simulé. Les derniers développements de la méthode d'acceptation-rejet semblent montrer que l'effort de développement des méthodes exactes vaut la peine : un article sous presse de M. Beaumont *et al.* [8] illustre la perte d'information engendrée par l'usage de statistiques sommaires : ce travail compare pour un modèle démographique simple la méthode bayésienne de I. Wilson et D. Balding [154], qui utilise la configuration exacte de l'échantillon, avec des méthodes bayésiennes approchées de type acceptation-rejet [145, 107, 39]. Les méthodes exactes sont bien plus gourmandes en temps de calcul, mais l'usage de statistiques sommaires dans les méthodes approchées engendre une perte d'information substantielle. Les méthodes probabilistes exactes sont déjà tout à fait utilisables pour le traitement d'un jeu de données de taille moyenne (une dizaine de loci, quelques dizaines d'individus). En revanche, un gain de temps de calcul de deux ordres de grandeur serait nécessaire pour que l'exploration des modèles sous-jacents soit confortable et puisse être complète. La combinaison des progrès algorithmiques et de ceux des matériels informatiques ne manqueront pas de rendre les méthodes exactes plus faciles d'usage.

Deuxième partie

Inférence bayésienne des paramètres d'une fondation-explosion

On cherche à faire des inférences sur les variations d'effectif passées d'une population, d'après l'observation de la diversité génétique actuelle dans cette population.

C'est l'existence d'une généalogie de l'échantillon qui permet de faire le lien entre diversité génétique à un instant et variations d'effectif passées (paragraphe 2.3.2). La forme de la généalogie dépend en effet des effectifs et se trouve mémorisée —malheureusement avec perte d'information— par le processus de mutation, sous la forme d'une distribution de comptages alléliques. Pourtant, la plupart des méthodes d'inférence développées jusqu'à présent laissent de côté une partie de l'information démographique enregistrée dans les jeux de données : les méthodes basées sur le calcul de moments de distributions [24], et les méthodes de type acceptation-rejet [107] résument en effet le jeu de données par un petit nombre de statistiques. Les mêmes valeurs des statistiques peuvent être obtenues pour des configurations différentes du jeu de données, qui possiblement correspondent à des surfaces de vraisemblance assez éloignées.

L'histoire généalogique et mutationnelle précise est inconnue, et son enregistrement par un marqueur ne nous informe que partiellement sur elle. Quand bien même l'histoire exacte à un locus serait connue, le scénario de variations d'effectif reste peu accessible. En effet, le lien entre démographie et généalogie est probabiliste, si bien que des inférences précises nécessitent l'observation de plusieurs réalisations indépendantes du processus aléatoire qui sous-tend les données de diversité. Malgré ce fait notoire la majorité des méthodes probabilistes se base sur une unique réalisation du processus stochastique à l'origine des données, c'est à dire sur l'observation d'une seule séquence d'ADN [107, 154], ou d'un groupe de loci liés [154]. Cette limitation n'est fatale ni pour les méthodes d'acceptation-rejet ni pour les méthodes probabilistes exactes (voir par exemple [39, 6]), mais pour ces dernières, le problème est celui de la complexité d'implémentation. Les méthodes des moments ont l'avantage apparent de pouvoir facilement traiter des données multilocus. Ces données multilocus sont toutefois grossièrement compactées, par exemple sous la forme d'un ou de quelques indices de déséquilibre génétique [71].

Enfin, on dispose souvent d'informations sur les valeurs de certains paramètres des modèles. On sait par exemple que le taux de mutation des microsatellites est relativement élevé, quelque part entre 10^{-6} et 10^{-3} par gène et par génération. On souhaiterait donc pouvoir incorporer à la méthode d'inférence des *a priori*, et les incertitudes sur ces *a priori*. Cela est permis (en fait requis) par les méthodes bayésiennes, que leur mise en oeuvre soit exacte [154] ou approchée [107, 39].

On a donc le choix entre une diversité de méthodes d'inférence, qui diffèrent par leur souplesse, par la complexité de leur mise au point et de leur implémentation, par leur gourmandise en temps de calcul et par leur précision. Le choix de l'une ou l'autre des méthodes dépend du problème et de contraintes techniques et humaines. Dispose-t-on du savoir-faire ? Quel temps peut-on allouer au développement de la méthode ? Disposera-t-on ensuite de l'équipement informatique suffisant pour exploiter la méthode ?

Le problème présentement posé est celui d'inférer conjointement les paramètres démographiques et mutationnels d'un modèle de fondation-explosion. Il est vraisemblable, d'après des études sur des modèles proches [46, 6, 39], que la fondation et l'explosion interfèrent, entre elles et avec le processus mutationnel, et que l'information contenue dans un jeu de données est très limitée. Dans ce contexte peu favorable, j'ai opté pour une méthode *probabiliste* —par opposition aux méthodes des moments—, exacte —par opposition aux méthodes probabilistes basées sur le calcul de statistiques sommaires. Afin de maximiser l'information empirique utilisée, j'ai souhaité une méthode pouvant exploiter des *données multilocus*. Face à l'extrême difficulté de

développement et d'implémentation des méthodes probabilistes, il est naturel de partir d'une méthode existante dont les modèles démographique et mutationnel sont plus simples. La méthode de M. Beaumont, implémentée dans le programme MSVAR, était la plus proche de ce que je recherchais [6]. Le chapitre 4 présente une extension de cette méthode, basée comme elle sur de l'échantillonnage de Monte Carlo par Chaîne de Markov. La méthode a l'avantage de ne pas nécessiter de développements mathématiques importants (par rapport à une méthode d'échantillonnage pondéré par exemple) : l'algorithme de Metropolis-Hastings peut en effet être utilisé pour des problèmes extrêmement variés, avec des adaptations relativement simples. La médaille a un revers, l'utilisation délicate de la méthode. Le chapitre 5.4 présente des outils pour l'exploration du modèle de fondation-explosion, et pour l'exploitation des échantillons de la loi *a posteriori* fournis par MSVAR. L'usage de ces outils est illustré sur l'exemple de la vérification de l'implémentation. Le chapitre 6 explore le modèle de fondation-explosion, et discute des améliorations possibles de MSVAR.

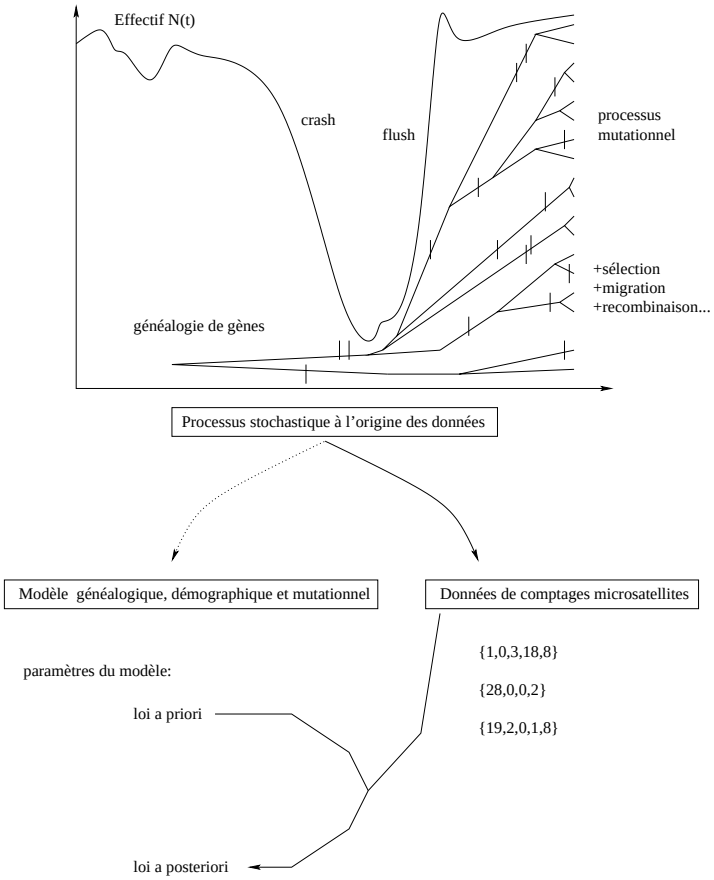


FIG. *3.1: Dans la méthode d'inférence bayésienne présentée dans cette partie, on modélise le processus stochastique à l'origine des données de diversité microsatellite. On incorpore les informations sur ce processus stochastique et les incertitudes sur ces informations, sous la forme de la loi jointe a priori des paramètres du modèle. Les données de diversité observées permettent de préciser les a priori. Les inférences sont exprimées par la loi jointe des paramètres, sachant les données, ou loi* a posteriori *des paramètres.*

Chapitre 4

Principe de la méthode d'inférence bayésienne

On dispose de données empiriques D sous la forme de comptages alléliques à un ou plusieurs loci microsatellites (encart page 78), dans un échantillon d'une population supposée close, et bien décrite par un modèle de Wright-Fisher pour lequel l'hypothèse de stabilité de l'effectif est relâchée, ou par tout autre modèle populationnel pour lequel le processus de coalescence discret est correctement approché par le coalescent standard avec variations d'effectif [99, 132].
On se donne, outre le modèle généalogique du coalescent avec variations d'effectif :

- un modèle démographique pour la population échantillonnée,
- un modèle mutationnel pour les marqueurs microsatellites,
- une loi *a priori* $p(\phi)$ sur les valeurs des paramètres du modèle (démographique et mutationnel).

On souhaite obtenir la loi *a posteriori* des paramètres du modèle, sachant la configuration exacte du jeu de données, et les *a priori*. Cette loi *a posteriori* $p(\phi|D)$ est en principe simplement obtenue selon la formule d'inversion de Bayes :

$$p(\phi|D) = \frac{p(\phi)p(D|\phi)}{p(D)}$$

On définit $p(D)$ par décomposition selon un système complet d'événements pour ϕ

$$p(D) = \int_{\Phi} p(D|\phi)p(\phi)d\phi\,.$$

Le terme $p(D)$ est donc une constante, pour des *a priori* fixés.
Dans la suite, nous travaillerons à cette constante près, et nous considèrerons simplement que $p(\phi|D)$ est proportionnel à $p(\phi)p(D|\phi)$. Cela n'est pas gênant du fait que nous obtiendrons des échantillons tirés selon $p(\phi|D)$, et que la loi *a posteriori* $p(\phi|D)$ sera obtenue par estimation de densité à partir de ces échantillons.

Reste donc à calculer $p(\phi)p(D|\phi)$. Plusieurs options pour l'implémentation de cette solution bayésienne sont possibles (*cf.* 3.3). La méthode utilisée par M. Beaumont [6] et reprise par C. Calmet et M. Beaumont (article en préparation annexé page 185) est basée sur de l'échantillonnage de Monte Carlo par Chaîne de Markov.

Jeux de données microsatellites

Lorsque l'on type un organisme diploïde à un locus microsatellite, on obtient en principe une information sur les deux copies génomiques de ce microsatellite, en termes de longueur de produits de PCR. Le nombre de paires de bases du motif répété étant connu, on peut traduire le polymorphisme éventuel en termes de différences dans le nombre de répétitions du motif. Pour le modèle démographique, généalogique et mutationnel utilisé dans la présente thèse, on ignore l'information génotypique pour considérer la distribution des comptages alléliques dans la population étudiée. Considérons le jeu de données suivant à un locus dinucléotide, représenté au format GENEPOP [112] :

Pop
indiv , 120120
indiv , 120122
indiv , 122122
indiv , 122126
indiv , 120120

La représentation normalisée des comptages alléliques dans ce jeu de données est $\{5, 4, 0, 1\}$, ce qui signifie qu'il comporte 5 copies d'une certaine longueur (produit de PCR de longueur 120, ce qui correspond à un certain nombre de répétitions du motif dinucléotide constitutif), 4 copies d'une unité plus longues (122), et une copie de deux unités plus longue encore (126). La classe allélique 124, non représentée dans l'échantillon, est représentée par un 0 pour que soit codé l'écart de taille entre les classes alléliques. Dans le sens inverse, une configuration de $\{1, 0, 0, 3\}$ sous forme normalisée peut par exemple correspondre à un échantillon de quatre gènes donnant des amplifiats de longueurs 100 et 106 pour un dinucléotide, ou de 203 et 215 pour un tétranucléotide. Ce codage normalisé ne peut rendre compte des comptages alléliques à des loci microsatellites pour lesquels on a des allèles hors échelle (par exemple, à un dinucléotide, des allèles donnant un produit de PCR de parité différente des autres).

Un script perl est disponible sur demande pour le formattage des données pour MSVAR, depuis le format GENEPOP

4.1 Processus stochastique à l'origine des données

On suppose que les généalogies des gènes échantillonnés sont obtenues par réalisation d'un processus stochastique bien décrit par le coalescent standard avec variations d'effectif. Cela définit un modèle généalogique dont la pertinence doit être discutée pour chaque cas d'application (*cf.* chapitres 7 et 8). On précise ci-après le modèle déterministe de variations d'effectif, et le modèle de mutation des marqueurs microsatellites qui ont enregistré ces variations d'effectif.

4.1.1 Modèle démographique

On considère une population close, panmictique, de $N(t)$ génomes haploïdes. La taille de la population à la date $t = 0$ où cette population est échantillonnée est notée N_0. On considère que le temps croît lorsque l'on remonte dans le passé, et on choisit pour unité de temps N_0 générations. $N(t)$ est supposé varier exponentiellement pour atteindre la taille N_1 à une date $t_f > 0$ (f comme fondation) dans le passé. Pour $t > t_f$, la taille de population ancestrale est supposée constante de valeur N_2. N_0, N_1 et N_2 peuvent être ordonnés arbitrairement, si bien que des variations d'effectif variées peuvent être décrites avec le même formalisme, avec ou sans discontinuité d'effectif à t_f. Le cas où $N_2 > N_1 < N_0$ correspond à une histoire démographique de type fondation-explosion (illustration figure 4.1), qui nous intéresse principalement. Si $N_2 > N_1 > N_0$, on a un crash-déclin, si $N_2 < N_1 < N_0$ un pic-explosion et enfin si $N_2 < N_1 > N_0$ un pic-déclin.

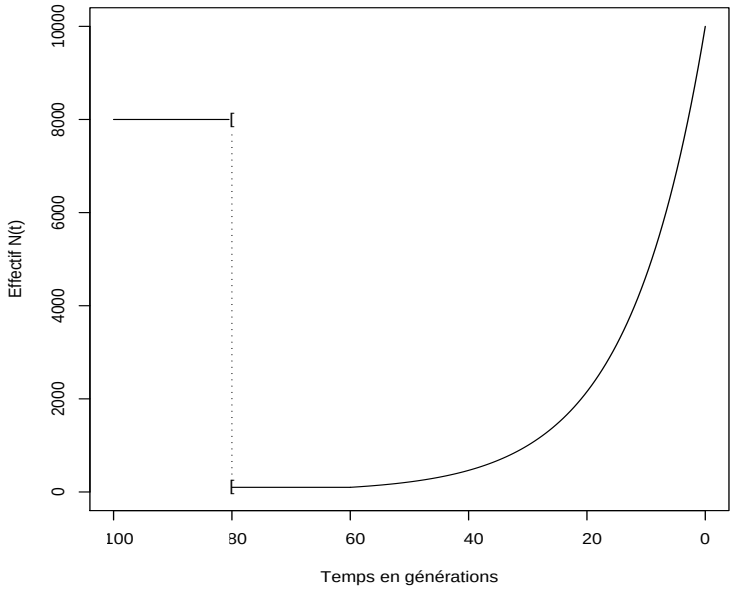


FIG. *4.1: Modèle démographique de fondation-explosion. Une population de taille jusque-là constante* $N_2 = 8\,000$ *gènes tombe brutalement à* $N_1 = 100$ *gènes,* 80 *générations avant d'être échantillonnée. La reprise démographique est immédiate et exponentielle et la population atteint l'effectif de* $N_0 = 10\,000$ *gènes à la date d'échantillonnage. Le modèle autorise n'importe quel ordre pour* N_0, N_1 *et* N_2, *et peut donc décrire des histoires démographiques variées. La seule contrainte est une variation d'effectif brutale suivie immédiatement d'une variation exponentielle.*

Pour des raisons pratiques —le coalescent standard ne permet d'inférer que certaines combinaisons des paramètres naturels—, on exprime l'effectif $N(t)$ relativement à la taille N_0, par $v(t) = N(t)/N_0$, où t est supposé croître quand on remonte dans le passé. Avec $r = N_0/N_1$ et $a = N_0/N_2$, on obtient

$$v(t) = \begin{cases} r^{-\frac{t}{t_f}} & \text{si } 0 \leq t \leq t_f, \\ \frac{1}{a} & \text{si } t > t_f. \end{cases}$$

On remarquera que ce modèle démographique n'autorise ni un délai entre la fondation et la reprise démographique, ni une stabilisation de l'effectif avant l'échantillonnage, ce qui peut limiter le nombre de cas d'application. Une variation d'effectif linéaire entre t_f et t_0 est également implémentée dans MSVAR. Cette possibilité n'a pas été explorée à l'heure actuelle et ne sera pas décrite dans le texte en français. Pour une description succinte, on se reportera à la publication correspondante (annexe page 185).

4.1.2 Modèle mutationnel pour les microsatellites

Le modèle mutationnel considéré est un modèle à deux phases (TPM) [31]. Soit μ la probabilité de mutation d'une copie physique d'un marqueur microsatellite, à chaque génération. Sous TPM, sachant qu'une mutation se produit, elle a probabilité p d'être une mutation "pas" (amplitude $a = 1$, perte ou gain d'une unité de répétition), et $(1 - p)$ d'être une mutation "saut", d'amplitude définie par une loi géométrique de moyenne $j \geq 1$ (encart page 39). Pour récapituler, la probabilité $p(a)$ pour qu'une mutation soit d'amplitude $a \geq 1$ est donnée par

$$p(a) = \begin{cases} p + \frac{1}{j}(1-p) & \text{si } a = 1, \\ (1-p)\frac{1}{j}\left(1 - \frac{1}{j}\right)^{a-1} & \text{si } a > 1. \end{cases} \tag{4.1}$$

Le modèle TPM dégénère en SMM si $p = 1$ et/ou $j = 1$, et en GSM si $p = 0$ (composante géométrique seulement). L'encart page 81 illustre la distribution d'amplitude pour quelques exemples de valeurs des paramètres du modèle TPM. Nous n'étudierons à aucun moment les effets de biais directionnels, de contraintes de taille, ou de taille-dépendance. La possibilité d'inclure un biais mutationnel est implémentée dans MSVAR (elle l'était déjà dans la version que m'a communiquée M. Beaumont). Les deux autres types d'extensions du modèle mutationnel ne modifieraient que le calcul de vraisemblance, ce qui les rend tout à fait envisageables avec des modifications localisées de l'implémentation.

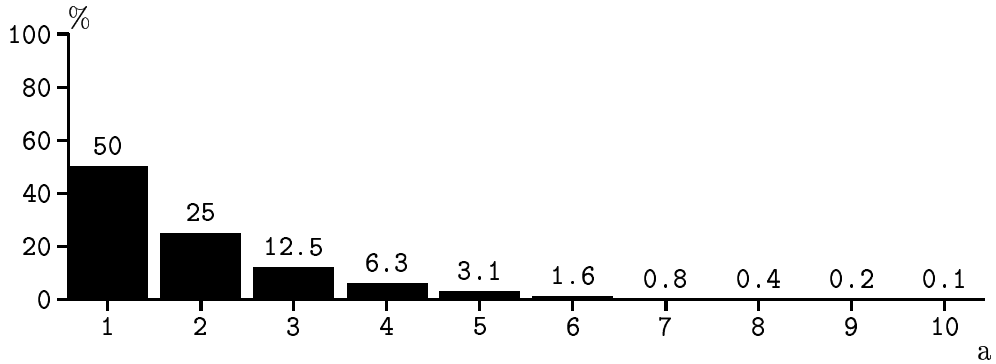
4.2 Calcul de la vraisemblance d'une généalogie

Pour échantillonner selon la loi *a posteriori* $p(\phi|D)$ par l'algorithme de Metropolis-Hastings, on a besoin de calculer la vraisemblance $p(D|\mathcal{H}, \phi)$, où $\mathcal{H}$ est une histoire généalogique et mutationnelle, entièrement spécifiée par les relations de parenté entre gènes (topologie de la généalogie), par la position et le type des mutations, par les dates des coalescences et des mutations, et par la classe allélique du MRCA. Les propriétés du coalescent standard (et de ses avatars avec variations d'effectif) permettent en fait de simplifier $\mathcal{H}$.

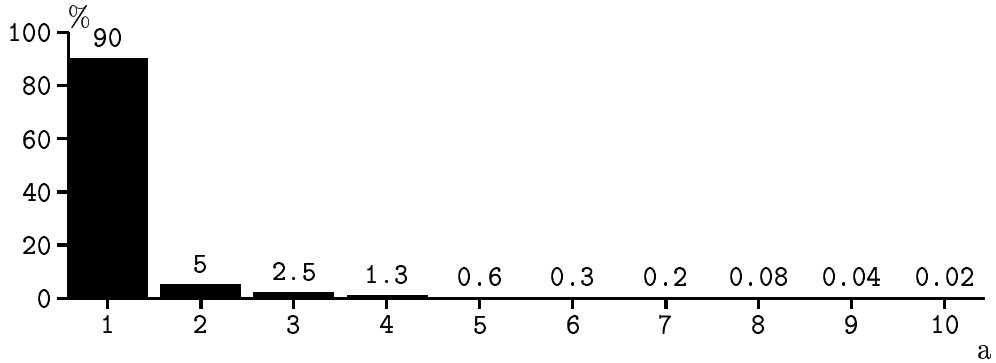
Distribution de l'amplitude des mutations sous TPM

On appelle amplitude d'une mutation le nombre de répétitions du motif microsatellite qui est perdu ou ajouté à l'allèle qui mute. La distribution de l'amplitude des mutations selon un modèle à deux phases (TPM) est déterminée par deux paramètres : un paramètre p qui détermine la proportion de la phase SMM (l'autre étant la phase géométrique), et un paramètre j qui correspond à la moyenne de l'amplitude des mutations, pour la phase géométrique. D'autres auteurs préfèrent utiliser comme paramètre de TPM la variance de l'amplitude pour la phase géométrique. Précisons pour faciliter la comparaison entre méthodes que cette variance s'exprime $j(j-1)$. Les mutations d'amplitude tirée dans la phase géométrique peuvent avoir pour amplitude 1, avec probabilité $1/j$. La proportion totale s de mutations d'amplitude $a > 1$ n'est donc pas égale à $1-p$, mais à $s = 1-p-(1-p)/j$. La moyenne d'amplitude m des mutations est inférieure à la moyenne j pour la phase géométrique, et vaut $m = p + (1-p)j$. Le couple (s, m) suffit à caractériser un modèle TPM, et la visualisation de la distribution correspondant à un couple est plus aisée qu'avec (p, j) :

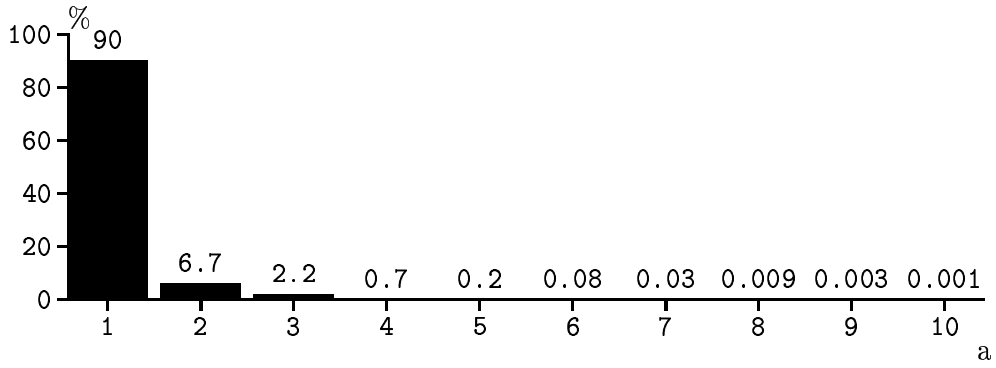
Distribution d'amplitude pour un modèle GSM avec $p = 0$ *et* $j = 2$ *soit* $(s, m) = (0.5, 2)$



Distribution d'amplitude pour un modèle TPM avec $p = 0.8$ *et* $j = 2$ *soit* $(s, m) = (0.1, 1.2)$



Distribution d'amplitude pour un modèle TSM avec $p = 0.7$ *et* $j = 1.5$ *soit* $(s, m) = (0.1, 1.15)$



4.2.1 Notion de séquence d'événements

Étant fixée la taille n_0 de l'échantillon, la probabilité de tirage d'une généalogie donnée, selon le coalescent standard, et pour un taux de mutation constant, ne dépend que des dates des événements de coalescence et de mutation successifs. En effet, dans le coalescent standard, toutes les lignées ont équiprobabilité de coalescer et d'être touchées par des mutations. On peut donc remplacer la structure arborescente des généalogies, et la position des mutations sur les branches, par une liste d'événements, chacun caractérisé par une date. Autrement dit, on projette la généalogie dont on veut calculer la probabilité sur l'axe des temps, et on la réduit à la séquence $\mathcal{E}$ des événements qui la caractérisent (figure 4.2).

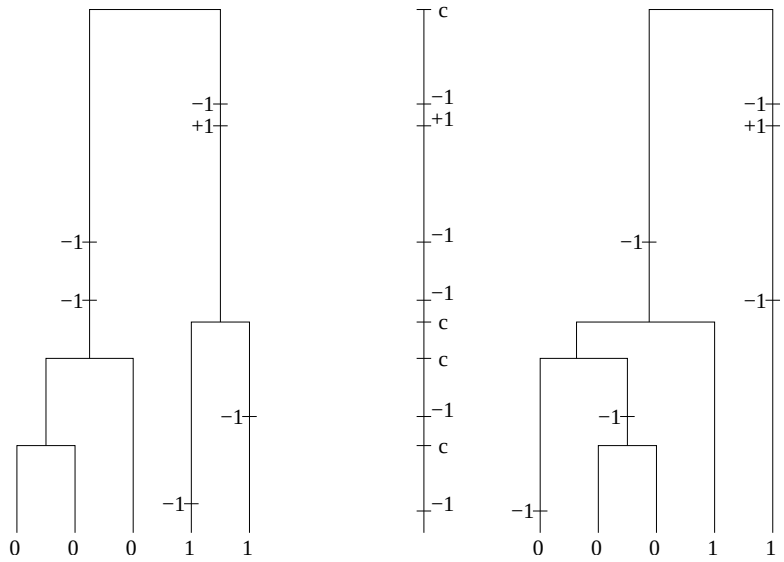


FIG. *4.2: La vraisemblance d'une histoire généalogique et mutationnelle compatible avec l'échantillon ne dépend que des temps d'attente entre événements successifs, indépendamment de la topologie exacte. Une histoire peut donc être projetée sur l'axe des temps, et remplacée par une séquence d'événements, pour le calcul de vraisemblance. Les deux histoires représentées correspondent à une même séquence d'événements* $\{-1, c, -1, c, c, -1, -1, +1, -1, c\}$*, où* c *représente les coalescences et où* $+/-i$ *désigne une mutation d'amplitude* i *telle que le gène mutant soit plus long/court que le gène parent. Dans les modèles mutationnels symétriques, on peut même ignorer le sens des mutations.*

4.2.2 Loi du temps d'attente entre deux événements

Dans le coalescent standard, lorsque les mutations sont neutres, il y a indépendance entre l'occurrence de coalescences et de mutations. Les deux types d'événements se produisent alors respectivement dans l'histoire d'un échantillon selon des taux $c(t)$ et $m(t)$. Le taux d'occurrence des coalescences est proportionnel au nombre $n_i(n_i-1)/2$ de paires de lignées pouvant coalescer, et inversement proportionnel à la taille de la population à l'instant t. Le taux d'occurrence des mutations est proportionnel au nombre n_i de lignées pouvant muter, et à la densité de probabilité de mutation pour une lignée. Si l'on appelle $\gamma(t)$ le taux des événements (coalescences et mutations confondus), on a [55] :

$$\gamma(t) = c(t) + m(t) = \frac{n_i(n_i-1)}{2 \quad v(t)} + \frac{n_i\theta}{2} \quad , \tag{4.2}$$

le temps étant mesuré en unités de N_0 générations. Le paramètre $\theta = 2N_0\mu$ (où μ est la probabilité de mutation de chaque gène à chaque génération) est appelé probabilité de mutation renormalisée. L'additivité des taux est due à la loi exponentielle des taux d'ettente, et à l'indépendance des événements de coalescence et de mutation dans le cas neutre qui nous intéresse. Remarquons qu'il est tout à fait équivalent de simuler les dates des événements (i) en tirant un temps d'attente selon le taux $\gamma(t)$, et en décidant si l'événement est une coalescence ou une mutation proportionnellement aux valeurs de $c(t)$ et de $m(t)$, ou bien (ii) de tirer indépendamment deux temps d'attente selon les taux $c(t)$ et $m(t)$, et de choisir le temps d'attente le plus court des deux, et son type.

Si le taux γ était constant entre deux événements, cela correspondrait tout simplement à un temps d'attente de loi exponentielle de paramètre γ (encarts page 38 et suivantes). Du fait des variations d'effectif selon $v(t)$, $\gamma(t)$ varie entre deux événements. La loi de la date t_{i+1} d'un nouvel événement, sachant que l'événement précédent s'est produit à t_i est donnée [55] par

$$p(t_{i+1}|t_i) = \gamma(t_{i+1}) \ \exp\left(-\int_{t_i}^{t_{i+1}} \gamma(t)dt\right). \tag{4.3}$$

Cela correspond à la densité d'occurrence de l'événement à la date t_{i+1}, multipliée par la probabilité pour qu'aucun événement n'ait eu lieu entre les dates t_i et t_{i+1}. Sachant qu'un événement s'est produit à l'instant t_{i+1}, il a probabilité $c(t)/\gamma(t)$ d'être une coalescence et probabilité $m(t)/\gamma(t)$ d'être une mutation.

De l'expression de $p(t_{i+1}|t_i)$, de la probabilité pour qu'un événement soit une coalescence ou une mutation et du modèle mutationnel considéré, on déduit la probabilité conditionnelle $p(\{e, t_{i+1}\}|t_i)$ d'occurrence d'un événement particulier e à la date t_{i+1} sachant la date t_i de l'événement précédent :

- Si cet événement est une coalescence c entre deux lignées particulières parmi les $n_i(n_i-1)/2$ paires possibles,
$$p(\{c, t_{i+1}\}|t_i) = \frac{\exp\left(-\int_{t_i}^{t_{i+1}} \gamma(t)dt\right)}{v(t)}.$$
- Si l'événement est une mutation m d'amplitude a sur une lignée particulière, dans une direction particulière,
$$p(\{m_{a=1}, t_{i+1}\}|t_i) = \frac{\theta}{4}\exp\left(-\int_{t_i}^{t_{i+1}} \gamma(u)du\right)\left[p + \frac{1}{j}(1-p)\right]$$
si $a = 1$, et si $a > 1$,
$$p(\{m_{a>1}, t_{i+1}\}|t_i) = \frac{\theta}{4}\exp\left(-\int_{t_i}^{t_{i+1}} \gamma(u)du\right)\left[(1-p)\left(1-\frac{1}{j}\right)^{a-1}\frac{1}{j}\right]$$

Des probabilités conditionnelles de ces deux événements élémentaires, on peut déduire celle d'événements composés (par exemple, celle d'une coalescence entre n'importe quelle paire de lignées, ou celle d'un quelconque événement de coalescence ou de mutation) Ces expressions peuvent être intégrées explicitement pour le modèle considéré. L'appendice A de l'article annexé page 185 donne les formes intégrées de $p(\{e, t_{i+1}\}|t_i)$ pour ce modèle . Le texte en français donne les formes intégrées pour un modèle démographique plus complexe (paragraphe 6.3.2) qui se réduit au présent modèle pour des valeurs particulières des paramètres.

4.2.3 Vraisemblance d'une séquence d'événements

Les temps d'attente et les types des événements successifs sont indépendants. Pour être plus précis, ils sont conditionnellement indépendants : en effet, le taux de coalescence et le taux d'occurrence des mutations dépendent tout de même du nombre de lignées non coalescées (donc du nombre de coalescence déjà réalisées), et des variations d'effectif depuis le dernier événement (donc de la date de ce dernier événement). La vraisemblance d'une séquence d'événements est donc le produit des densités élémentaires des événements et de la probabilité de l'état ancestral. Dans le modèle considéré, tous les états ancestraux sont équiprobables : les configurations sont considérées comme identiques par translation. Comme la vraisemblance n'intervient dans la méthode présentée —basée sur l'algorithme de Metropolis-Hastings— que sous forme de rapports (ou plutôt de différences entre des log-vraisemblances), on peut travailler à une constante multiplicative près, et ignorer le terme correspondant à la probabilité de l'état ancestral. La vraisemblance $L(\mathcal{E})$ d'une séquence d'événements est alors proportionnelle au seul produit des $p(\{e_{i+1}, t_{i+1}\}|t_i)$. Si n_0 est la taille de l'échantillon, D la configuration du jeu de données, $\mathcal{E}$ la séquence des événements et ϕ la valeur du vecteur de paramètres, on a donc :

$$L(\mathcal{E}) = p(D, \mathcal{E}|n_0, \phi) \propto \prod_{i=0}^{card(\mathcal{E})-1} p(\{e_{i+1}, t_{i+1}\}|t_i) \tag{4.4}$$

4.3 Échantillonnage MCMC selon la loi a posteriori $p(\phi|D)$

L'échantillonnage MCMC selon la loi *a posteriori* $p(\phi|D) \propto p(D|\phi)p(\phi)$ comporte plusieurs étapes. La première est la définition d'un état de départ dans l'espace d'échantillonnage, ici l'espace des paramètres et des généalogies. La seconde est le choix d'un noyau de proposition, qui permette d'explorer l'espace d'échantillonnage, et assure l'irréductibilité de la chaîne de Markov. La troisième est le calcul de la probabilité d'acceptation des propositions, qui permet de définir les états successifs de la chaîne de Markov. Nous allons donner des éléments de compréhension du protocole adopté pour chacune de ces trois étapes. La seconde, l'exploration de l'espace d'échantillonnage, détermine crucialement la corrélation entre états successifs de la chaîne de Markov —ou autocorrélation de la chaîne— et donc la rapidité des simulations.

4.3.1 Construction d'un premier arbre compatible avec les données

L'échantillonnage selon la loi *a posteriori* $p(\phi|D)$, par MCMC, nécessite de partir d'un état dans l'espace probabilisé des paramètres et des généalogies. L'utilisateur fournit des valeurs initiales pour les paramètres démographiques et mutationnels N_0, N_1, N_2, t_a, μ, p et j (en réalité, des combinaisons de ces paramètres naturels). En revanche, il ne donne pas de généalogie de départ, mais seulement des données de comptage allélique à des loci microsatellites (supposons un seul locus pour simplifier). Il faut donc construire une généalogie compatible avec le jeu de données, pour pouvoir mettre en oeuvre l'algorithme de Metropolis-Hastings. Un bon protocole de construction de la première généalogie doit assurer de trouver un MRCA, et de partir d'une généalogie qui aurait pu, avec une probabilité aussi grande que possible, être tirée selon $p(D|\phi)$.

L'algorithme de la procédure de construction du premier arbre (repris sans modification de [6]) permet de tirer les dates des coalescences et des mutations approximativement selon $c(t)$ et $m(t)$. Il répète les opérations suivantes jusqu'à trouver un MRCA pour l'échantillon :

1. Tirage d'une date de coalescence conformément à $c(t)$.

2. Tirage d'une date de mutation conformément à $m(t)$ (loi exponentielle).

3. Comparaison des deux dates, et choix de la plus récente, et de son type.

4. Si l'événement choisi est une mutation, choisir uniformément un des gènes patriarches[1], faire muter le gène, en direction de la moyenne de taille parmi les gènes patriarches.

5. Si l'événement choisi est une coalescence, choisir deux lignées de même état allélique, les faire coalescer ; s'il n'existe pas de telle paire, faire à la place de la coalescence une mutation (il faut être pragmatique).

Ce schéma de construction du premier arbre est loin d'être idéal, mais il assure de trouver un MRCA, et essaie de tirer les temps d'attente selon leur distribution théorique. Le protocole de construction de la chaîne de Markov se chargera de modifier cet arbre. Pour des jeux de données monolocus, la configuration initiale n'influence que très peu le devenir de la chaîne et sa vitesse de convergence. Pour des jeux de données multilocus, la MCMC peut mettre longtemps à se sevrer de la configuration initiale, et le biais d'initialisation doit être supprimé en tronquant le début de l'échantillon. Mais lorsque des problèmes de biais d'initialisation se posent, cela va généralement de pair avec des problèmes d'autocorrélation (figure 5.1 page 99).

4.3.2 Exploration de l'espace des généalogies et des paramètres

La matrice de proposition utilisée doit permettre l'exploration de l'ensemble de l'espace des paramètres Φ et de l'espace des généalogies, ainsi que le calcul de la probabilité d'acceptation et de rejet de la proposition, conformément à l'algorithme de Metropolis-Hastings. Le protocole choisi, de type 'random-sweep' [15] permet un déplacement désordonné dans l'espace d'échantillonnage. Dans ce protocole, on se donne la probabilité de différents types de modifications, et à chaque itération, un des types de modifications est tiré, indépendamment du type choisi à l'itération précédente. Les probabilités ont été choisies par essais-erreurs afin de maximiser le taux d'acceptation et de permettre un déplacement aussi rapide que possible dans l'espace des généalogies. Elles sont données ci-dessous entre parenthèses :

(0.99) Modification de la généalogie d'un locus (nombre, ordre des événements).

(0.0095) Modification de certains paramètres dans ϕ, et avec probabilité 0.5, des dates T :
(0.4) θ, p, j, r, a, t_f ;
(0.05) θ ; (0.025) p ; (0.025) j ; (0.1) θ, p, j ;
(0.05) r ; (0.05) a ; (0.05) t_f ; (0.1) r, a, t_f ;
(0.05) θ, r ; (0.05) θ, a ; (0.05) θ, t_f.

(0.0005) Modification des dates des événements, T.

La probabilité de modifier seulement θ et t_f est par exemple égale à 0.0095×0.05 (et avec probabilité 0.5, on modifie en même temps les dates des événements).

[1] Les gènes patriarches sont les gènes ancêtres de l'échantillon localisés aux extrémités pas encore coalescées de la généalogie.

Modifications des paramètres et des dates des événements

Pour se déplacer dans l'espace Φ, on modifie la valeur de certains paramètres par une déviation aléatoire de loi donnée. Les modalités détaillées des modifications ont été choisies par essais-erreurs au vu des résultats de la méthode. Les paramètres r, a, t_f et μ peuvent *a priori* varier de zéro (exclu) à $+\infty$, et on s'interesse à leur ordre de grandeur : on n'espère pas savoir si une population a été fondée il y a 100 ou 101 générations, mais plutôt départager entre 10, 100 ou 1000 générations. Pour cette raison, on reporte aux points d'échantillonnage successifs le logarithme décimal des paramètres. Afin d'explorer rapidement l'intervalle *a priori* pour le logarithme décimal, on le modifie par ajout d'un tirage aléatoire dans une loi normale centrée, d'écart-type σ. On constate empiriquement que la convergence de la MCMC est meilleure si l'on réduit cet écart-type lorsque le nombre ℓ de loci augmente. M. Beaumont a observé que l'autocorrélation de la MCMC est sensiblement diminuée si σ est augmenté lorsque t_f est le seul paramètre modifié. La synthèse des observations empiriques conduit à proposer $\sigma = 1/\sqrt(\ell)$, et $\sigma = 5/\sqrt(\ell)$ lorsque seul t_f est modifié. Lorsque plusieurs paramètres sont modifiés conjointement, la déviation est la même en valeur absolue pour tous les paramètres, mais afin d'augmenter le taux d'acceptation, son signe est inversé pour μ (car μ est inversement corrélé aux autres paramètres). Le paramètre j du modèle TPM est forcément supérieur ou égal à 1, mais on reporte aussi son logarithme décimal, et on modifie ce paramètre comme les précédents. Le paramètre p qui spécifie la fraction de SMM a un statut un peu différent : il varie potentiellement dans $[0; 1]$, et les bornes —incluses— correspondent au SMM strict et au GSM. On a donc choisi de reporter la valeur naturelle de p, et d'explorer l'intervalle $[0; 1]$ comme on explore l'intervalle *a priori* pour le logarithme décimal des autres paramètres. Autrement dit, on modifie p (et non $\log_{10}(p)$) en lui ajoutant un tirage selon une loi normale centrée. Un écart-type de 0.05 permet une bonne exploration de $[0; 1]$ par p. Lorsque les paramètres modifiés sont hors de l'intervalle de densité *a priori* non nulle, la proposition sera bien sûr rejetée —cela n'est pas satisfaisant, et une version de MSVAR en développement intègrera des lois *a priori* et des modalités d'exploration qui éviterons de telles troncatures.

Pour les modifications de T, de nouvelles dates sont tirées conformément aux valeurs des taux de coalescence et de mutation $c(t)$ et $m(t)$, calculées d'après l'état de ϕ, comme expliqué dans le détail à la section 5.1, au sujet de la simulation de jeux de données.

Modifications des généalogies de gènes

Le choix de la matrice de proposition pour la modification des généalogies de gènes détermine de façon cruciale si l'unique distribution invariante de la MCMC est $p(\phi|D)$, comme souhaité, et donc si, sous réserve de convergence, les états successifs de la chaîne seront des tirages corrélés selon $p(\phi|D)$. Ce choix est du domaine technique, et n'a aucun fondement biologique. La chaîne de Markov spécifiée doit être irréductible, c'est à dire que tout point de l'espace des paramètres doit pouvoir être atteint en un nombre fini de pas, depuis n'importe quel autre point. Une contrainte majeure est qu'il faut être capable de calculer la probabilité P_f de proposition des modifications, et de calculer la probabilité P_r de la proposition inverse. Ces conditions sont remplies pour des modèles de mutation des microsatellites autorisant des sauts (et donc en particulier pour le modèle TPM présentement considéré), en rendant possibles des modifications des 8 types suivants (figure 4.3). Les types 1, 2, 3, 4, 7 et 8 étaient déjà utilisés dans [6]. Seuls les types 5 et 6 ont été nouvellement implémentés, afin que puissent être explorées les histoires mutationnelles comportant des sauts, tels qu'autorisés par le modèle TPM. Le type 7 n'est

pas strictement nécessaire à l'irréductibilité de la chaîne de Markov, mais diminue fortement son autocorrélation. Bien entendu, les modifications donnent des histoires mutationnelle et généalogique qui demeurent compatibles avec l'échantillon :

1/2 : addition/retrait de 2 mutations pas (amplitude 1) en sens contraire, sur une lignée.
3/4 : addition/retrait de trois mutations s'annulant autour d'un noeud.
5/6 : fusion/fission de deux/une mutation(s) en une/deux.
7 : échange de l'ordre de deux événements successifs dans une séquence d'événements.
8 : échange des lignées ancestrales de deux événements successifs de même état allélique.

La généalogie qui illustre les types de modifications (figure 4.3) a bien sûr été choisie pour pouvoir se prêter à tous. Mais ce n'est pas le cas général : alors que les ajouts de mutations (types 1 et 3) sont toujours possibles, quelle que soit la généalogie de départ, les autres types de modifications ne sont pas toujours envisageables. Par exemple, on ne peut pas mettre en oeuvre le type 2 s'il n'existe aucune lignée comportant deux mutations pas en sens contraire. On ne peut pas appliquer le type 4 si aucun noeud n'est entouré de trois lignées ayant des triplets annulables de mutations pas, soit respectivement des mutations +1, -1, -1 ou bien -1, +1, +1 pour les lignées mère, et les deux lignées filles. On ne peut pas appliquer le type 5 s'il n'existe pas de lignée ayant une paire de mutations fusionnables, ni le type 6 s'il n'existe aucune mutation d'amplitude $a > 1$ dans la généalogie. On ne peut pas appliquer le type 8, si par exemple il n'existe pas de paires d'événements successifs qui soient de même état allélique (il y a d'autres contraintes), ni le type 7 s'il n'y a pas de paires d'événements successifs autres que "père et fils dont au moins une coalescence". De ce fait, on se donne un jeu fixé de poids pour les 8 types de modifications : $(W_1, \ldots, W_8) = (0.15, 0.15, 0.15, 0.15, 0.1, 0.1, 0.1, 0.1)$. L'état de la généalogie à modifier détermine un jeu de valeurs binaires $(B_1, \ldots, B_8)$ indiquant pour chaque type s'il est possible ($B_i = 1$) ou pas ($B_i = 0$). La probabilité de proposer une modification de type i est alors donnée par

$$P_i = B_i U_i / \sum_j B_j U_j.$$

Dans chaque classe, le calcul de la probabilité de la proposition directe[2] P_f et de la proposition inverse[3] P_r doit correspondre parfaitement à ce que fait l'algorithme (par exemple, si l'on calcule P_f pour le type 1 en supposant que n'importe quelle paire de mutations pas s'annulant peut être choisie avec même probabilité sur une lignée, il faut s'assurer que le programme permet bien de choisir n'importe quelle paire de façon équiprobable). Les détails des modalités d'exploration des généalogies présentés ci-après sont une composante importante de l'algorithme, qui déterminent son efficacité :

Ajout ou retrait d'une paire de mutations pas, sur une lignée : les types de modifications 1 et 2 sont l'inverse l'un de l'autre. Si P_f est calculé selon le type 1, P_r le sera selon le type 2 et inversement.

Une modification de type 1 est spécifiée par le choix d'une lignée parmi les $n_\ell = 2(n_c - 1)$ lignées possibles —où n_c est le nombre de coalescences), et par le choix de dates pour les deux mutations pas à ajouter. On gagne à pondérer le choix des lignées, non pas uniformément (probabilité $1/n_\ell$ d'être choisie pour chaque lignée), mais proportionnellement au carré de leur durée δt. En effet, cela permet d'ajouter préférentiellement des mutations sur les lignées les plus

[2] P_f pour forward
[3] P_r pour reverse

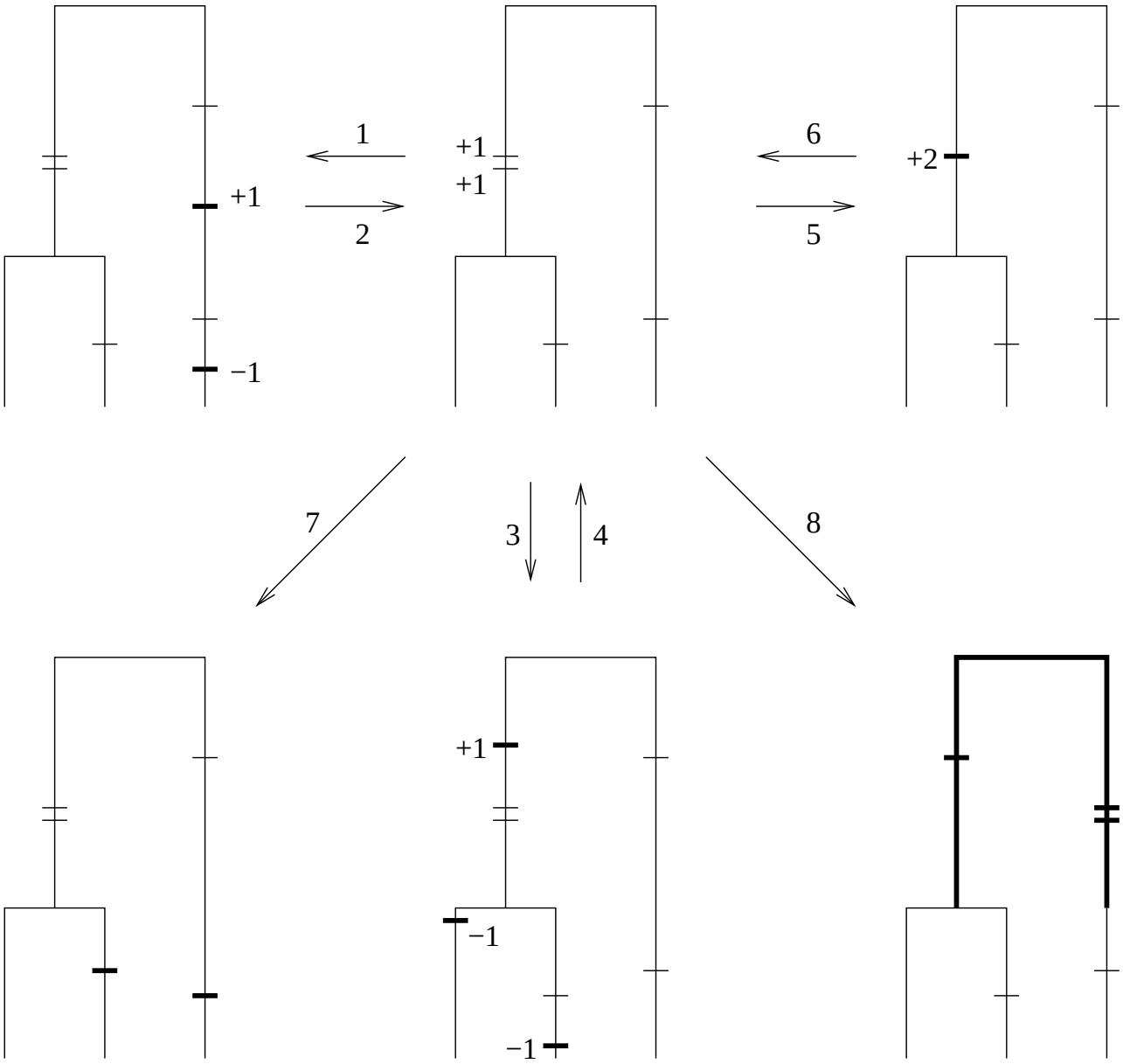


FIG. *4.3: Modifications élémentaires des histoires généalogique et mutationnelle d'un échantillon monolocus. L'arbre choisi peut se prêter à chacun des 8 types de modifications implémentés : ajout et retrait de deux mutations s'annulant sur une lignée (types 1 et 2), ajout et retrait d'un triplet de mutations autour d'un événement de coalescence (types 3 et 4), fusion de mutations de même direction (type 5) et fission de mutations sauts (type 6), échange des dates de deux événements successifs sur des lignées différentes ou de deux mutations successives sur une même lignée (type 7) et échange des lignées ancestrales de deux événements successifs de même état allélique (type 8). Les 8 types sont suffisants pour l'exploration de l'espace des histoires possibles de l'échantillon. Le type 7 est le seul qui ne soit pas nécessaire, mais il améliore sensiblement la vitesse de convergence MCMC.*

susceptibles de les accueillir, et d'augmenter la probabilité d'acceptation de la proposition. Les dates des deux mutations sont tirées uniformément et indépendamment sur la lignée choisie, événement de densité $1/(\delta t)^2$. Le bilan est que la probabilité d'une modification particulière de type 1 est

$$P_f = \frac{P_1}{\sum_{j=1}^{n_\ell} (\delta t_j)^2}.$$

Une modification de type 2 est spécifiée par le choix d'une lignée parmi les $n_\ell > 0$ lignées pouvant se prêter à annulation de deux mutations pas, et par le choix d'une paire de telles mutations. On choisit la lignées uniformément parmi celles qui comptent au moins une mutation $+1$ et une mutation -1, et on choisit la paire de mutations parmi les $n_p > 0$ paires possibles sur cette lignée. Le nombre de paires n_p le produit du nombre de mutations $+1$ et du nombre de mutations -1 dans la lignée. La probabilité d'une modification particulière de type 2 est donc donnée par

$$P_f = \frac{P_2}{n_\ell n_p}.$$

Dans les deux cas, suite à modification, les classes alléliques entre les mutations ajoutées doivent être mises à jour. Les états aux noeuds qui encadrent la lignée modifiée sont inchangés.

Ajout ou retrait d'un triplet de mutations pas, autour d'un noeud : ces deux types de modifications (types 3 et 4) sont l'inverse l'un de l'autre. Si P_f est calculé selon le type 3, P_r le sera selon le type 4 et inversement.

Une modification de type 3 est spécifiée par le choix d'un noeud parmi les n_c que comporte l'histoire, par le choix des dates pour les trois mutations pas à ajouter, et par le choix du sens des mutations (remarquons que quand le noeud est le MRCA, seules les deux lignées filles sont en fait concernées). Sur le même principe que pour le type 1 (et pour les mêmes raisons), on pondère le choix des noeuds, proportionnellement au produit des durées des trois lignées encadrantes (appelées m, g, d comme mère, fille gauche et fille droite). Les dates des mutations sont tirées uniformément sur chacune de ces trois lignées. Les alternatives d'ajouter $+1$, -1 et -1 ou d'ajouter -1, $+1$ et $+1$ sont équiprobables. Le bilan est que la probabilité d'une modification particulière de type 3 est

$$P_f = \frac{P_3}{2 \sum_{j=1}^{n_c} \delta t_m \, \delta t_g \, \delta t_d}.$$

Une modification de type 4 est caractérisée par le choix d'un noeud parmi les $n_n > 0$ pouvant se prêter à annulation de trois mutations pas, et par le choix d'un triplet de telles mutations parmi $n_t > 0$ triplets. Un noeud peut être choisi pour peu que la lignée ancestrale comporte au moins une mutation $+1$ et ses deux lignées filles au moins une mutation -1 chacune, ou le contraire. Le nombre n_t de triplets est donné par le produit du nombre de mutations $+1$ dans la lignée mère et des nombres de mutations -1 dans chacune des deux lignées filles, plus le produit des nombres de mutations pour la configuration -1, $+1$, $+1$. La probabilité d'une modification particulière de type 4 est donc donnée par

$$P_f = \frac{P_4}{n_n n_t}.$$

Dans les deux cas, suite à modification, les classes alléliques entre les mutations ajoutées doivent être mises à jour. L'état au noeud est incrémenté ou décrémenté de 1.

Fusion ou fission de mutations existantes sur une lignée : ces deux types de modifications vont eux aussi par paire, l'un étant l'inverse de l'autre.

Une modification de type 5 est spécifiée par le choix uniforme d'une lignée parmi les n_f lignées pouvant se prêter à fusion de mutation, par le choix uniforme d'une paire de mutations parmi les n_p possibles, et par le choix de la date de la mutation obtenue par fusion, laquelle a même direction que les mutations fusionnées, et amplitude la somme de leurs amplitudes. Une lignée peut donner lieu à fusion de mutations si elle comporte au moins deux mutations de même direction, dont l'une doit être un pas. On aurait aussi bien pu autoriser la fusion de sauts entre eux, mais ce ne serait pas très utile, car les généalogies vraisemblables pour un modèle mutationnel raisonnable comportent peu de mutations sauts, ce qui rend peu probable qu'il y en ait plusieurs sur une même lignée. En tout cas, dans l'implémentation actuelle, on peut fusionner +3 et +1 en +4, ou alors -1 et -1 en -2, mais pas -2 et -2 en -4. Le calcul de n_p est donc simplement la somme des nombres de quatre types différents de paires : des paires +1 +1, des paires -1 -1, des paires +1 +a et des paires -1 -a, avec $a > 1$. La mutation fusion est placée de façon équiprobable à l'une des deux dates des mutations fusionnées, ce qui donne la probabilité de proposition suivante :

$$P_f = \frac{P_5}{2 n_f \, n_p}.$$

Une modification de type 6 est caractérisée par les choix uniformes d'une lignée parmi les n_s lignées comportant au moins une mutation saut (scindable) et d'une mutation saut parmi les n_j présentes sur cette lignée. On décrémente l'amplitude de la mutation saut de 1 en conservant sa direction, on crée une mutation pas de même direction (l'amplitude totale est donc inchangée), on tire une date uniformément sur la lignée de longueur δt_i, et on y place l'une des deux mutations résultant de la fission (avec même probabilité), l'autre mutation restant à la date initiale. Une subtilité peut échapper : si la mutation à scinder est d'amplitude $a = 2$, on n'a qu'une configuration possible, une fois choisie la date de la mutation à ajouter. Si l'amplitude est $a > 2$, on a un saut et un pas à placer, et donc deux configurations équiprobables pouvant être distinguées. De ce fait, la probabilité de proposition est, avec $C = 1$ si $a = 2$, et $C = 0.5$ si $a > 2$:

$$P_f = \frac{P_6}{n_s n_j \delta t_i} \times C.$$

Échange de l'ordre de deux événements temporellement adjacents : on s'autorise à intervertir l'ordre de deux événements temporellement adjacents, à la condition que ces deux événements ne soient pas ancêtre et descendant l'un de l'autre ou, s'ils le sont, que ce soient deux mutations. Cette restriction est nécessaire pour que ce type de modification (type 7) donne des topologies cohérentes. Un événement parmi les n_e pouvant échanger leur date avec l'événement juste plus ancien dans la séquence est choisi uniformément. Les deux dates sont échangées, ce qui donne à une modification du type 7 la probabilité

$$P_f = \frac{P_7}{n_e}.$$

La probabilité de proposer la modification retour est également calculée selon le type 7, qui est son propre inverse.

Échange des lignées ancestrales de deux événements temporellement adjacents : on modifie la topologie des histoires généalogiques par échange des lignées ancestrales de deux événements temporellement adjacents. Pour cela, on choisit uniformément un événement parmi les n_i qui ont même classe allélique que l'événement juste plus ancien dans la séquence, lequel ne doit pas être leur père. Chacun des deux événements est alors branché, avec son sous-arbre descendant, à l'ancêtre de l'autre événement. Le type 8 est son propre inverse, et la probabilité d'un échange particulier de lignées est calculé par

$$P_f = \frac{P_8}{n_i}.$$

4.3.3 Calcul de la probabilité d'acceptation des propositions

Pour que la procédure MCMC fournisse un échantillonnage selon la loi *a posteriori* $p(\phi|D)$, il faut que, conformément à l'algorithme de Metropolis-Hastings, les propositions de passage de la chaîne de Markov d'un état à un autre de l'espace Φ des paramètres soient acceptées avec probabilité

$$\min\left(1, \frac{p(\phi'|D)P_r}{p(\phi|D)P_f}\right)$$

où P_f est la probabilité de la proposition de passage de l'état 1 à l'état 2 (ce dernier étant dénoté par des prime), et P_r est la probabilité que l'on aurait de proposer le passage retour de 2 à 1.

On réécrit $p(\phi|D)$ sous sa forme inversée par la formule de Bayes, $p(D|\phi)p(\phi)$ à une constante multiplicative près (le dénominateur $p(D)$ est omis). Comme précédemment, on remplace $p(D|\phi)$ par $p(D|\phi,\mathcal{G})p(\mathcal{G}|\phi)$. En choisissant pour $\mathcal{G}$ les séquences d'événements $\mathcal{E}$ compatibles avec le jeu de données, $\mathcal{G} = \{D, \mathcal{E}\}$, on a toujours $p(D|\phi,\mathcal{G}) = 1$.

En omettant D dans l'écriture, la probabilité d'acceptation devient

$$\min\left(1, \frac{p(\mathcal{E}'|\phi')p(\phi')P_r}{p(\mathcal{E}|\phi)p(\phi)P_f}\right) = \min\left(1, \frac{L_2}{L_1} \times \frac{\pi_2}{\pi_1} \times \frac{P_r}{P_f}\right).$$

Dans cette formule, L_2/L_1 est le terme de rapport des vraisemblances, π_2/π_1 est le terme de rapport des densités *a priori*, et P_r/P_f est le rapport des probabilités de proposition, ou terme correctif de Hastings-Green [62]. Le terme L_2/L_1 s'exprime alors

$$\frac{L_2}{L_1} = \frac{p(\mathcal{E}'|\phi')}{p(\mathcal{E}|\phi)} \times \frac{p(\phi')}{p(\phi)}.$$

Puisque les paramètres modifiés sont la séquence d'événements $\mathcal{E}$ et les paramètres de Φ, le terme P_r/P_f est

$$\frac{P_r}{P_f} = \frac{p(\mathcal{E}, \Phi|\mathcal{E}', \Phi')}{p(\mathcal{E}', \Phi'|\mathcal{E}, \Phi)}.$$

Conclusion

La méthode qui vient d'être présentée est implémentée dans le programme `msvar0.5.0.c`. Le code commenté est disponible sur demande ainsi que des exécutables Unix-Linux et Windows. Un fichier `README` donne des éléments sur l'implémentation de la méthode et sur son usage.

Notons que la version utilisée dans la suite de cette thèse, et conforme à ce qui a été décrit précédemment, ne permet de supposer que des lois *a priori* en créneau. Comme nous allons le voir, cela facilite la vérification de l'implémentation par comparaison avec une méthode de Monte Carlo standard. Mais une version en développement permettra d'utiliser des lois *a priori* plus pertinentes [139].

Chapitre 5

Outils pour l'exploitation de MSVAR

Le programme MSVAR permet d'obtenir des échantillons de points de l'espace Φ des paramètres d'un modèle démographique et mutationnel (annexe page **??**). Ces points sont tirés de façon corrélée selon la loi *a posteriori* sur Φ pour des *a priori* $p(\phi)$ et un jeu de données D. L'objectif de la méthode est inférentiel, mais des préalables à ce type d'utilisation consistent (i) à en vérifier l'implémentation, (ii) à explorer la relation entre jeux de données, loi *a priori* et loi *a posteriori*, (iii) à déterminer quelles tailles de jeux de données constituent un bon compromis entre information démographique ou mutationnelle d'une part, et temps de calcul pour obtenir cette information d'autre part, (iv) à évaluer la robustesse de la méthode aux violations des hypothèses sous-jacentes.

L'objectif (i) peut être accompli en utilisant une méthode alternative pour le calcul de vraisemblance, la méthode de Monte Carlo directe. Les objectifs (ii), (iii), (iv) nécessitent le développement d'outils pour la simulation de jeux de données. La méthode de Monte Carlo directe reposant elle-même sur la simulation à répétition de jeux de données, nous présenterons l'algorithme de simulation de données, puis la méthode de Monte Carlo directe. Les préalables requièrent en outre, comme d'ailleurs l'objectif inférentiel, l'usage de méthodes de traitement des échantillons corrélés que nous aborderons ensuite, essentiellement pour le diagnostic de convergence et pour l'estimation de densité. Les outils présentés seront en fin de chapitre appliqués à la vérification de l'implémentation de MSVAR.

5.1 Simulation de jeux de données

Des données sont obtenues en simulant un coalescent standard avec variations d'effectif, additionné de mutations neutres, selon l'algorithme de R. Hudson [64]. Cet algorithme a été adapté par M. Beaumont pour la simulation de données microsatellites [6], puis par moi-même pour généraliser les modèles démographique et mutationnel, calculer des statistiques sommaires sur les jeux de données simulés et permettre la genèse d'échantillons corrélés.

On suppose que la taille de la population de gènes est passée il y a t_a générations de N_2 gènes à N_1 gènes, avant de varier de façon exponentielle pour atteindre N_0 gènes à la génération actuelle. Cela correspond à des valeurs des paramètres $r = N_0/N_1$, $a = N_0/N_2$ et $t_f = t_a/N_0$. On souhaite simuler un jeu de données de taille n gènes sous ce modèle démographique, pour un marqueur microsatellite mutant selon le modèle TPM. Cela suppose de générer un coalescent vierge de mutations pour n gènes, d'affecter un état allélique au MRCA, et de tirer les nombres et les types de mutations pour chacune des branches afin de déterminer les états alléliques dans

l'échantillon.

5.1.1 Construction d'un n-coalescent

Rappelons que la taille de population relative à la taille à $t = 0$ est donnée par

$$v(t) = \begin{cases} r^{-\frac{t}{t_f}} & \text{si } 0 \leq t \leq t_f, \\ \dfrac{1}{a} & \text{si } t > t_f. \end{cases}$$

Le taux de coalescence instantané s'exprime alors

$$c(t) = \frac{n_i(n_i-1)}{2 \quad v(t)} = \begin{cases} \dfrac{n_i(n_i-1)}{2} \times r^{\frac{t}{t_f}} & \text{si } 0 \leq t \leq t_f, \\ \dfrac{n_i(n_i-1)}{2} \times a & \text{si } t > t_f. \end{cases}$$

On souhaite simuler la date des événements de coalescence successifs, selon le taux $c(t)$. On résout pour cela en t l'équation $C(t) = \mathcal{U}$ où $\mathcal{U}$ est un tirage selon une loi uniforme sur $[0;1]$ et $C(t)$ est la fonction de répartition correspondant à $c(t)$ (méthode tout à fait classique d'inversion de la fonction de répartition de la loi à simuler [106], dont l'encart page 96 rappelle le principe).

Selon les positions relatives de la date t_1 de la dernière coalescence et de la date t_2 de la prochaine coalescence, par rapport à t_f, on obtient :

$$t_2 = \begin{cases} \dfrac{t_f}{\log r} \times \log\left(r^{t_1/t_f} - 2 \times \dfrac{\log(\mathcal{U})}{n_i(n_i-1)}\right) \times \dfrac{\log(r)}{t_f} & \text{si } 0 \leq t_1 < t_2 \leq t_f, \\ t_f - 2 \times \dfrac{\log(\mathcal{U})}{n_i(n_{i-1})} \times \dfrac{1}{a} - \dfrac{t_f}{a\log(r)}\left(r - r^{t_1/t_f}\right) & \text{si } 0 \leq t_1 \leq t_f < t_2, \\ t_i - 2 \times \dfrac{\log(\mathcal{U})}{n_i(n_i-1)} \times \dfrac{1}{a} & \text{si } t_f < t_1 < t_2. \end{cases} \tag{5.1}$$

Dans la pratique, pour réaliser le tirage de la prochaine date de coalescence, on tire uniformément selon $\mathcal{U}$ et on calcule le terme

$$-2\frac{\log(\mathcal{U})}{n_i(n_i-1)} \quad , \tag{5.2}$$

conformément au nombre n_i de lignées ancestrales de l'échantillon. Pour $t_1 < t_f$, la réalisation de la condition $t_2 \leq t_f$ nécessite que

$$-2\frac{\log(\mathcal{U})}{n_i(n_i-1)} \leq \left(r - r^{t_1/t_f}\right)\frac{t_f}{\log(r)} \quad , \tag{5.3}$$

ce qui décide du calcul de t_2 selon la première ou la seconde expression du système. Pour $t_1 \geq t_f$, t_2 est calculé selon la troisième expression. On choisit uniformément une paire de gènes parmi les $n_i(n_i-1)/2$ paires pouvant coalescer, et on les fusionne à la date t_2. On répète le processus jusqu'à avoir fait coalescer les deux dernières lignées à une date t_{MRCA}. On dispose donc d'un n-coalescent, vierge de mutations, dont on a mémorisé la topologie (chaque noeud coalescent connait son noeud coalescent ancêtre et ses deux noeuds coalescents descendants) et les dates (chaque noeud coalescent a une date absolue mesurée en unités de N_0 générations).

5.1.2 Affectation d'états alléliques à l'échantillon

Pour les modèles considérés, la classe allélique n'influence pas le processus mutationnel. On affecte donc arbitrairement la taille 0 au MRCA, et on en déduit la taille des noeuds coalescents descendants, jusqu'à l'échantillon, en nombre de répétitions de plus ou de moins que le MRCA (méthode "allele dropping"). Un moyen simple de considérer toutes les branches est de parcourir la liste des noeuds coalescents, du plus ancien (proche du MRCA) au plus récent. La taille allélique du noeud ancêtre est toujours connue, et il suffit de tirer le nombre et le type des mutations sur chaque branche pour déterminer la classe allélique du noeud en cours de traitement. Le nombre de mutations sur une branche de longueur δt est tiré selon une loi de Poisson de paramètre $\delta t \times \theta/2$. Les types de mutations sont tirés selon la distribution d'amplitude voulue. Pour le modèle SMM symétrique, il y a équiprobabilité pour les mutations gain et perte d'une unité de répétition. Pour un TPM, la distribution d'amplitude est déterminée par les paramètres de probabilité de saut p et de moyenne d'amplitude des sauts j.

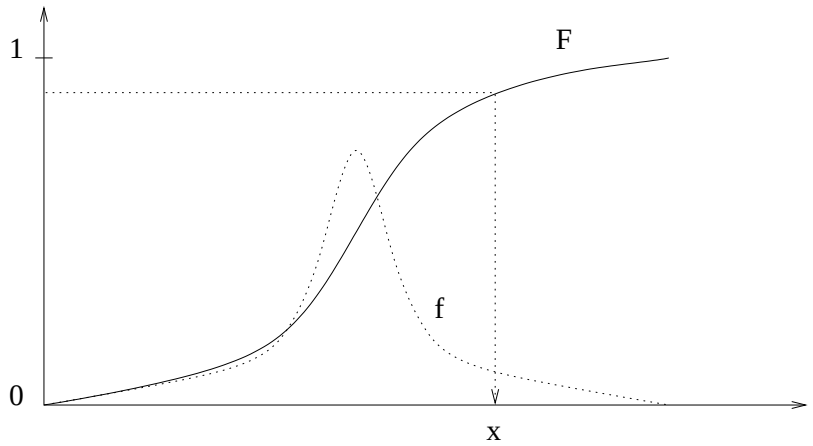
5.1.3 Simulation de généalogies corrélées

Le temps de simulation nécessaire pour estimer précisément la loi *a posteriori* des paramètres par la méthode bayésienne empêche d'explorer le modèle de façon systématique. Par exemple, si l'on souhaite savoir si supposer un modèle SMM alors que les marqueurs mutent selon TPM nuit à l'inférence sur les paramètres démographiques. On pourrait simuler 100 jeux de données sous SMM, 100 jeux de données sous TPM, pour un même jeu de valeurs des paramètres démographiques, et comparer l'erreur moyenne faite sur les inférences démographiques, dans un cas et dans l'autre. Malheureusement, si 24 heures de calcul sont nécessaires pour chaque jeu de données, il faut en tout 200 journées de calcul. On ne peut alors pas se permettre de tâtonner, et d'explorer plusieurs valeurs des paramètres.

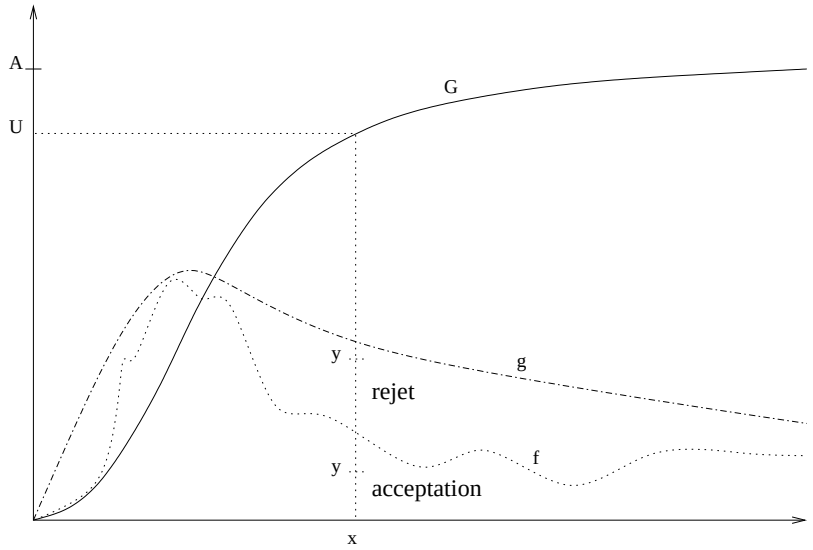
Dans ce contexte délicat, il m'est apparu intéressant de pouvoir simuler des jeux de données qui ne diffèrent que par certaines des composantes du processus stochastique. Par exemple, on peut vouloir fixer la topologie du n-coalescent et ne faire varier que les dates des coalescences, et l'histoire mutationnelle. On peut vouloir fixer la topologie datée et modifier le nombre, la position et le type des mutations, ou encore fixer en plus le nombre et la position de mutations, voire leur sens et la valeur de la fonction de répartition de l'amplitude, et ne faire varier que la distribution d'amplitude. Ces différentes possibilités sont implémentées, et leur usage sera signalé au fur et à mesure de l'exploration du modèle.

Simulation de tirages selon une loi de densité explicite

Deux grandes méthodes permettent de simuler des tirages pseudo-aléatoires selon une loi de densité f, à partir d'un générateur uniforme [134]. La première méthode est applicable dès lors que la fonction de répartition F de la loi peut être inversée explicitement. Rappelons que la fonction de répartition est continue à droite, croissante, et telle que $F(-\infty) = 0$ et $F(+\infty) = 1$. L'inverse généralisé de F sur $[0;1]$ défini par $F^{-1}(t) = \inf\{x \in \mathbb{R}; F(x) > t\}$ est également continu à droite, et si $\mathcal{U}$ est une loi uniforme sur $[0;1]$, $F^{-1}(\mathcal{U})$ a pour loi f.



La seconde méthode, plus générale, ne nécessite pas que F^{-1} soit explicite (ni même F d'ailleurs). Elle consiste à réaliser un tirage uniforme en 2 dimensions, sous la courbe représentative de f, en réalisant un tirage uniforme sur une surface incluant celle-ci, par la méthode précédente. Pour cela, on choisit une fonction g définie sur le même domaine que f, partout supérieure à elle, d'intégrale finie $A > 1$ et pouvant être explicitement inversée. On réalise par inversion de la primitive adéquate un tirage d'une abscisse x (toutefois, le tirage uniforme est sur $[0;A]$ et non pas sur $[0;1]$, où A est la valeur de la surface sous g), puis un tirage d'une ordonnée y, uniforme sur $[0;g(x)]$. Le point de coordonnées (x,y) est alors tiré uniformément sous la courbe représentative de g. Si $y \leq f(x)$ (et on répète le tirage jusqu'à ce que ce soit le cas), il est également tiré uniformément sous f, c'est à dire que son abscisse x est tirée selon f. On a intérêt à minimiser la probabilité de retirage, et donc à choisir g de façon que A soit aussi proche de 1 que possible.



5.2 Échantillonnage de Monte Carlo direct

Supposons fixé le vecteur de paramètres démographiques et mutationnels à une valeur ϕ. La vraisemblance de cette valeur est la probabilité d'observer une configuration identique à celle du jeu de données, par simulation selon le processus stochastique déterminé par ϕ. Ce processus stochastique consiste en le coalescent avec variations d'effectif, additionné de mutations distribuées uniformément sur le coalescent, et la vraisemblance peut être estimée comme la proportion de simulations du processus stochastique donnant la configuration de l'échantillon, comme illustré ci-après sur des exemples concrets.

Considérons un jeu de données de taille 5 gènes et de configuration normalisée $\{1, 3, 0, 1\}$ (encart page 78). Fixons tous les paramètres démographiques et mutationnels à des valeurs simples, soit $N_0 = N_1 = N_2$ (effectif stable, t_f n'a alors aucune influence), et $p = 0$ (modèle mutationnel SMM). Le processus stochastique est dans ce cas réduit au *coalescent standard*, avec des mutations SMM.

Seul le taux de mutation renormalisé θ sera dans notre exemple autorisé à varier, mais on aurait aussi bien pu proposer un maillage multidimensionnel de Φ. Fixons pour commencer θ à la valeur 10. Simulons avec ces valeurs des paramètres un grand nombre de 5-coalescents et ajoutons-y des mutations uniformément, conformément à la valeur de θ. Une première simulation fournit par exemple un échantillon simulé de configuration $\{1, 1, 3\}$. Une seconde simulation donne la configuration $\{3, 2\}$. À la 78-ième simulation, on obtient la configuration $\{1, 3, 0, 1\}$, identique à celle de l'échantillon. Sur 1 million de simulations, on atteint 11097 fois la configuration cible $\{1, 3, 0, 1\}$. Sur un autre million de simulations, on l'atteint 11041 fois. Un estimateur à 10^{-5} près de la vraisemblance de la valeur de paramètres $\phi = \{N_0 = N_1 = N_2; p = 0; \theta = 10\}$ est donc 0.0110, pour la configuration d'échantillon $\{1, 3, 0, 1\}$.

On peut estimer de la même façon la vraisemblance des données pour une gamme de valeurs de θ. Sur un million de simulations pour chaque valeur de θ choisie, on a atteint la configuration cible un nombre de fois reporté dans le tableau 5.1 (la variabilité intersimulation ne change pas l'ordre de grandeur de ce nombre) :

θ	0.1	0.5	1	3	5	7	10	50	100	1000
$10^6 \times L(\theta)$	30	1807	5207	13215	14238	13167	11041	2180	766	10

TAB. *5.1: Estimation par échantillonnage de Monte Carlo direct de la vraisemblance des jeux de paramètres* $\phi = \{N_0 = N_1 = N_2; p = 0; \theta\}$*, pour un échantillon de configuration* $\{1, 3, 0, 1\}$*. La valeur de* θ *qui donne la plus forte probabilité au jeu de données est située à proximité de* $\theta = 5$*.*

La méthode de Monte Carlo directe permet donc, pour de petites tailles d'échantillon, d'obtenir des estimations de la valeur de $L(\theta)$ —où plus généralement de $L(\phi)$—, pour différentes valeurs de θ —ϕ.

5.3 Traitement des échantillons MCMC

5.3.1 Diagnostic de convergence

Les méthodes fondées sur de l'échantillonnage par MCMC sont extrêmement satisfaisantes sur le papier, mais peuvent s'avérer très difficiles d'utilisation. En effet, la MCMC est construite

pour que sa distribution stationnaire soit la loi *a posteriori* recherchée. Mais les états successifs de la MCMC sont des tirages corrélés selon cette loi, et non pas des tirages indépendants comme dans les méthodes de Monte Carlo statiques. Si l'autocorrélation le long de la chaîne de Markov est trop forte (c'est le cas du fait de l'algorithme d'exploration des généalogies, qui procède par petites modifications), il est nécessaire d'espacer les points d'échantillonnage, afin qu'ils fournissent une bonne représentation de la loi *a posteriori*, sans que leur nombre augmente trop fortement.

Des critères visuels permettent de se convaincre de la convergence d'une simulation MCMC. Une première possibilité consiste à visualiser les variations des valeurs des paramètres d'intérêt en fonction du nombre d'itérations (figure 5.1). Lorsque le degré d'espacement (entre points considérés pour l'estimation de densité *a posteriori*) est trop faible, les valeurs des paramètres se modifient mollement avec le nombre d'itérations. Idéalement, l'espacement doit être suffisant pour que les points d'échantillonnage successifs soient indépendants. Malheureusement, avec des jeux de données de taille raisonnable (quelques loci, quelques dizaines d'individus), l'explosion du temps de calcul empêche d'augmenter l'espacement suffisamment pour frôler l'indépendance. Un critère de convergence consiste alors à comparer les estimations de la loi a posteriori obtenues d'après des simulations indépendantes. Ces simulations doivent être caractérisées par des points de départ dispersés dans l'espace des paramètres et des généalogies, afin d'éviter des corrélations entre chaînes prétendument indépendantes.

Une littérature abondante discute de moyens plus formels de vérification de la convergence (*e.g.* [16]). Certains critères sont implémentés dans le package CODA [14] utilisable dans R [65]. CODA permet notamment de représenter l'autocorrélation d'une chaîne de Markov, et de calculer l'indice de convergence de Gelman-Rubin (encart 5.3.1). Tous les résultats de simulation présentés dans cette thèse ont fait l'objet d'un diagnostic de convergence, soit à l'oeil (pour les données simulées monolocus, peu problématiques), soit par le critère de Gelman-Rubin (pour les données simulées multilocus et les données réelles).

Critères de convergence MCMC

Deux critères de convergence dominent la littérature MCMC. Le critère de *Raftery-Lewis* [110] est basé sur la simulation d'une unique chaîne surdimensionnée, et détermine le nombre d'itérations nécessaire pour estimer des quantiles des distributions *a posteriori* avec un degré de précision donné. Le critère de *Gelman-Rubin* [48] utilise plusieurs chaînes indépendantes plus courtes, et compare la variabilité intra-chaîne et inter-chaîne pour les paramètres. La convergence est typiquement considérée comme satisfaisante lorsque pour chacun des paramètres d'intérêt, la variance interchaîne V_{inter} de la moyenne représente moins de 5% de la moyenne des variances intra-chaîne V_{intra}, ce qui correspond à [49] :

$$\sqrt{\frac{V_{intra} + V_{inter}}{V_{intra}}} < 1.1$$

Le critère de Raftery-Lewis est plus exigent que le critère de Gelman-Rubin, et il est difficile d'y satisfaire dans la pratique, pour des MCMC basées sur le coalescent. Même lorsque les critères visuels et le critère de Gelman-Rubin indiquent une convergence suffisante, le critère de Raftery-Lewis n'est en général pas rempli.

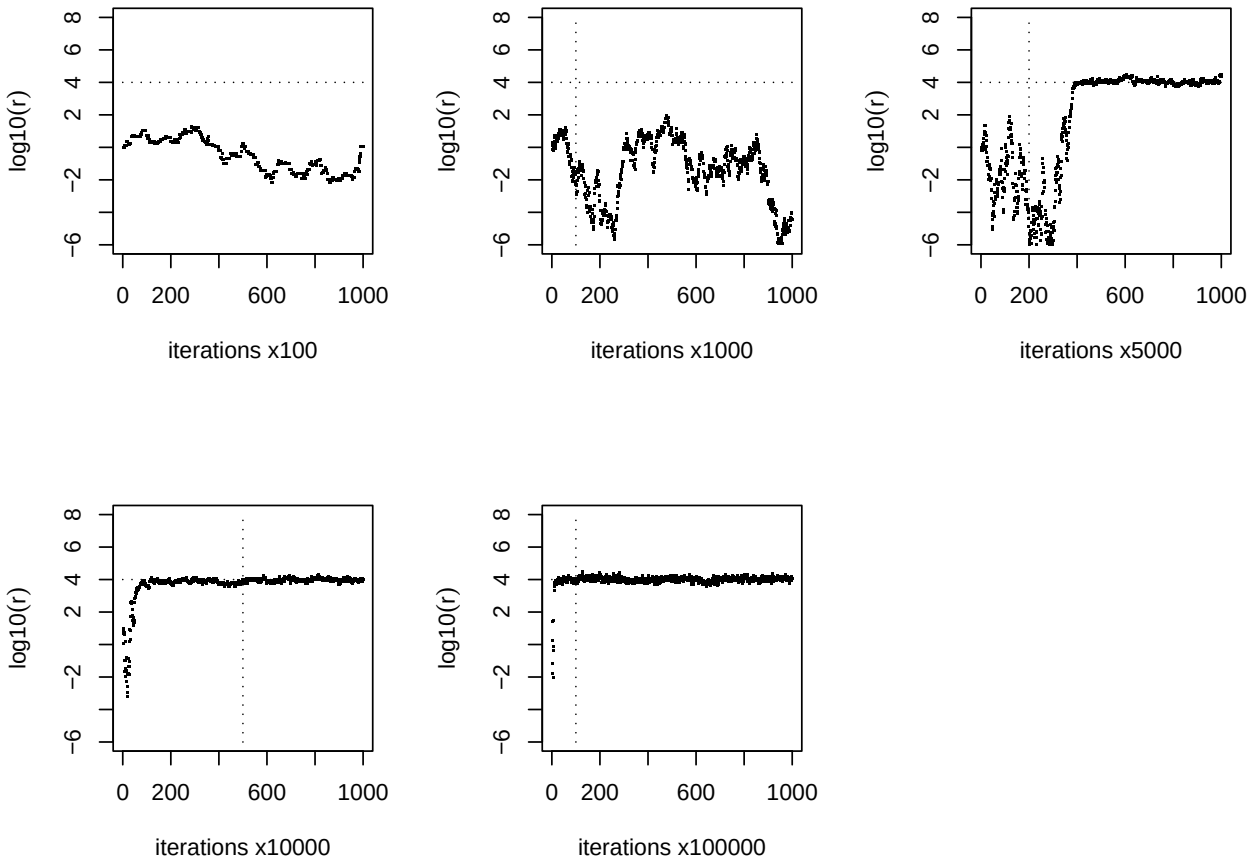


FIG. *5.1: Effet de l'espacement des points d'échantillonnage sur l'autocorrélation le long d'une MCMC. Un jeu de données de* 5 *loci indépendants, de taille* 50 *gènes par locus, a été généré en supposant une histoire de fondation-explosion avec* $r = 10\,000$, $a = 1$, $t_f = 0.001$ *et* $\theta = 100$ *(configuration exacte donnée dans le tableau 6.3). La loi a priori pour les paramètres* r, a, *et* θ *a été choisie uniforme pour leur logarithme en base* 10, *sur les intervalles respectifs* $[-6;8]$, $[-5;5]$ *et* $[-3;5]$. *La loi* a priori *pour le paramètre* t_f *est une masse de probabilité* 1 *pour la "vraie" valeur* $t_f = 0.001$. *Les valeurs d'initialisation de la simulation MCMC pour les paramètres sont communes aux* 4 *simulations (*$r = 1$, $a = 1$, $\theta = 0.1$ *et bien sûr* $t_f = 0.001$*) ainsi que la généalogie initiale. Le modèle mutationnel supposé est le SMM. Chaque cadre de la figure montre les variations du paramètre* r *en fonction du nombre de points d'échantillonnage. L'espacement entre points (c'est à dire le nombre d'itérations non reportées entre deux points plus* 1*) vaut* 100, 1000, 5000, 10 000 *et* 100 000 *pour les cadres successifs. Chaque simulation est donc résumée dans les premières itérations de la suivante (la zone répétée est limitée à droite par un pointillé vertical). Pour une valeur d'espacement de* 100 *et* 1000, *la valeur de* r *aux points d'échantillonnage est fortement influencée par les valeurs d'initialisation. Lorsque l'espacement augmente, on s'affranchit des valeurs initiales, et la part de la simulation qui leur correspond se réduit peu à peu. Dans le cas illustré un espacement de* 10000 *est suffisant à condition de tronquer les premiers points d'échantillonnage (durée de simulation :* $\sim 1h$ *sur un PC équipé d'un processeur 1GHz pour la valeur d'espacement de* 10 000*).*

5.3.2 Estimation de densités a posteriori

On peut représenter une loi *a posteriori*, jointe ou marginale, par la collection des points échantillonnés (par MCMC) selon cette loi. Nous le ferons abondamment dans le chapitre 6, pour des représentations bivariées. Il peut toutefois être pratique, pour faire la synthèse de résultats, d'estimer la loi *a posteriori* à partir de l'échantillon : cela permet de calculer des modes, des quantiles ou des intervalles de densité *a posteriori*, sur lesquels on peut baser une prise de décision (comparaison de modèles, sélection d'hypothèses). Les logiciels de traitement statistique courants (R [65], Mathematica) implémentent des méthodes d'estimation de densité. Une méthode standard consiste à sommer des noyaux affectés à chaque point. Les noyaux sont de petites distributions continues (*e.g.* en créneau, gaussienne), centrés sur les points et plus ou moins étalés autour d'eux. La densité est estimée comme la somme des noyaux (les noyaux de points suffisamment proches se chevauchent). Le choix de la variance des noyaux détermine à quel point on lisse la densité estimée : si la variance est trop faible, la densité estimée risque de comporter du bruit ; si elle est trop grande, on lisse trop et on risque de manquer des caractéristiques importantes de la collection de points.

Dans la présente thèse, les densités *a posteriori* sont estimées en utilisant un critère de vraisemblance locale, implémenté dans le package LOCFIT [84] disponible pour R. LOCFIT permet également de tracer des courbes de niveau de densité *a posteriori*, qui définissent des domaines de probabilité *a posteriori* (ou intervalles HPD, pour Highest Posterior Density).

5.4 Application à la vérification de MSVAR

L'implémentation du programme MSVAR est extrêmement complexe, et il est indispensable de pouvoir en vérifier le bon fonctionnement. La méthode de Monte Carlo directe est très utile dans ce contexte (paragraphe 3.3.1). En effet, pour des jeux de données monolocus de configuration D, d'au maximum une dizaine de gènes, cette méthode permet de calculer $p(D|\phi)$. En choisissant une loi *a priori* rectangulaire pour chaque paramètre dans Φ, on peut obtenir par la méthode bayésienne par MCMC une loi *a posteriori* qui coïncide avec la vraisemblance dans le domaine de densité *a priori* uniforme (à une constante multiplicative près). Une estimation de la surface de vraisemblance dans le même domaine, par échantillonnage de Monte Carlo direct sur un maillage de cet espace, permet donc une vérification de la méthode bayésienne. Une vérification supplémentaire consiste à comparer des caractéristiques des généalogies échantillonnées par MC et par MCMC, cette fois tous paramètres fixés.

Des éléments de vérification de la version `msvar0.4.1b.c` du programme ayant déjà été publiés dans la référence [6], je ne détaillerai que la vérification des nouvelles implémentations. Les extensions démographique et mutationnelle de [6] ont été validées séparément. D'une part, la surface univariée de vraisemblance a été comparée entre MC et MCMC , pour les paramètres r, a, t_f et θ successivement, les autres paramètres étant fixés et le modèle mutationnel étant réduit à un SMM strict. D'autre part, pour des valeurs fixées des paramètres démographiques et un modèle TPM, la distribution du nombre de mutations par généalogies, et la distribution intergénéalogies des amplitudes des mutations ont été comparées entre les deux méthodes.

5.4.1 Comparaison des inférences démographiques

Un échantillon cible de configuration $\{1, 9\}$ a été choisi arbitrairement. Sa taille $n = 10$ est suffisamment faible pour que la méthode de Monte Carlo directe estime efficacement la vraisemblance. Les paramètres sont fixés aux valeurs $r = 6\,608$, $a = 1$, $t_f = 0.00075$, $\theta = 40$ (pour information, ces valeurs décrivent l'histoire documentée du chat introduit aux Kerguelen). Les paramètres p et j sont tels que les mutations suivent le modèle SMM ($p = 1$ et $j = 1$ par exemple). Chaque cadran de la figure 5.2 correspond à l'estimation de la surface de vraisemblance, conditionnellement aux valeurs des paramètres fixés, obtenue en faisant varier un seul des paramètres θ, r, t_f et a, à la fois (figure 5.2). Pour la méthode de Monte Carlo directe, la vraisemblance est estimée pour des valeurs du paramètre variable régulièrement espacées sur une échelle logarithmique. Pour la méthode de Monte Carlo par chaîne de Markov, seul ce paramètre est autorisé à varier d'une itération à la suivante. La surface sous la courbe est normalisée entre les bornes d'exploration du paramètre qui varie, pour mettre les résultats des deux méthodes à la même échelle. La figure 5.2 montre la bonne adéquation entre les deux méthodes, pour une estimation par MC et 5 estimations indépendantes par MCMC, à partir de généalogies initiales différentes.

5.4.2 Comparaison de la distribution des mutations

Des échantillons cibles de configurations normalisées $\{2\}$, $\{3\}$, $\{4\}$, $\{1, 3, 0, 1\}$ et $\{1, 4, 0, 2, 0, 3\}$ ont été successivement utilisés pour vérifier l'implémentation du modèle TPM. Un échantillon trivial de deux gènes identiques a d'abord permis de vérifier l'implémentation des probabilités de transition pour les modifications généalogiques (en particulier pour les types 5 et 6 nouvellement implémentés). Une fois la MCMC ajustée à la MC pour cet échantillon minimal, un échantillon de trois gènes identiques a été utilisé, et ainsi de suite pour les échantillons de quatre gènes identiques, et les échantillons plus complexes $\{1, 3, 0, 1\}$ et $\{1, 4, 0, 2, 0, 3\}$. Les deux méthodes concordent bien pour toutes ces configurations d'échantillons (illustration figure 5.3 pour l'échantillon de configuration $\{1, 3, 0, 1\}$). Les valeurs choises des paramètres sont $a = r = 1$, $t_f = 0$, $\theta = 1$, $p = 0.5$ et $j = 2$, ce qui correspond à une population d'effectif stable N_0 et de taux de mutation μ tels que $\theta = 2N_0\mu = 1$. Le modèle TPM est volontairement forcé, avec 25% de mutations d'amplitude $a > 1$ et une amplitude moyenne des sauts de 1.5. Cela permet une comparaison des deux méthodes meilleure qu'avec un TPM plus réaliste, pauvre en mutations sauts. On notera que la vraisemblance du jeu de valeurs des paramètres est assez élevée, de l'ordre de 6.10^{-3}. L'innocuité des modifications successives du programme a été vérifiée régulièrement sur la configuration $\{1, 3, 0, 1\}$.

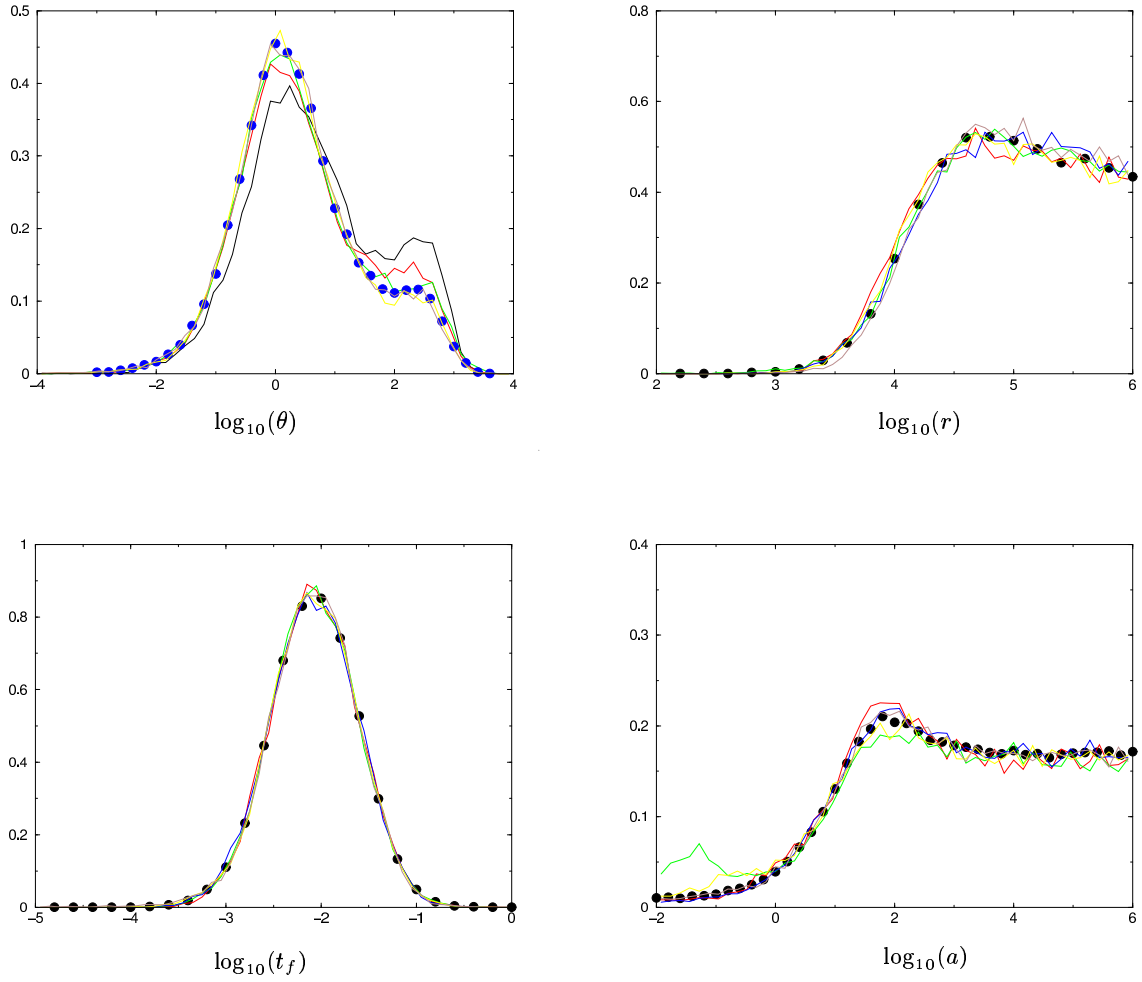


FIG. *5.2: Adéquation entre la vraisemblance conditionnelle des paramètres θ, r, t_f et a, telles qu'estimées pour une configuration d'échantillon $\{1, 9\}$ par MCMC (5 itérations de 20 000 points avec un espacement de 10 000, traits pleins) et par MC (estimation sur 10 0000 itérations, points noirs). N.B. : simulations et figure réalisées par M. Beaumont.*

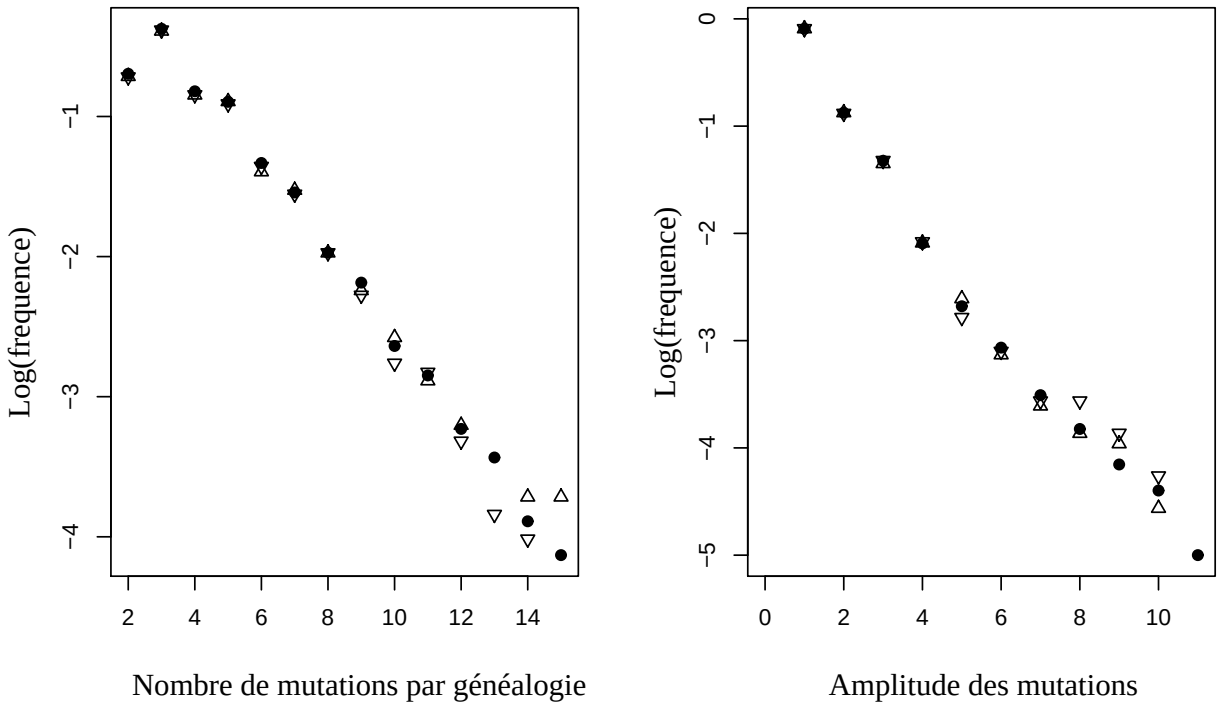


FIG. *5.3: Adéquation entre la distribution du nombre de mutations par généalogie (à gauche) et la distribution de l'amplitude des mutations toutes généalogies confondues (à droite), telles qu'estimées pour une configuration d'échantillon* $\{1, 3, 0, 1\}$ *sur des généalogies échantillonnées respectivement par MCMC (deux estimations sur* 20000 *points chacune, triangles et triangles renversés) et par MC (estimation sur* 100000 *itérations, points noirs). Le modèle démographique et mutationnel est défini par les valeurs des paramètres* $a = r = 1$, $t_f = 0$, $\theta = 1$, $p = 0.5$ *et* $j = 2$, *ce qui correspond à une population d'effectif stable, typée à un marqueur mutant selon un modèle TPM.*

Conclusion

L'échantillonnage de Monte Carlo direct, naïf lorsqu'il s'agit de développer une méthode d'inférence, s'avère extrêmement utile pour vérifier l'implémentation de méthodes plus élaborées, et en particulier celles basées sur un échantillonnage MCMC. Le tirage aléatoire des temps d'attente entre événements est un outil commun aux deux méthodes, donc la comparaison ne permet pas de détecter des erreurs sur l'implémentation de ces tirages (il faut pour cela vérifier directement la distribution des temps d'attente simulés). La comparaison entre MC et MCMC permet en revanche de détecter les erreurs de programmation dans les procédures d'exploration des généalogies, qui constituent la part la plus délicate à programmer, et les erreurs dans le calcul des probabilités de transition, point clé de l'algorithme de Metropolis-Hastings.

Chapitre 6

Exploration du modèle de fondation-explosion

Avant d'utiliser la méthode d'inférence sur des données de typage de populations réelles, on explore le comportement du modèle et la précision de la méthode, sur des données simulées. Ces données sont obtenues sous les modèles démographique et mutationnel supposés dans la méthode, puis dans des conditions de violation de certaines des hypothèses de ces modèles.

6.1 Relation entre histoire démographique et vraisemblance

Notre objectif ici est de rechercher une relation entre histoire démographique et surface de vraisemblance pour les paramètres démographiques et mutationnels. Le temps d'obtention à l'aide de MSVAR d'un échantillon de 20 000 points représentatif de $p(D|\phi)$ rend cette tâche difficile. Ce temps est en effet de l'ordre de la journée pour un échantillon monolocus de taille 100 sur un PC récent, et atteint facilement la semaine pour un échantillon multilocus. Une dizaine de scénarios ont été partiellement explorés, et pour chacun plusieurs configurations d'échantillons. Il apparaît *a posteriori* que l'essentiel peut être montré en considérant seulement trois scénarios bien typés.

6.1.1 Scénarios démographiques et statistiques sommaires

Des jeux de données de taille 100 gènes (*i.e.* 50 individus diploïdes) ont été simulés avec $\theta = 2N_0\mu = 10$ et $p = 0$ (modèle de mutation SMM). Trois histoires démographiques ont été supposées (figure 6.1) par choix des valeurs de $r = N_0/N_1$, de $a = N_0/N_2$ et de la date $t_f = t_a/N_0$. On rappelle que t_a est un nombre de générations, que N_0 est la taille de la population —en nombre de génomes haploïdes— à la date $t = 0$ de l'échantillonnage, que N_1 est la taille de la population juste après l'événement qui se produit à t_f, et que N_2 est la taille de la population ancestrale, jusqu'à t_f. (i) Avec $a = r = 1$, l'effectif de la population est stable ; cette situation peut par exemple correspondre à une population de taille 5 000 individus diploïdes typés à des marqueurs microsatellites de taux de mutations 5.10^{-4}. (ii) Avec $a = r = 100$ et $t_f = 0.05$, on suppose un passé de croissance exponentielle pendant $t_f = 0.05$ unités g/N_0 de temps, avec un rapport d'effectif $r = 100$ entre le début de la croissance exponentielle et la date d'échantillonnage. L'effectif avant t_f est stable, et en continuité avec la phase de croissance exponentielle. Cela correspond par exemple à une population de taille efficace 50 individus il y

a 500 générations (et antérieurement) et 5 000 individus au moment de l'échantillonnage, typée à des marqueurs microsatellites de taux de mutations 5.10^{-4}. (iii) Avec $a = 1$ et $r = 100$, on a le même profil de croissance exponentielle, mais avec une discontinuité à t_f, puisque l'effectif de la population ancestrale est le même que dans la population échantillonnée. Cela correspond par exemple à une population de taille efficace stable 5000 individus, qui chute brutalement à 50 individus il y a 500 générations et retrouve son effectif initial de 5 000 individus au moment de l'échantillonnage, suite à une phase de croissance exponentielle. Le taux de mutation des marqueurs microsatellites demeure 5.10^{-4}. Les trois scénarios sont donc bien tranchés, en restant réalistes (taux de mutation des marqueurs standard, taux de croissance des populations raisonnable).

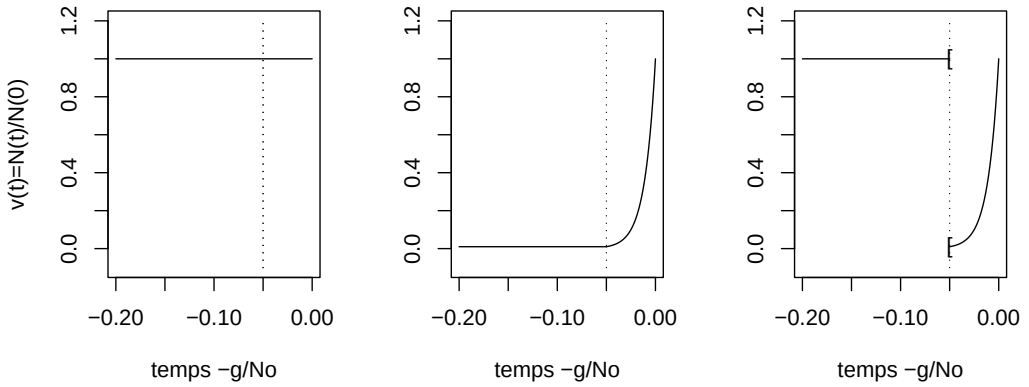


FIG. *6.1: Représentation de $v(t) = N(t)/N_0$ pour chacun des trois scénarios démographiques considérés. Le scénario de stabilité démographique ($v(t) = 1$, à gauche), le scénario de croissance exponentielle ($v(t) = 100^{t/t_f}\ \mathbb{I}_{t\geq -0.05} + 0.01\ \mathbb{I}_{t<-0.05}$, au centre) et le scénario de fondation-explosion ($v(t) = 100^{t/t_f}\ \mathbb{I}_{t\geq -0.05} + \mathbb{I}_{t<-0.05}$, à droite) sont rendus comparables par le partage de l'effectif à $t = 0$. Cela se traduit par une même valeur du taux de mutation renormalisé $\theta = 2N_0\mu$.*

Ces histoires démographiques conduisent à des généalogies simulées du même type que celles représentées aux figures 2.6, 2.8 et 2.10 (les scénarios sont les mêmes). Pour les deux scénarios qui comportent une phase de croissance exponentielle avec $r = 100$ et $\theta = 10$, on a une dérive notable dans les générations qui suivent la fondation, malgré la croissance exponentielle qui débute immédiatement. Considérons pour le montrer uniquement la fraction des généalogies plus récente que $t_f = 0.05$. En l'absence de dérive, les généalogies seraient en étoile. Pour des échantillons de taille $n = 100$, la longueur totale des généalogies serait donc $n \times t_f = 5$ unités de temps, et le nombre espéré de mutations sur les généalogies $n \times t_f \times \theta/2 = 25$. Au lieu de cela, les généalogies simulées ont des branches profondes de longueur non négligeable, sur lesquelles des mutations peuvent se produire. Du fait de cette structure généalogique, la longueur totale moyenne des généalogies est seulement 1.44 unités, et le nombre espéré de mutations est réduit à 7.2.

Il n'est pas innocent qu'en croissance exponentielle, le temps de coalescence de l'échantillon soit distribué autour de la date t_f. C'est en effet une situation dans laquelle les configurations d'échantillons obtenues sont très différentes selon que la croissance est précédée ou non d'une fondation. La figure 6.2 montre l'effet des scénarios avec variation d'effectif sur le nombre d'allèles, la variance de taille allélique et l'homozygotie, estimées sur 10000 jeux de données simulés pour chaque scénario. Les échantillons générés en croissance exponentielle ont des généalogies

plus courtes, donc moins variables. Le scénario de fondation-explosion a une variabilité intermédiaire. La distribution jointe de statistiques sommaires distingue nettement les trois scénarios (figure 6.3).

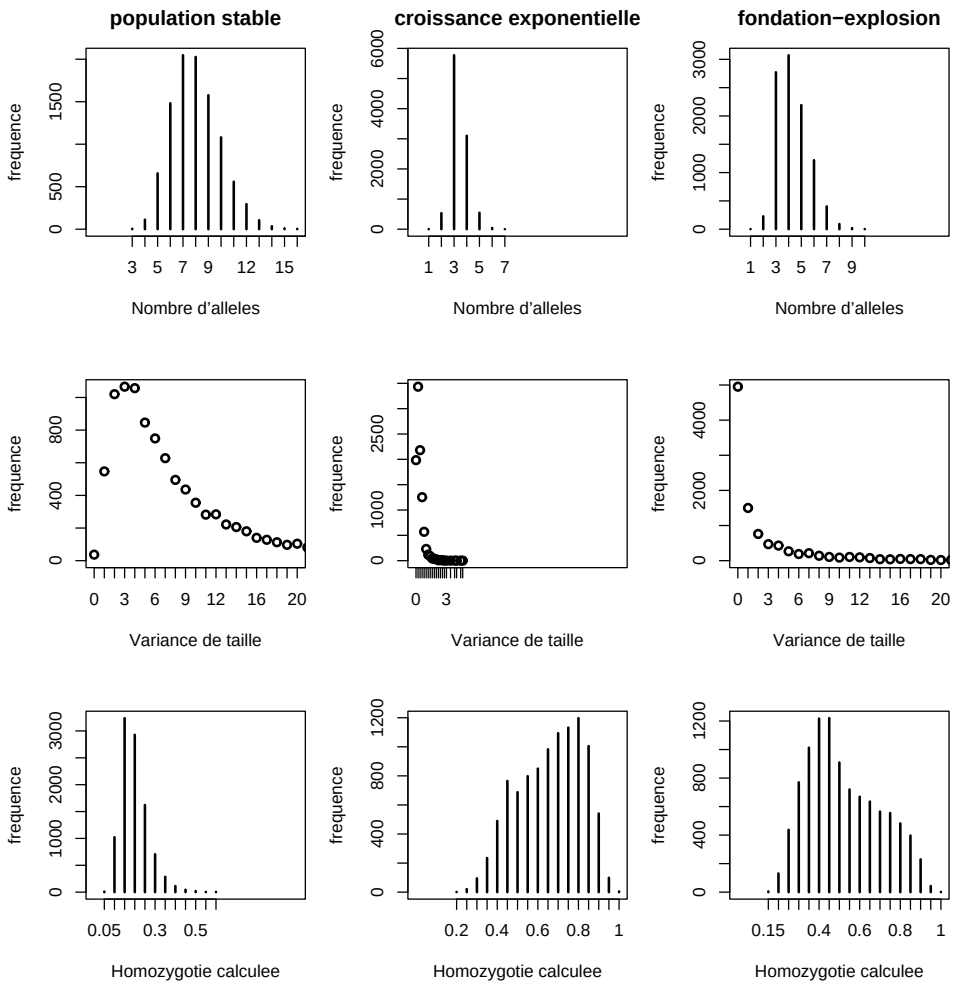


FIG. *6.2: Distribution de statistiques sommaires obtenue à partir de* 10 000 *échantillons de taille* 100 *gènes simulés sous SMM en population stable* ($r = a = 1$, *colonne de gauche), en supposant une croissance exponentielle* ($r = a = 100$, *au milieu) et une histoire de fondation-explosion* ($r = 100$ *et* $a = 1$, *à droite) avec* $\theta = 10$ *dans les trois cas. La première ligne donne la distribution du nombre* A *d'allèles dans les échantillons : on a* $A_s = 8.0 \pm 1.9$ *en population stable,* $A_e = 3.4 \pm 0.7$ *en régime de croissance exponentielle et* $A_f = 4.3 \pm 1.2$ *suite à la fondation-explosion. La seconde et la troisième ligne donnent une discrétisation de la distribution de la variance de taille allélique, et de l'homozygotie calculée respectivement. La distribution de la variance de taille est très dissymétrique, avec des couples (mode,moyenne) de* $(3, 9.9)$, $(0.20, 0.45)$ *et* $(0.2, 4.0)$ *pour les trois modèles.*

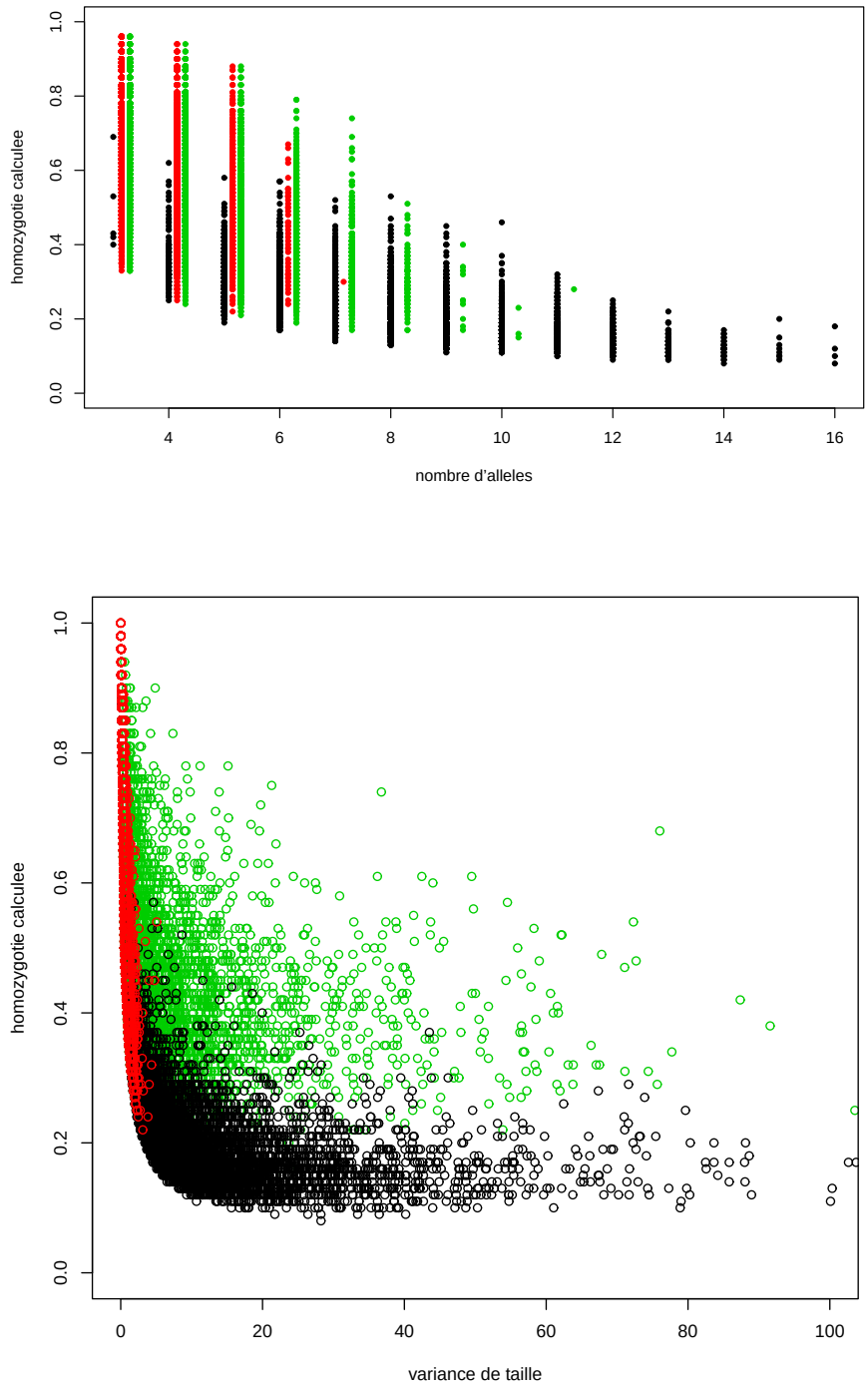


FIG. *6.3: Distribution jointe de statistiques sommaires obtenue à partir de* 10000 *échantillons de taille* 100 *gènes simulés sous SMM en population stable* ($r = a = 1$, *en noir), en supposant une croissance exponentielle* ($r = a = 100$, *en rouge) et une histoire de fondation-explosion* ($r = 100$ *et* $a = 1$, *en vert) avec* $\theta = 10$ *dans les trois cas. En haut, statistiques utilisées dans le programme* BOTTLENECK *[24, 105] : le nombre d'allèles* A *et l'homozygotie calculée* $\widehat{H}$. *En bas, statistiques utilisées par Kimmel* et al. *pour définir l'index de déséquilibre* β *[71] : la variance de taille* $\widehat{V}$ *et l'homozygotie calculée* $\widehat{H}$.

6.1.2 Typologie de la loi a posteriori des paramètres démographiques

Un grand nombre d'échantillons monolocus, tous simulés selon le modèle sous-jacent à la méthode d'inférence, ont été traités avec le programme MSVAR qui implémente cette méthode. Une observation attentive des surfaces de vraisemblance obtenues pour les paramètres démographiques et mutationnels a montré que toutes peuvent être rattachées à 4 grands types. Certaines correspondent vraiment à l'un de ces types, d'autres sont un peu hybrides entre deux types. Afin d'illustrer et d'interpréter les 4 catégories de surfaces de vraisemblance, 4 configurations d'échantillons de taille 100 gènes ont été choisies. Chacune a été prélevée dans un pool simulé en supposant l'un des trois scénarios démographiques précédents, et donne une beaucoup plus forte vraisemblance à ce scénario qu'aux deux autres (illustration partielle au tableau 6.1). La configuration $S = \{10, 8, 12, 2, 53, 11, 4\}$, simulée en supposant le scénario de stabilité, comporte 7 allèles et elle est trimodale, sans allèles manquants : tous les nombres de répétitions entre les deux extrêmes sont représentés. La configuration $E = \{1, 98, 1\}$, simulée en supposant le scénario de croissance exponentielle, comporte un allèle fortement majoritaire, et deux allèles contigus en une seule copie. Les deux allèles minoritaires ont été obtenus par mutation sur une branche terminale de la généalogie. La configuration $D = \{37, 0, 63\}$, simulée en supposant le scénario de fondation-explosion, comporte deux allèles fréquents, non consécutifs, qui tous deux étaient présents à la fondation. Enfin, la configuration $\mathcal{F} = \{55, 1, 0, 0, 0, 0, 1, 42, 1\}$ comporte deux allèles fréquents non-consécutifs fondateurs, et 3 allèles contigus en faible fréquence, des mutants post-fondation.

Configuration	C pour le scénario		
	stable	exponentiel	fondation
$S = \{10, 8, 12, 2, 53, 11, 4\}$	0	0	0
$E = \{1, 98, 1\}$	0	31190	11065
$D = \{37, 0, 63\}$	0	0	10
$F = \{55, 1, 0, 0, 0, 0, 1, 42, 1\}$	0	0	0

TAB. *6.1: La vraisemblance des échantillons de configurations S, E, D et F est estimée par échantillonnage de Monte Carlo direct, pour trois scénarios démographiques tranchés : un scénario de stabilité ($a = r = 1$), un scénario de croissance exponentielle ($a = r = 100$, $t_f = 0.05$) et un scénario de fondation ($a = 1$, $r = 100$ et $t_f = 0.05$). Pour cela, 10^7 échantillons de taille $n = 100$ ont été simulés pour chacun des trois scénarios, et le nombre C de configurations identiques à S, E, D et F a été compté. Pour l'échantillon E, le scénario le plus vraisemblable est le scénario de croissance exponentielle, et le scénario de fondation est seulement trois fois moins vraisemblable. Pour l'échantillon D, seul le scénario de fondation a permis de simuler la configuration cible, et la vraisemblance est au moins 10 fois plus forte pour le scénario de fondation que pour les autres. Les échantillons S et F n'ont été simulés aucune fois sur 10^7 itérations (la méthode de Monte Carlo standard montre encore une fois ses limites). On peut avancer que la configuration S a plus de chances d'être simulée pour le scénario de stabilité et la configuration F plus de chances d'être simulée pour le scénario de fondation, simplement d'après le nombre d'allèles et l'espacement entre les modes. On pourrait le vérifier par échantillonnage d'importance [55, 138], ou en utilisant une méthode d'acceptation-rejet [107, 39].*

Loi *a posteriori* caractéristique d'une histoire de stabilité

La figure 6.6 illustre la loi *a posteriori* de type P_s ici inférée à partir de l'échantillon S. À la page suivante, cette figure est reproduite en affectant des couleurs aux points de l'échantillon MCMC, selon les valeurs de certains paramètres.

La loi jointe de $\log_{10}(r)$ et $\log_{10}(a)$ (visible sur le cadran **d**) est approximativement uniforme, sans même une plus forte densité autour des valeurs utilisées pour simuler les données, c'est à dire $a = r = 1$. La considération du paramètre de date t_f est plus informative (voir cadrans **a**,**b** et **c**). Une valeur charnière est $log_{10}(t_f) \simeq 1$: grosso modo, pour t_f plus grand que ce seuil (zone **A**), le MRCA est plus récent que t_f (en fait, la frontière entre généalogies ayant 1 gène fondateur ou plus dépend de a et r, *cf.* coloration des points échantillonnés en fonction du nombre de lignées fondatrices). La loi *a posteriori* donne plus de poids (probabilité 0.61) au cas d'une seule lignée à t_f, et donc à de grandes valeurs de t_f (figure 6.4), puisque t_f et le nombre de lignées non coalescées à t_f sont négativement corrélés. Pour ces points (zone **A**) , comme le MRCA de l'échantillon est plus récent que t_f, l'échantillon ne contient aucune information sur la taille de la population ancestrale (c'est à dire sur a). Plus étonnant à première vue, on n'a aucune information non plus sur la valeur de r. On peut le comprendre en remarquant qu'une date t_f reculée correspond de toute façon à une relative stabilité de la population depuis cette date (même pour les valeurs extrêmes de r, la croissance ou la décroissance, qui sont étalées sur une longue période, ne sont pas fortes). Cette pseudo-stabilité n'est compatible avec la configuration de l'échantillon que pour un taux de mutation dans un intervalle étroit (zone **A** sur le cadran **c**) : l'intervalle de plus forte densité *a posteriori* de probabilité 95% pour $\log_{10}(\theta)$ est $[0.23, 1.10]$, conditionnellement au fait que $\log_{10} t_f > 1$ (figure 6.4). Cela correspond à une probabilité de 95% pour que θ soit dans l'intervalle $[1.69, 12.6]$ (intervalle qui contient la valeur utilisée pour la simulation, $\theta = 10$). Pour des dates de transition démographique suffisamment récentes ($\log_{10}(t_f) < 0$, zone **B**), il y a une incompatibilité du jeu de données avec les valeurs conjointement positives de r et de a, sauf pour les événements très récents (trou de densité sur les cadrans **a** et **b**). L'identification des points d'échantillonnage en fonction du signe de $\log_{10}(r)$, de $\log_{10}(a)$, et de $\log_{10}(r) - \log_{10}(a)$ permet d'identifier des corrélations. La plus marquante est la relation linéaire entre $\log_{10}(a)$ et $\log_{10}(\theta)$, conditionnellement à $\log_{10}(t_f) < 0$ (zone **B** sur le cadran **f**. Intuitivement : pour obtenir la variabilité du jeu de données avec une population ancestrale de petite taille, il faut un fort taux de mutation.

Nous l'avons vu, la majorité des points échantillonnés correspondent à un t_f si ancien que la généalogie de l'échantillon ne reflète que la période plus récente que t_f (zone **A**). Dans ce cas, toute combinaison de a et de r conduit à des généalogies peu déformées par rapport au coalescent standard. Le cas $a \simeq r \simeq 1$ correspond lui aussi à une stabilité, quelle que soit la valeur de t_f (c'est cette situation qui a servi à générer les données). Pourquoi alors la loi *a posteriori* donne-t-elle plus de poids à la situation avec t_f grand ? La réponse vient de la considération de la loi *a priori*. Cette loi a été choisie uniforme sur un domaine D_p de dimension 4 : le $\log_{10}$ de chacun des paramètres a, r, t_f et θ est *a priori* uniforme sur $[-5, 5]$. Dans le domaine qui définit une stabilité avec t_f grand, a et r peuvent prendre n'importe quelle valeur, et θ est dans un intervalle restreint. Le sous-domaine de D_p qui correspond à cette forme de stabilité a donc deux dimensions complètes (r et a), une demi-dimension ($t_f > 1$) et un intervalle étroit pour la dimension qui correspond à θ. Par comparaison, le domaine qui définit la stabilité avec $a \simeq 1$ et $r \simeq 1$ est petit : seul t_f est libre de varier. La loi *a priori* donne donc plus de poids à la stabilité avec t_f grand qu'à la stabilité avec a et r proches de 1, ce qui se répercute sur la loi *a posteriori*.

In fine, pour le jeu de données monolocus S généré en population stable, la méthode donne l'essentiel de la probabilité à une histoire de changement monotone sur le long terme, sans préférence pour un déclin ou une croissance au long cours (zone **A***).*

Pour un jeu de données de 5 loci simulés en supposant ce même scénario de stabilité, les points échantillonnés se distribuent selon deux zones, une zone avec t_f grand et une seule lignée ancestrale à t_f (équivalente à la zone **A**), et une zone —qui se présente comme un secteur angulaire sur la représentation de t_f en fonction de r— avec t_f petit et plus de 10 lignées fondatrices (équivalente à la zone **B**). Ces deux zones sont séparées par des zones de densité virtuellement nulle (aucun point échantillonné sur 20 000), et le seul raccordement correspond à des généalogies avec 2 à 10 lignées fondatrices, obtenues avec $r \simeq a \simeq 1$. Ce sont donc des généalogies peu déformées par rapport au cas stable. La proportion de points échantillonnés avec t_f grand et une seule lignée à t_f est plus grande qu'avec l'échantillon monolocus S (76% des points). Il serait intéressant de vérifier si cette tendance se vérifie sur un plus grand nombre de jeux de données, et si cette proportion tend vers 1 lorsque le nombre de loci augmente. Autrement dit, détecte-t-on de façon certaine que le scénario ne comporte pas de transition démographique, pour un grand jeu de données ?

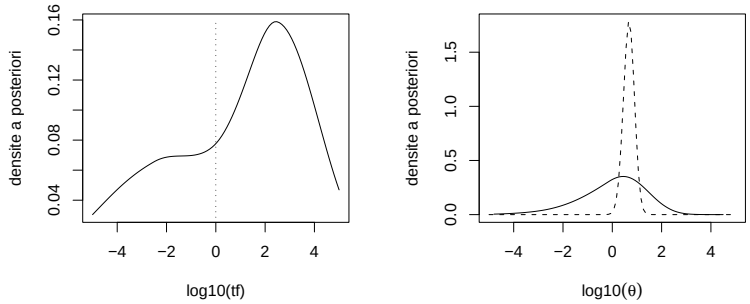


FIG. *6.4: À gauche : densité marginale* a posteriori *de* $\log_{10} t_f$*, estimée à partir d'un échantillon MCMC de* 20 000 *points. Près de* 65% *des points de l'échantillon ont une valeur de* $\log_{10} t_f$ *positive. Les points correspondant à des généalogies avec une seule lignée fondatrice sont tous avec* $\log_{10} t_f > 0$*, et représentent* 61% *du total des points. À droite en traits pleins, densité marginale* a posteriori *de* $\log_{10}(\theta)$*. En traits pointillés, même chose conditionnellement à* $\log_{10}(t_f) > 1$.

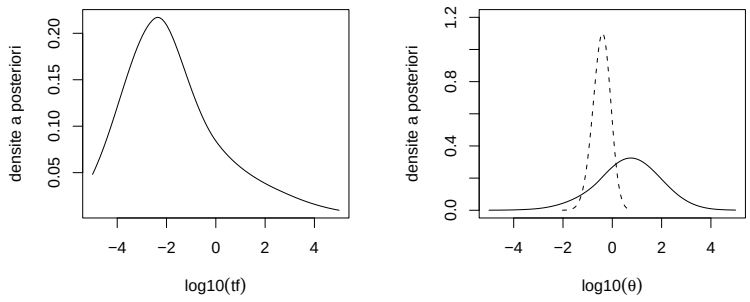


FIG. *6.5: À gauche : densité marginale* a posteriori *de* $\log_{10} t_f$*. 18% des points de l'échantillon ont une valeur de* $\log_{10(} t_f)$ *positive. À droite en traits pleins, densité marginale* a posteriori *de* $\log_{10}(\theta)$*. En traits pointillés, même chose conditionnellement à* $\log_{10}(t_f) > 1$.

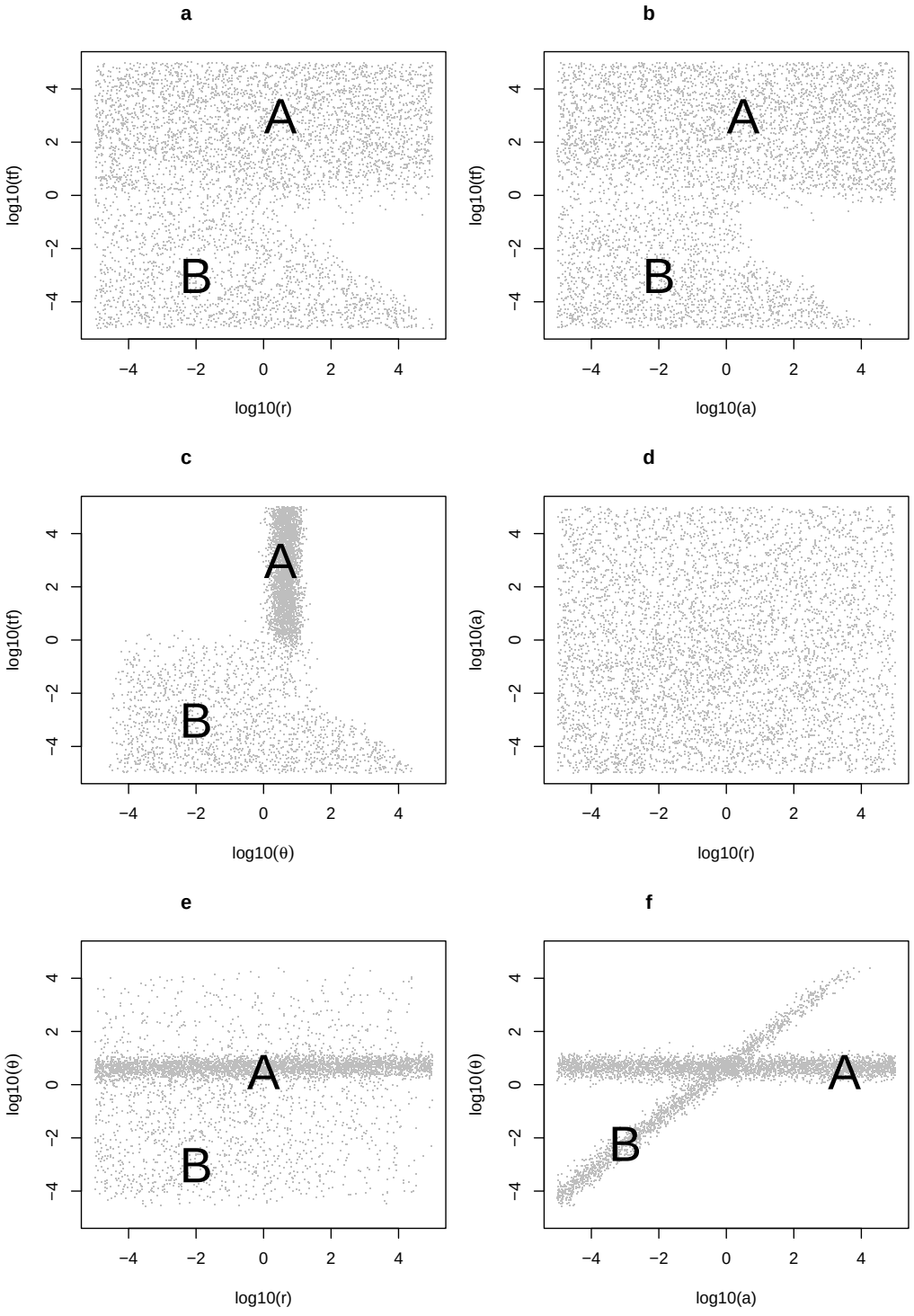


FIG. *6.6: Échantillon de* 5 000 *points représentatif de la loi jointe* a posteriori *des paramètres* r, a, t_f *et* θ, *pour un jeu de données de configuration* $S = \{10, 8, 12, 2, 53, 11, 4\}$. *La loi* a priori *est uniforme sur* $[-5; 5]$, *pour le* $\log_{10}$ *de chacun des* 4 *paramètres, si bien que la loi jointe* a posteriori *coincide avec la vraisemblance sur le domaine de densité uniforme* a priori.

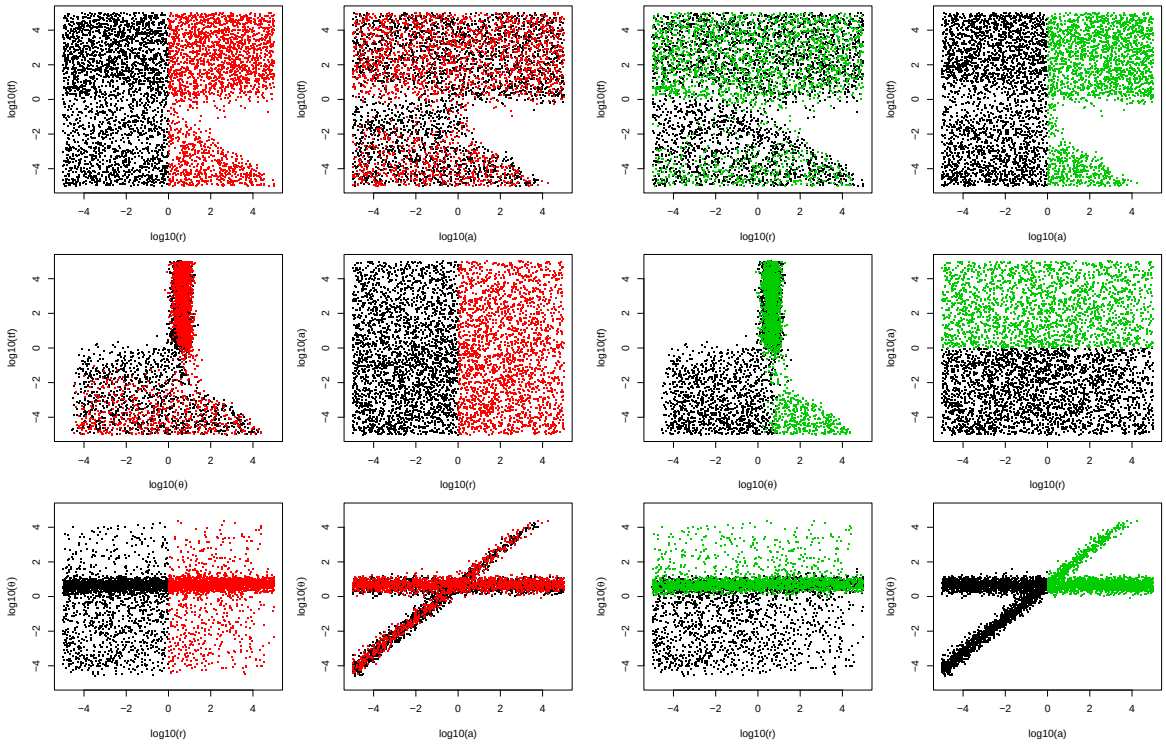


FIG. *6.7: Duplicata de la figure 6.6, avec une coloration différente des points échantillonnés, selon le signe de* $\log_{10}(r)$ *(à gauche) et de* $\log_{10}(a)$ *(à droite)*

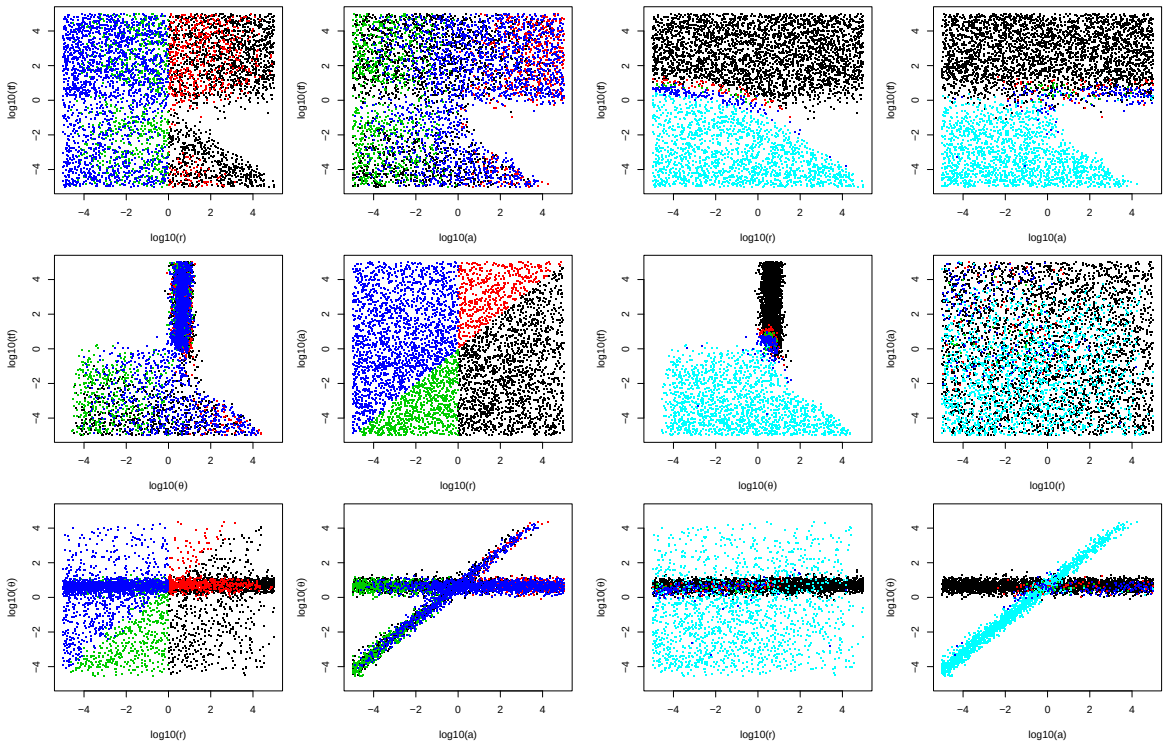


FIG. *6.8: Duplicata de la figure 6.6. À gauche, coloration selon le signe de* $\log_{10}(r)$ *et de* $\log_{10}(r) - \log_{10}(a)$*. À droite, coloration selon le nombre de lignées non coalescées à* t_f *(noir/rouge/vert/bleu/ciel :* $1/2/3/4\text{-}10/\geq 11$ *lignées fondatrices).*

Loi *a posteriori* caractéristique d'une histoire de croissance exponentielle

La figure 6.9 illustre la loi *a posteriori* de type P_e, obtenue pour un échantillon de configuration $\{1, 98, 1\}$ généré en supposant le scénario d'explosion démographique. Comme dans le cas stable et pour la même raison, les valeurs de t_f anciennes ($\log_{10}(t_f) > 1$) correspondent à des valeurs quelconques et non corrélées de a et r, et à un taux de mutation renormalisé θ dans un intervalle restreint (zone **A** sur les cadrans **a**, **b** et **c**). Par rapport à ce que l'on a observé pour l'échantillon S, l'intervalle de plus forte densité *a posteriori* est décalé vers les faibles valeurs de θ (probabilité 95% pour $\log_{10}(\theta)$ de se trouver dans $[-1.16, 0.27]$ soit $[0.07, 1.86]$ pour θ). Cela est dû à la faible diversité de l'échantillon $E = \{1, 98, 1\}$, comparée à celle de $S = \{10, 8, 12, 2, 53, 11, 4\}$. La faible diversité s'explique en fait —on le sait pour avoir simulé l'échantillon— par une taille de population ancestrale faible, et une explosion démographique récente. Les histoires de stabilité définies par t_f grand sont donc associées à des valeurs de θ fortement sous-estimées (pour obtenir cette configuration quasi monomorphe en population stable, il faudrait une faible valeur de θ). Les points d'échantillonnage pour lesquels $\log_{10}(t_f) < 0$ (zones **C** et **D**) sont en proportion beaucoup plus grande que dans le cas stable (figure 6.5) : l'échantillon E est fortement en faveur d'un événement démographique récent. Deux zones de forte densité sont à distinguer pour $\log_{10}(t_f) < 0$. L'une (zone **C**) correspond à des valeurs de a positives, et à n'importe quelle valeur de r, c'est à dire à des scénarios avec une population de taille ancestrale faible. La valeur de $\log_{10}(\theta)$ pour les points échantillonnés dans cette zone remplit l'intervalle $[-1; 3]$ (zone **C** sur les cadrans **e** et **f**) , avec une forte corrélation négative entre $\log_{10}(t_f)$ et $\log_{10}(\theta)$ (zone **C** sur le cadran **c**) : plus l'augmentation de taille de population est récente, plus fort doit être le taux de mutation pour générer les deux allèles mutants. La seconde zone remarquable (zone **D**) correspond à des fondation-explosion récentes. Là aussi, le taux de mutation est borné, dans $[0; 2]$ pour le logarithme décimal. Dans la zone **C**, avec $\log_{10}(a) > 0$ le nombre de lignées non coalescées à t_f pouvait être grand : les coalescences restant à faire étaient en effet forcément très concentrées juste avant t_f, du fait de la petite taille de population ancestrale. Dans les cas des fondation-explosion (zone **D**), l'événement de fondation doit être suffisamment drastique pour que l'échantillon E soit dérivé d'une seule lignée fondatrice, ou de quelques-unes, sans doute de même état allélique.

En conclusion, un échantillon généré en supposant une histoire de croissance exponentielle soutient fortement des valeurs positives de a (zone **C***), c'est à dire des scenarios dans lesquels la taille de la population ancestrale est beaucoup plus faible que celle de la population échantillonnée. Toutefois, à peu près n'importe quel scenario à t_f est possible, tant que a est positif.*

Loi *a posteriori* caractéristique d'une histoire de déclin

La figure 6.12 illustre la loi *a posteriori* de type P_d, ici pour un échantillon de configuration $D = \{37, 0, 63\}$ généré en supposant le scénario de fondation-explosion, mais sans mutations post-fondation (on n'a donc que les signatures du déclin et pas celles de l'explosion). Comme pour les échantillons S et E et pour la même raison, les valeurs de t_f anciennes (zone **A**) correspondent à des valeurs quelconques et non corrélées de a et r, et à un taux de mutation renormalisé θ dans un intervalle restreint (t_f ancienne signifie toutefois ici $\log_{10}(t_f) > 1.5$ et non pas $\log_{10}(t_f) > 1$: le jeu de données donne du poids à l'existence de plusieurs lignées fondatrices). Comme pour l'échantillon E, l'événement $\log_{10}(t_f) > 1.5$ a une faible probabilité *a posteriori*, et l'intervalle de plus forte densité *a posteriori* pour $\log_{10}(\theta)$, conditionnellement à $\log_{10}(t_f) > 1.5$, sous-estime fortement la valeur de θ. Chose nouvelle par rapport aux types P_s et P_e, une bande de forte densité contenant 81% des points échantillonnés est observable sur la représentation bivariée de r et t_f, avec une corrélation négative entre ces deux paramètres (zones **E** et **F** sur le cadran **a**). Schématiquement, pour les valeurs de $\log_{10}(r) < -2$, les points situés sur cette bande correspondent à des pic-déclin (zone **E**), puis pour $-2 < \log_{10}(r) < 0$, on a des crash-déclin, et enfin pour $\log_{10}(r) > 0$, on a des fondation-explosion (zone **F**). Dans les trois cas, on n'a pas de borne inférieure nette pour θ (zones **E** et **F** sur le cadran **c**) : il n'y a pas d'allèles mutants, et les deux allèles fondateurs sont de taille proches, ce qui, pour a petit, peut avoir été obtenu avec un faible taux de mutation. On a en revanche une borne supérieure pour θ. Cette borne supérieure est négativement corrélée à t_f (zone **F** sur le cadran **c**).

L'échantillon D soutient donc fortement un déclin de population en marche d'escalier, suivi ou non d'explosion.

Loi *a posteriori* caractéristique d'une histoire de fondation-explosion

Enfin, la figure 6.15 illustre la loi *a posteriori* de type P_f obtenue pour un échantillon de configuration $F = \{55, 1, 0, 0, 0, 0, 1, 42, 1\}$ généré en supposant le scénario de fondation-explosion. La différence majeure avec le cas précédent concerne le poids relatif des pic-déclin et crash-déclin d'une part (zone **G**), et des fondation-explosion d'autre part (zone **H**), au niveau de la bande de forte densité observable sur le cadran **a** (loi bivariée de t_f et de r) : les points d'échantillonnage qui correspondent à une fondation-explosion (zone **H**) sont beaucoup plus nombreux (87% du total des points). Dans le même temps, les fondation-explosion sont associées à des valeurs de θ dans une gamme restreinte, avec une corrélation négative entre t_f d'une part, et θ et r d'autre part (zone **H** sur les cadrans **c** et **e**) : plus l'explosion a commencé récemment, plus la présence de trois allèles mutants nécessite que θ et r soit grands.

On notera que toutes les configurations d'échantillon donnant une loi a posteriori *de type P_f ne sont pas si démonstratives, mais la configuration F illustre le fait qu'un unique échantillon puisse être très fortement en faveur d'une fondation-explosion.*

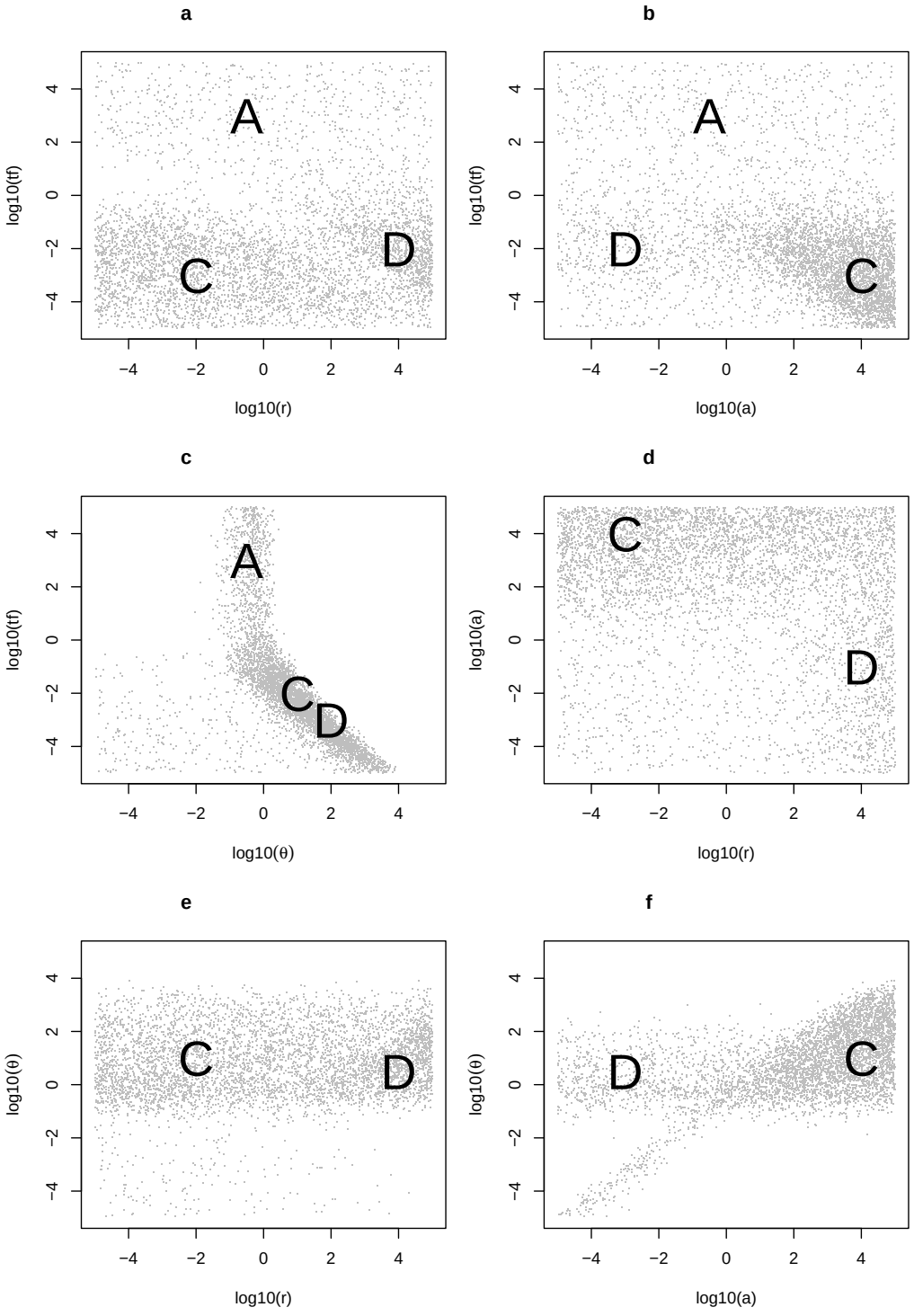


FIG. *6.9: Loi jointe* a posteriori *des paramètres* r, a, t_f *et* θ, *inférée pour un jeu de données de configuration* $\{1, 98, 1\}$, *simulé en supposant un scénario de croissance exponentielle, sous SMM, avec* $\theta = 10$, $r = a = 100$ *et* $t_f = 0.05$. *La loi a priori est uniforme sur* $[-5; 5]$, *pour le* $\log_{10}$ *de chacun des* 4 *paramètres.*

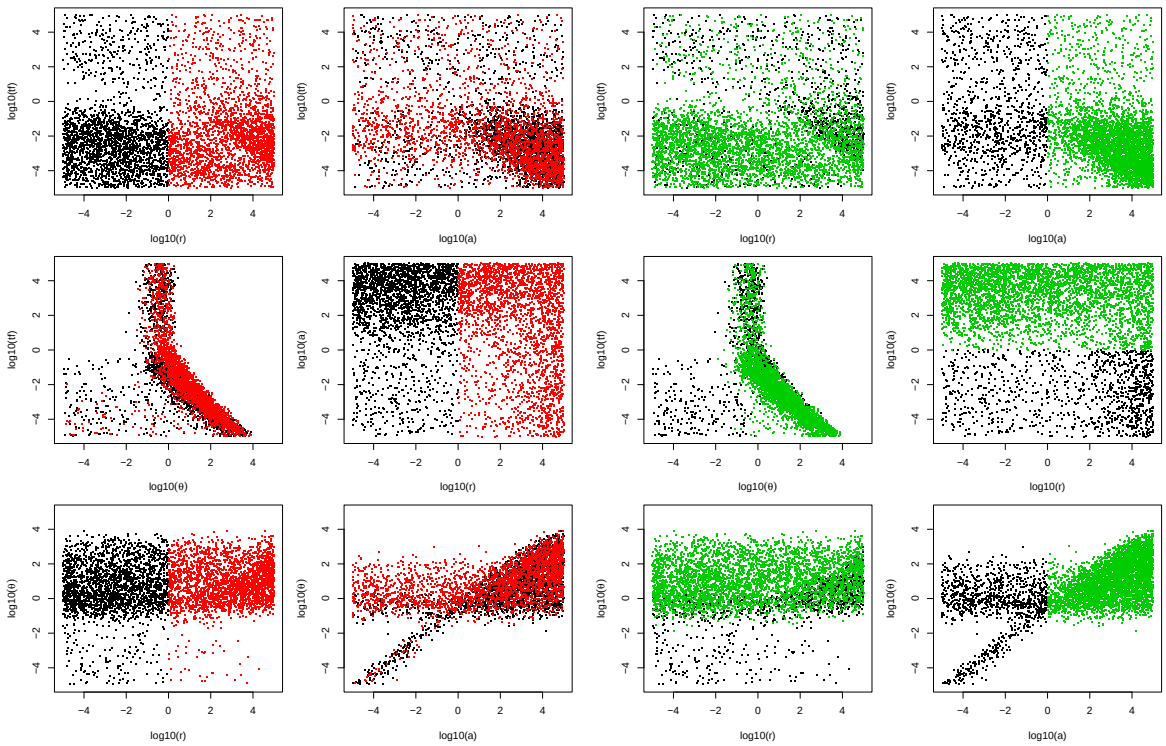


FIG. *6.10: Duplicata de la figure 6.9, avec une coloration différente des points échantillonnés, selon le signe de* $\log_{10}(r)$ *(à gauche) et de* $\log_{10}(a)$ *(à droite)*

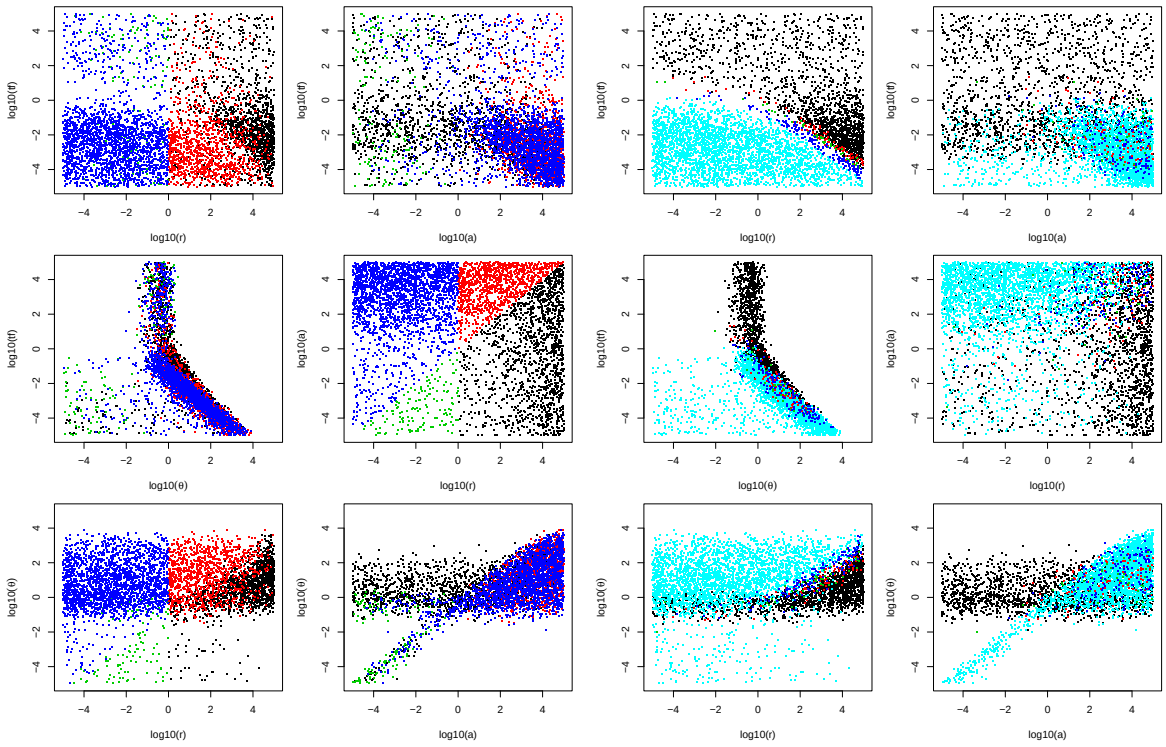


FIG. *6.11: Duplicata de la figure 6.9. À gauche, coloration selon le signe de* $\log_{10}(r)$ *et de* $\log_{10}(r) - \log_{10}(a)$*. À droite, coloration selon le nombre de lignées non coalescées à* t_f *(noir/rouge/vert/bleu/ciel :* $1/2/3/4\text{-}10/\geq 11$ *lignées fondatrices).*

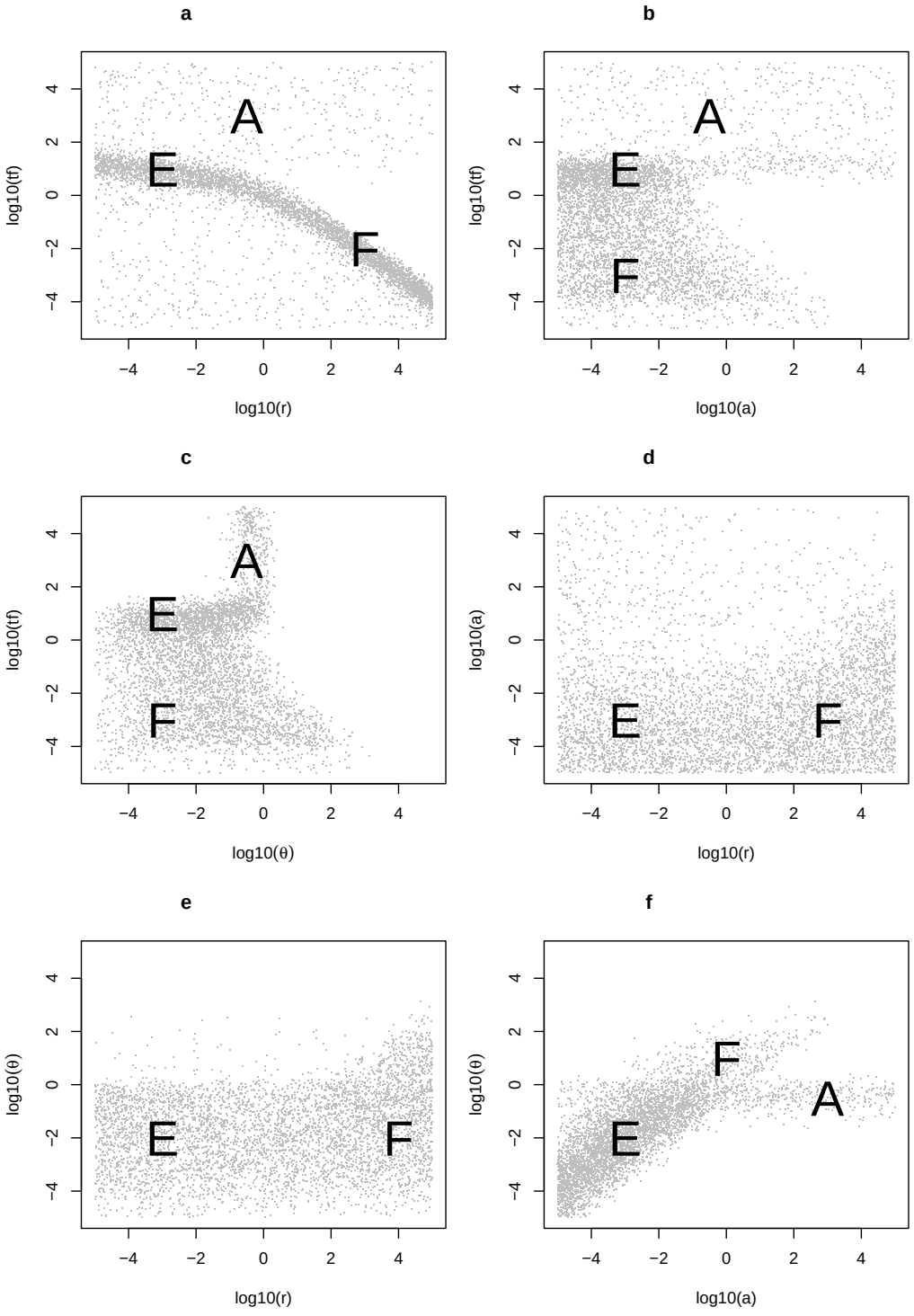


FIG. *6.12: Loi jointe* a posteriori *des paramètres* r, a, t_f *et* θ, *inférée pour un jeu de données de configuration* $\{37, 0, 63\}$, *simulé en supposant un scénario de fondation-explosion, sous SMM, avec* $\theta = 10$, $r = 100$, $a = 1$ *et* $t_f = 0.05$. *La loi* a priori *est uniforme sur* $[-5; 5]$, *pour le* $\log_{10}$ *de chacun des* 4 *paramètres.*

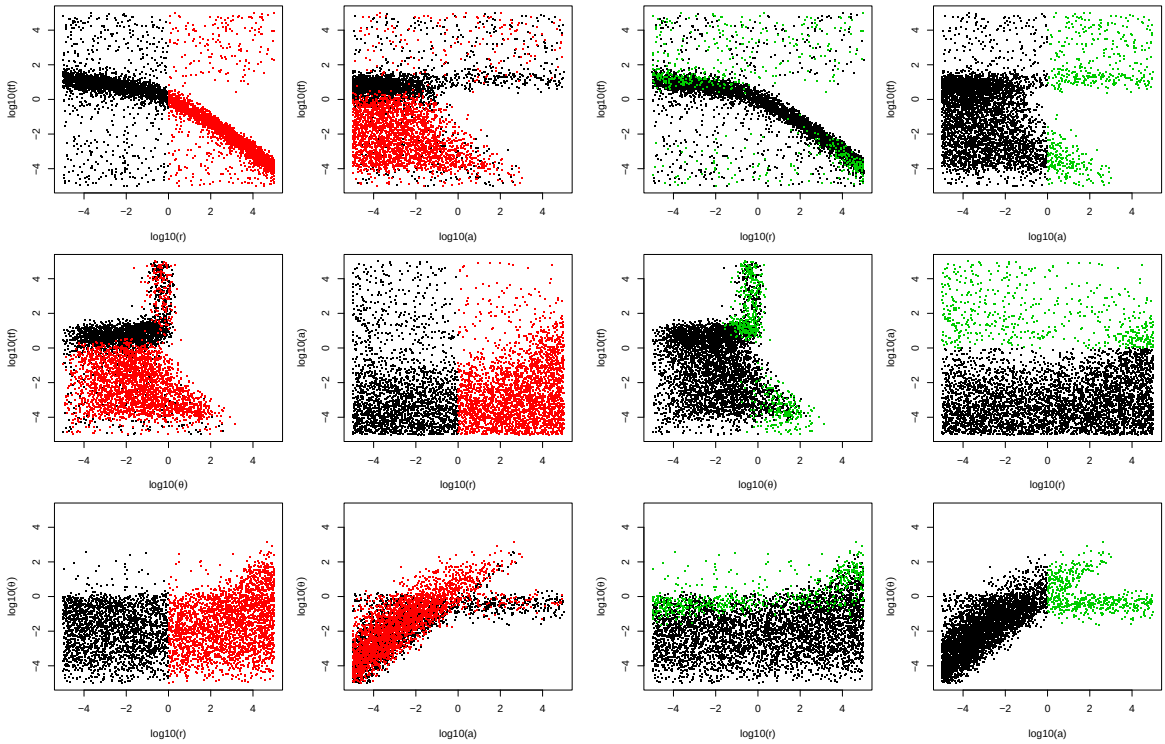


FIG. *6.13: Duplicata de la figure 6.12, avec une coloration différente des points échantillonnés, selon le signe de* $\log_{10}(r)$ *(à gauche) et de* $\log_{10}(a)$ *(à droite)*

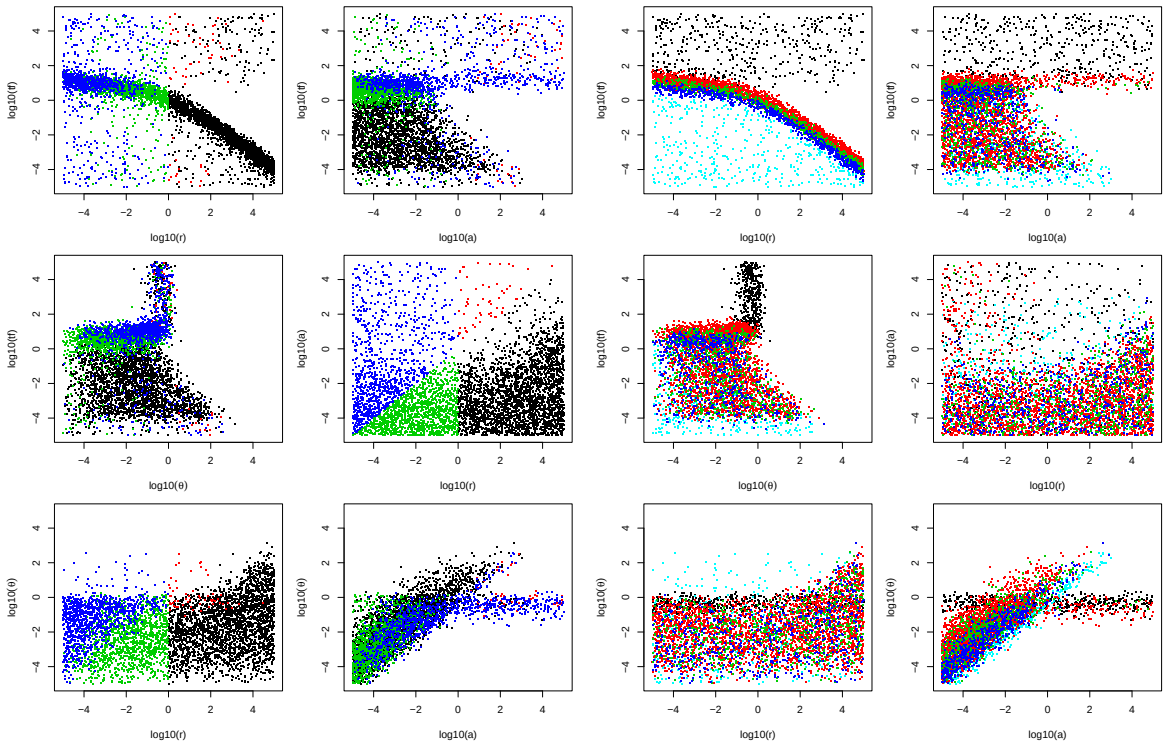


FIG. *6.14: Duplicata de la figure 6.12. À gauche, coloration selon le signe de* $\log_{10}(r)$ *et de* $\log_{10}(r) - log_{10}(a)$*. À droite, coloration selon le nombre de lignées non coalescées à* t_f *(noir/rouge/vert/bleu/ciel :* $1/2/3/4\text{-}10/\geq 11$ *lignées fondatrices).*

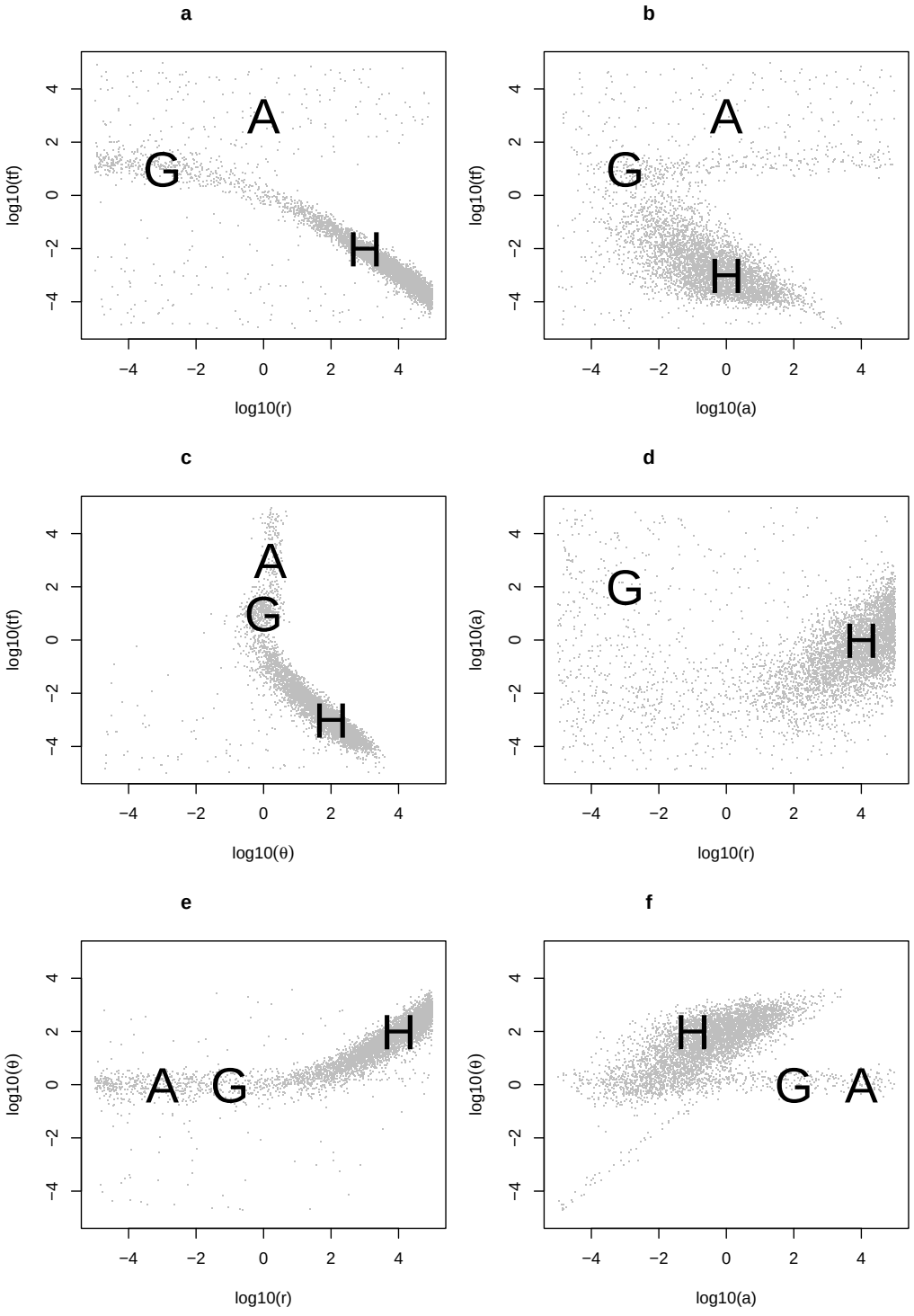


FIG. *6.15: Loi jointe* a posteriori *des paramètres* r, a, t_f *et* θ, *inférée pour un jeu de données de configuration* $\{55, 1, 0, 0, 0, 0, 1, 42, 1\}$, *simulé en supposant un scénario de fondation-explosion, sous SMM, avec* $\theta = 10$, $r = 100$, $a = 1$ *et* $t_f = 0.05$. *La loi* a priori *est uniforme sur* $[-5; 5]$, *pour le* log_{10} *de chacun des* 4 *paramètres.*

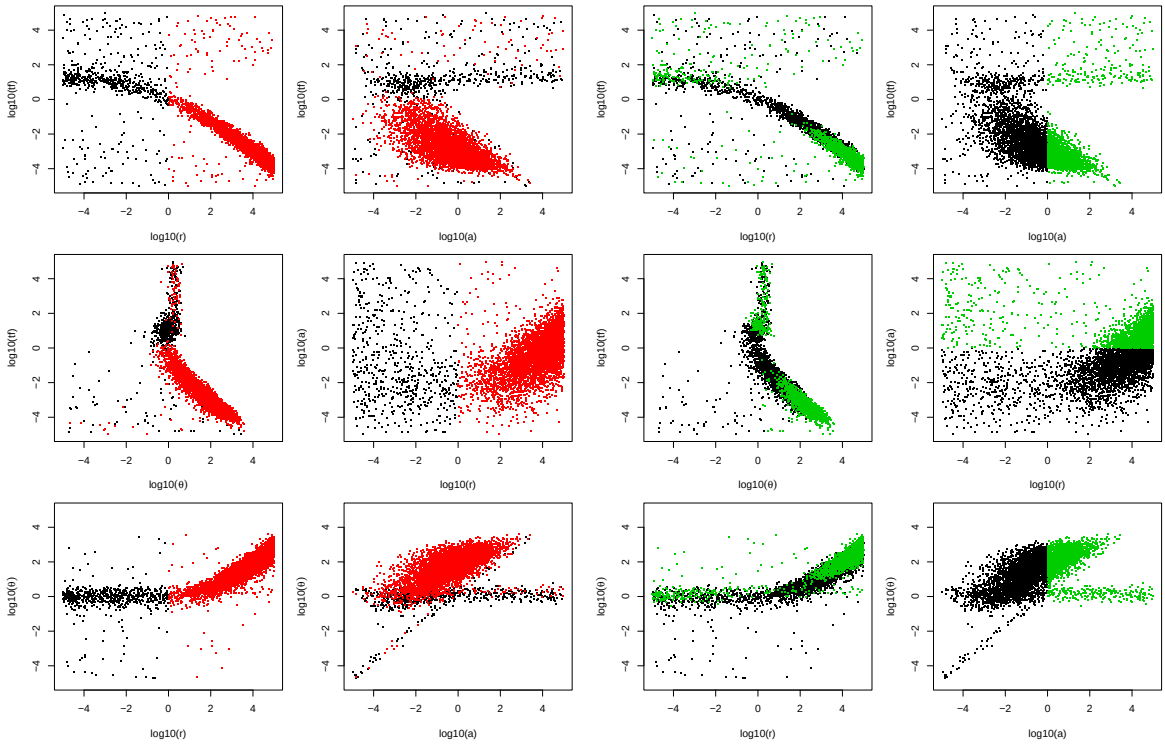


FIG. *6.16: Duplicata de la figure 6.15, avec une coloration différente des points échantillonnés, selon le signe de* $log_{10}(r)$ *(à gauche) et de* $log_{10}(a)$ *(à droite)*

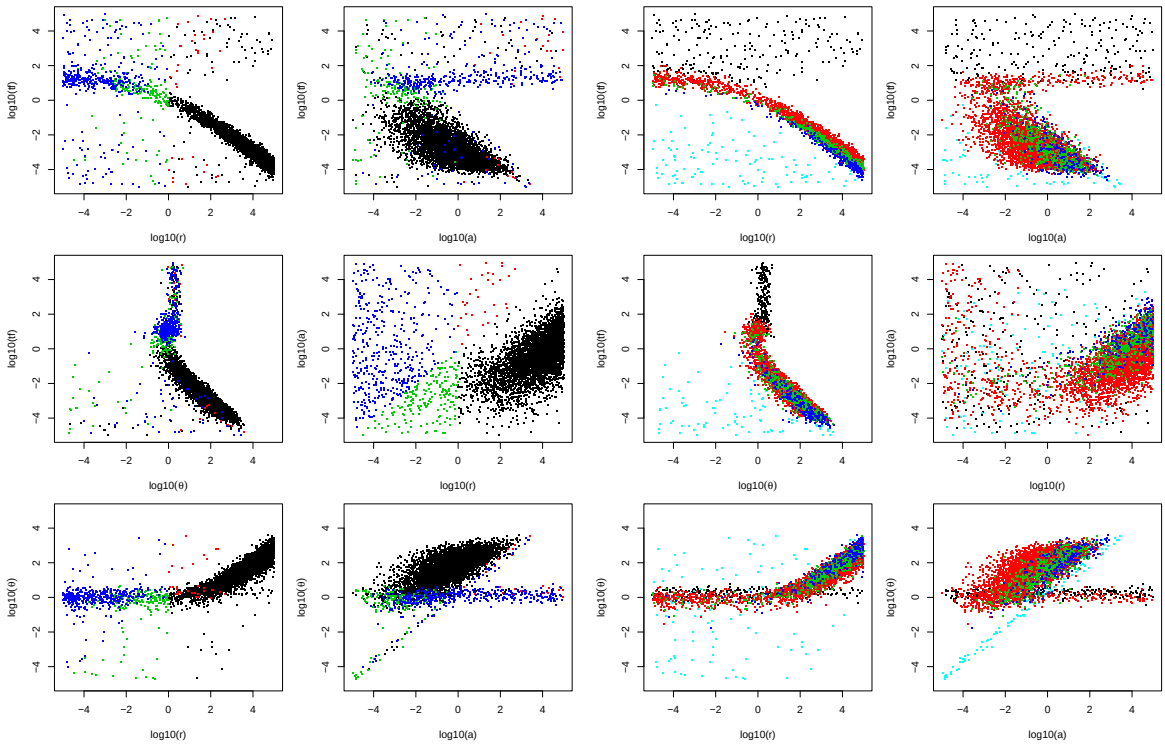


FIG. *6.17: Duplicata de la figure 6.15. À gauche, coloration selon le signe de* $log_{10}(r)$ *et de* $log_{10}(r) - log_{10}(a)$*. À droite, coloration selon le nombre de lignées non coalescées à* t_f *(noir/rouge/vert/bleu/ciel :* $1/2/3/4\text{-}10/\geq 11$ *lignées fondatrices).*

6.2 Précision des inférences sur une fondation-explosion

Afin de se faire une idée de la répartition optimale de l'effort de typage dans les études empiriques, on compare la précision des inférences obtenue pour des gammes de variation de certaines caractéristiques des jeux de données D, simulés sous le modèle démographique de fondation-explosion. Le modèle mutationnel est SMM aussi bien pour la genèse des données que pour les inférences.

6.2.1 Effet de la taille d'échantillon monolocus

Un jeu de données monolocus généré par une histoire de fondation-explosion ne contient d'information sur la fondation que si le MRCA de l'échantillon est antérieur à cet événement (c'est à dire si l'on a au moins deux lignées fondatrices, non coalescées à la date t_f de la fondation). On peut se demander dans quelle mesure le fait d'élargir l'échantillon augmente la probabilité pour que celui-ci soit dérivé d'au moins deux lignées fondatrices. Un résultat analytique concernant le coalescent nous informe sur ce point [125] : la probabilité pour que le MRCA d'un échantillon de taille n soit le MRCA de la population entière est $(n-1)/(n+1)$. Cette probabilité se rapproche très rapidement de 1 lorsque n augmente, et varie donc peu en fonction de n, dès que n dépasse quelques dizaines de gènes. D'après ce résultat, rien ne sert —pour détecter une fondation— de typer des échantillons de grande taille. En effet, comme l'illustre la figure 6.18, la taille d'échantillon change très peu la distribution du nombre de lignées non coalescées à la fondation. Mieux vaut, pour augmenter ses chances de trouver les signatures génétiques d'une éventuelle fondation, augmenter le nombre de loci, pour disposer de plusieurs tirages selon la loi de la date du MRCA. Pour une histoire de fondation-explosion, on augmente alors nos chances d'avoir des loci avec plusieurs lignées fondatrices, et donc potentiellement avec plusieurs allèles fondateurs. Une grande taille d'échantillon augmente en revanche fortement la probabilité d'observer des allèles en faible fréquence, obtenus par mutation post-fondation. La figure 6.19 illustre ce point en montrant la distribution du nombre de mutations post-fondation, pour des n-coalescents avec n croissant.

Dans la pratique, on est rapidement limité par le temps de calcul, et au delà d'une dizaine de loci et d'une centaine de gènes, on est confronté à de graves problèmes de convergence de la MCMC.

Pour illustrer l'effet de la taille d'échantillon sur la précision des inférences, on a simulé un jeu de données de taille $n = 200$, dont on a extrait par tirage uniforme sans remise un échantillon de chacune des tailles $n = 100$, 50, 30 et 10. Les 5 configurations obtenues sont données dans le tableau 6.2. Les valeurs des paramètres utilisées pour la genèse des données sont les mêmes que précédemment ($r = 100$, $a = 1$, $t_f = 0.05$, $\theta = 10$).

Les échantillons de taille $n = 10$ et 30, qui sont tout à fait du même type que la configuration $D = \{37, 0, 63\}$, avec deux allèles fondateurs non consécutifs, donnent logiquement des lois a posteriori de type P_d (seul le déclin a des signatures détectables dans les petits échantillons). Les échantillons de taille $n = 50$, 100 (configuration F illustrée précédemment) et 200, qui comportent des mutations post-fondation, donnent tous les trois des lois *a posteriori* de type P_f (on a à la fois les signatures de la fondation —2 gènes fondateurs distants— et celles de l'explosion —des mutants post-fondation. Plus que de la taille d'échantillon, les inférences semblent donc dépendre de la distribution des allèles fondateurs et mutants. Si cela est vrai, on peut prédire qu'un échantillon de taille par exemple 30 ayant l'un des allèles mutants devrait conduire à une loi *a posteriori* de type P_f. Pour le vérifier, de nouveaux échantillons de taille 30

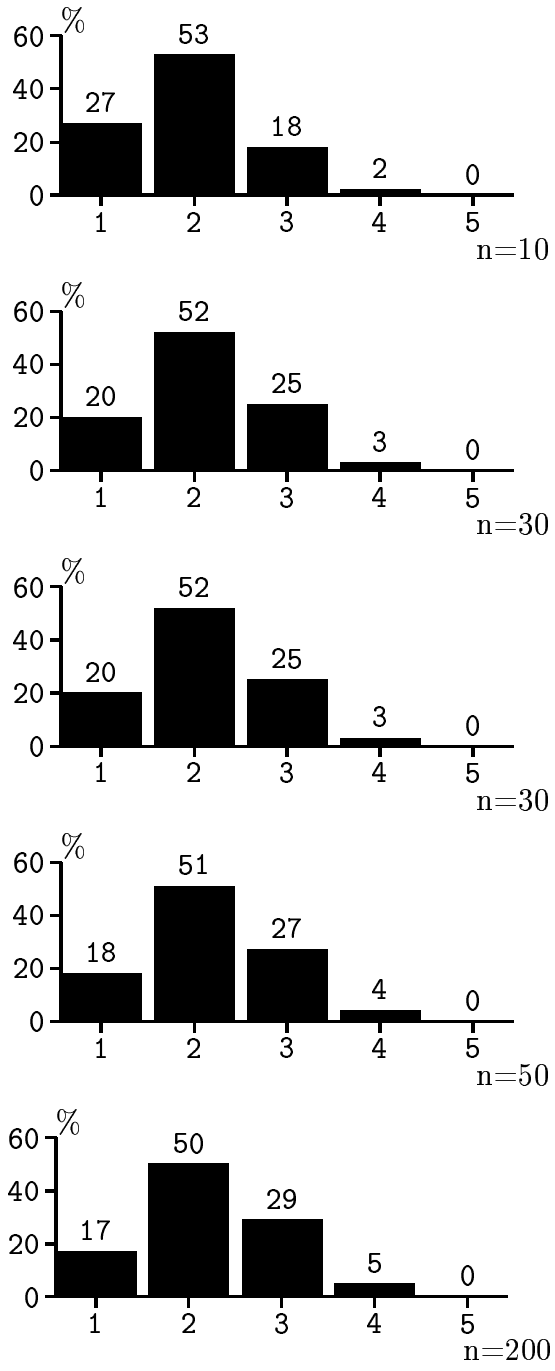


FIG. *6.18: Effet de la taille d'échantillon n sur le nombre de lignées fondatrices. Des coalescents sont générés comme décrit au paragraphe 5.1, en supposant une histoire de fondation-explosion avec* $r = 100$, $a = 1$ *et* $t_f = 0.05$. *On compare la distribution du nombre* f *de lignées ancestrales de l'échantillon encore présentes à la date* t_f, *pour des tailles d'échantillons croissantes (*$n = 10$, 30, 50, 100, 200*). Des tailles de quelques dizaines d'individus typés sont suffisantes pour remonter aux coalescences les plus profondes dans la généalogie de la population totale. Les inférences sur la taille de la population ancestrale doivent donc assez peu dépendre de la taille d'échantillon.*

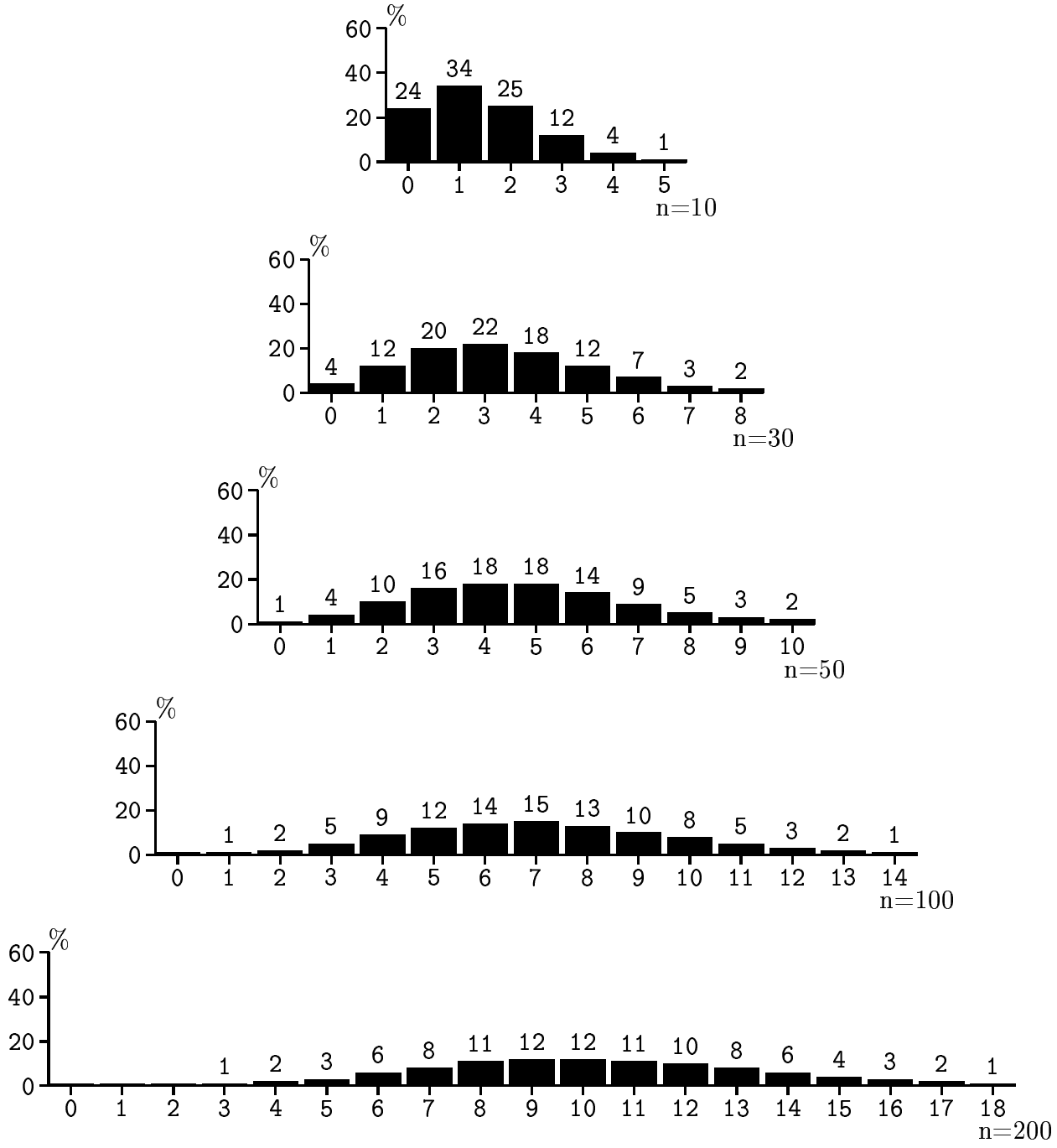


FIG. *6.19: Effet de la taille d'échantillon n sur le nombre de mutations post-fondation. Des coalescents sont générés dans les mêmes conditions qu'à la figure 6.18. On compare la distribution du nombre m de mutations plus récentes que t_f, pour des tailles d'échantillons croissantes ($n = 10, 30, 50, 100, 200$). Ce nombre moyen augmente fortement avec n. On s'attend donc à pouvoir déterminer plus précisément la phase de croissance exponentielle, pour n grand. Toutefois, quand le nombre de mutations post-fondation croit, l'homoplasie entre allèles mutants est inévitable, surtout sous SMM. L'augmentation de la taille d'échantillon est donc moins rentable au delà d'un seuil —50 gènes suffisent pour le scénario envisagé, à en juger par la loi a posteriori (figure 6.20).*

n	configuration	f	m
200	$\{112, 2, 0, 0, 0, 0, 3, 79, 4\}$	2	3
100	$\{55, 1, 0, 0, 0, 0, 1, 42, 1\}$	2	3
50	$\{27, 0, 0, 0, 0, 0, 1, 21, 1\}$	2	2
30	$\{20, 0, 0, 0, 0, 0, 0, 10\}$	2	0
10	$\{5, 0, 0, 0, 0, 0, 0, 5\}$	2	0

TAB. *6.2: Un échantillon de taille* $n = 200$ *chromosomes a été simulé pour les valeurs des paramètres* $r = 100$, $a = 1$, $t_f = 0.05$ *et* $\theta = 10$, *sous le modèle de mutation SMM. Des sous-échantillons de tailles* 100, 50, 30 *et* 10 *en ont été extraits indépendamment par tirage uniforme sans remise. L'échantillon complet dérive de* $f = 2$ *lignées fondatrices et* 6 *événements mutationnels post-fondation ont fourni* $m = 3$ *allèles nouveaux. Dans les sous-échantillons, les* $f = 2$ *lignées fondatrices sont —c'est le cas le plus fréquent— toujours présentes, alors que le nombre d'allèles mutants est réduit à* $m = 2$ *dans l'échantillon de taille* 50, *puis à* $m = 0$ *dans les échantillons de tailles* 30 *et* 10.

ont été extraits de l'échantillon de taille 200, par tirage sans remise, et le premier contenant au moins un allèle mutant a été choisi. Sa configuration est $\{21, 1, 0, 0, 0, 0, 0, 8\}$. Comme attendu, la loi *a posteriori* obtenue pour cet échantillon est nettement de type P_f. Ce qui différencie les valeurs de n dans la gamme choisie est surtout la probabilité pour que les échantillons de cette taille comportent effectivement au moins un allèle mutant (pour les tailles $n = 10$, 30, 50 et 100 tirés de notre échantillon-école de taille 200, cette probabilité vaut respectivement 0.38, 0.78, 0.93 et 0.99).

*En conclusion, il semble suffisant de traiter des échantillons de quelques dizaines de gènes (*50 *à* 100*) par locus, pour saisir les signatures génétiques des variations d'effectif passées.*

6.2.2 Effet du nombre de loci indépendants

Nous venons de voir que de petits sous-échantillons peuvent comporter l'essentiel de l'information sur une fondation-explosion apportée par des échantillons plus grands, si par chance ils comportent des représentants des classes alléliques mutantes. En traitant en parallèle plusieurs échantillons de petite taille à des loci différents, on augmente la probabilité que certains soient typiques d'une fondation-explosion. La capacité de détecter une fondation-explosion dépendra alors de la loi *a posteriori* pour l'échantillon multilocus : quand on a un *mélange de loci* qui soutiennent une *expansion* —c'est le cas pour les loci dont le MRCA est plus récent que t_f—, un *déclin* —c'est le cas pour les loci ayant plusieurs allèles fondateurs de tailles distantes, mais pas de classes alléliques mutantes—, ou une *fondation-explosion* —c'est le cas des loci ayant à la fois plusieurs allèles distants et des mutants post-fondation—, que soutient l'échantillon multilocus ?

Pour aborder cette question, un jeu de données de 5 loci indépendants et de taille 50 chromosomes a été simulé pour les valeurs des paramètres de fondation-explosion suivantes : $r = 10000$, $a = 1$, $t_f = 0.001$, $\theta = 100$, sous SMM. Le tableau 6.3 donne les configurations normalisées pour les 5 loci, et pour information, le nombre de lignées fondatrices (nombre de lignées non coalescées à t_f) et le nombre de mutations post-fondation dans l'arbre (mutations qui en principe nous informent sur l'expansion de la population).

La figure 6.21 donne des représentations de la loi *a posteriori* des paramètres, pour une loi *a priori* de (r, a, t_f, θ) uniforme sur $[-6; 8] \times [-5; 5] \times [-7; 4] \times [-5; 3]$ et une loi *a priori* ponctuelle

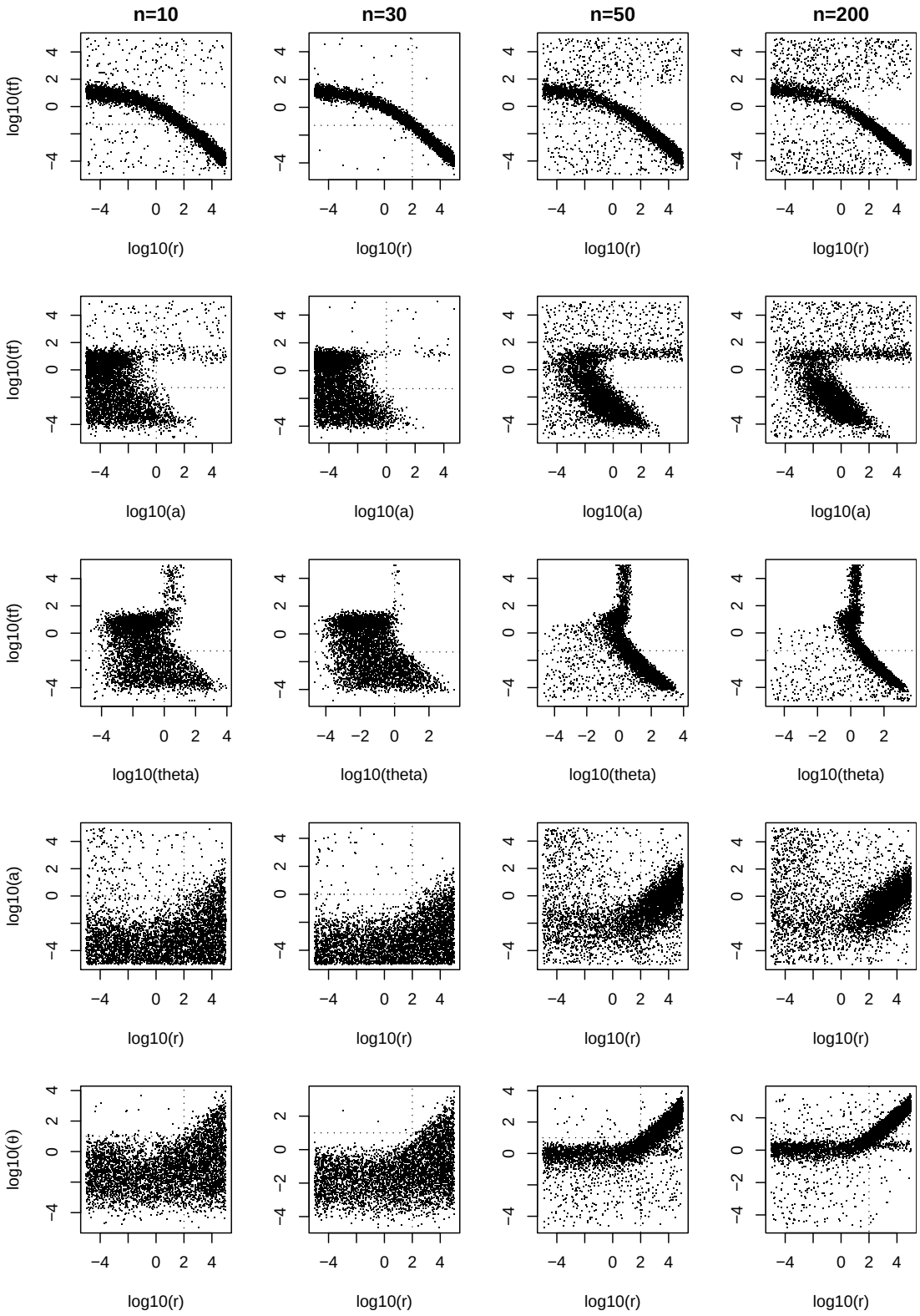


FIG. *6.20: Les échantillons corrélés de configurations présentées au tableau 6.2 ont été utilisés pour mettre à jour des* a priori *uniformes sur* $[-5; 5]$*, pour le* log_{10} *des* 4 *paramètres d'intérêt. Une valeur d'espacement de* 10000 *s'est avérée suffisante pour assurer une bonne convergence dans les* 5 *cas. Les échantillons de taille* 10 *et* 30 *donnent des lois* a posteriori *de type* P_d *et les échantillons de taille* 50 *et* 200 *donnent des lois* a posteriori *de type* P_f *(interprétations dans le texte).*

nom	configuration	lignées fondatrices	mutations post-fondation	type
N_1	$\{49, 1\}$	1	1	P_e
N_2	$\{47, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3\}$	2	0	P_d
N_3	$\{14, 0, 0, 36\}$	2	0	P_d
N_4	$\{1, 22, 1, 1, 24, 1\}$	2	4	P_{sf}
N_5	$\{21, 0, 0, 27, 2\}$	2	1	P_{df}

TAB. *6.3: Cinq échantillons de taille* 50 *chromosomes ont été simulées indépendamment pour les valeurs des paramètres* $r = 10000$, $a = 1$, $t_f = 0.001$ *et* $\theta = 100$, *sous le modèle de mutation SMM. Cela correspond à une fondation-explosion récente et drastique (crash d'un facteur* 10000 *et retour exponentiel à la taille initiale en* $0.001 \times N_0$ *générations. On indique la configuration normalisée pour les* 5 *loci, le nombre de lignées fondatrices de l'échantillon, le nombre de mutations dans la partie de la généalogie plus récente que la date* t_f *de fondation, et le type —hybride pour* N_4 *et* N_5*— de la loi* a posteriori.

pour les paramètres qui définissent le modèle mutationel SMM. Les inférences sont beaucoup plus précises que pour des échantillons monolocus simulés sous le même scenario (non-illustré, car très semblable aux figures 6.6, 6.9. 6.12 et 6.15), en ce sens qu'on a moins de points dispersés hors des zones de forte densité. La vraie valeur du vecteur ϕ de paramètres se trouve dans la zone de plus forte densité. Toutefois, les intervalles de plus forte densité marginale *a posteriori* restent larges, en raison de l'étalement de la loi *a posteriori*, et de corrélations entre les paramètres. Par exemple, la loi jointe *posteriori* de r et t_f a une forte densité sur une très fine bande, mais cette bande couvre plusieurs ordres de grandeurs des valeurs marginales de r et t_f. La légende de la figure 6.21 discute le même problème pour le paramètre θ. Les données populationnelles ne permettent donc pas d'inférer de valeurs précises pour l'un des paramètres en particulier. En revanche, les données génétiques imposent des corrélations fortes entre paramètres, donc si l'on a des informations *a priori* sur certains paramètres, on peut inférer précisément ceux qui leur sont corrélés.

En conclusion, l'augmentation du nombre de loci concentre considérablement l'échantillon MCMC dans de petites zones de l'espace des paramètres, sans toutefois permettre des inférences à l'ordre de grandeur près pour chaque paramètre.

6.2.3 Effets de l'intensité et de la date de la fondation-explosion

Dans les paragraphes précédents, deux jeux de valeurs des paramètres ont été utilisés pour la genèse des données. L'un correspond à une fondation-explosion violente ($r = 10\,000$, $a = 1$, $t_f = 0.001$, $\theta = 100$, ce qui est assez proche de nos *a priori* concernant l'histoire de la population de chats harets des Kerguelen, *cf.* chapitre 7). L'autre décrit une fondation-explosion moins drastique et plus ancienne, avec $r = 100$, $a = 1$, $t_f = 0.05$ et $\theta = 10$. Ces deux cas sont en fait très comparables en ce qu'ils placent t_f à proximité du temps espéré de coalescence en l'absence de l'événement de fondation (*i.e.* pour $r = a = 10\,000$). On a donc, dans des jeux de données multilocus, un mélange de loci ayant un allèle fondateur, et de loci ayant deux ou trois allèles fondateurs, rarement plus. Dans les deux situations, θ est tel qu'on a en moyenne quelques mutations post-fondation, et parfois des allèles fondateurs assez distants (en tout cas non consécutifs). La distribution jointe du nombre f de fondateurs et du nombre m de mutations post-fondation est représentée sur la figure 6.22, pour les deux scénarios. Il semble, d'après l'exploration que j'ai pu faire du modèle, que n'importe quel scénario de fondation-

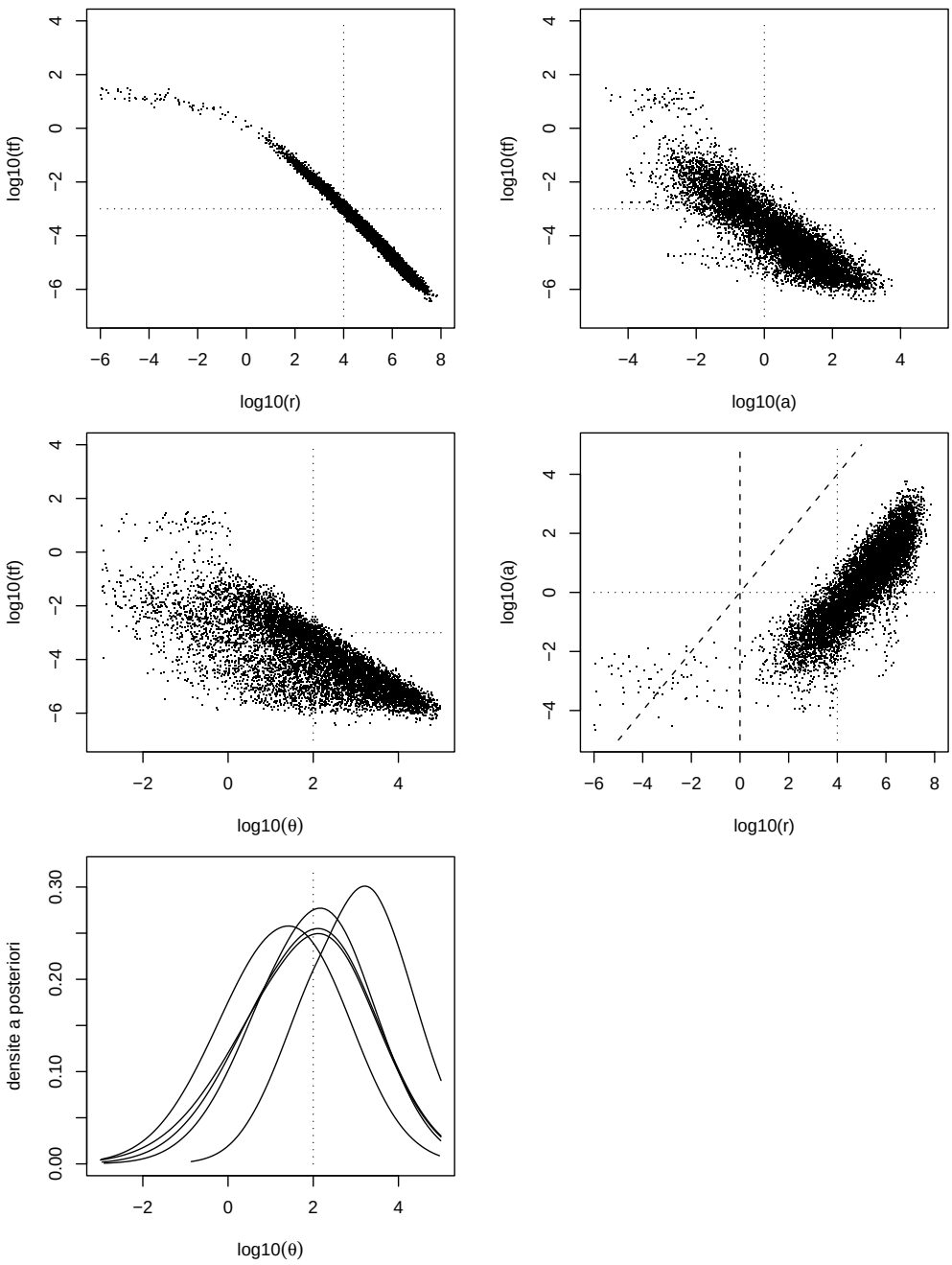


FIG. *6.21: Représentations de la loi* a posteriori *pour l'échantillon de* 5 *loci de configuration donnée dans le tableau 6.3. Les valeurs utilisées pour la simulation des jeux de données sont indiquées en traits pointillés pour les quatre représentations bivariées. Pour la loi bivariée de* a *et* r*, le secteur qui correspond à une inférence de fondation-explosion (*$\log_{10}(r) > 0$ *et* $\log_{10}(r) > \log_{10}(a)$*) est délimité à gauche par deux lignes discontinues. Les données sont fortement en faveur d'une histoire de fondation-explosion :* 99% *des points de l'échantillon correspondent à des fondation-explosion, alors que leur probabilité* a priori *est de* 0.48*). La densité marginale* a posteriori *pour* θ *est figurée pour chacun des* 5 *loci. Le mode est très proche de la "vraie" valeur pour* 3 *des* 5 *loci, et il est légèrement décalé pour* N_3 *vers les faibles valeurs de* θ *et pour* N_4 *vers les fortes valeurs de* θ*. Dans les* 5 *cas, l'intervalle de plus forte densité* a posteriori *est très large (intervalle* $HPD_{0.95}$ $[-0.6, 4.4]$ *pour* $\log_{10}(\theta)$ *au locus* N_2*, qui donne l'intervalle le plus restreint). Les données populationnelles ne permettent donc pas d'inférer précisément le taux de mutation des marqueurs.*

explosion donnant une distribution jointe de ce type pour f et m a des signatures franches et détectables.

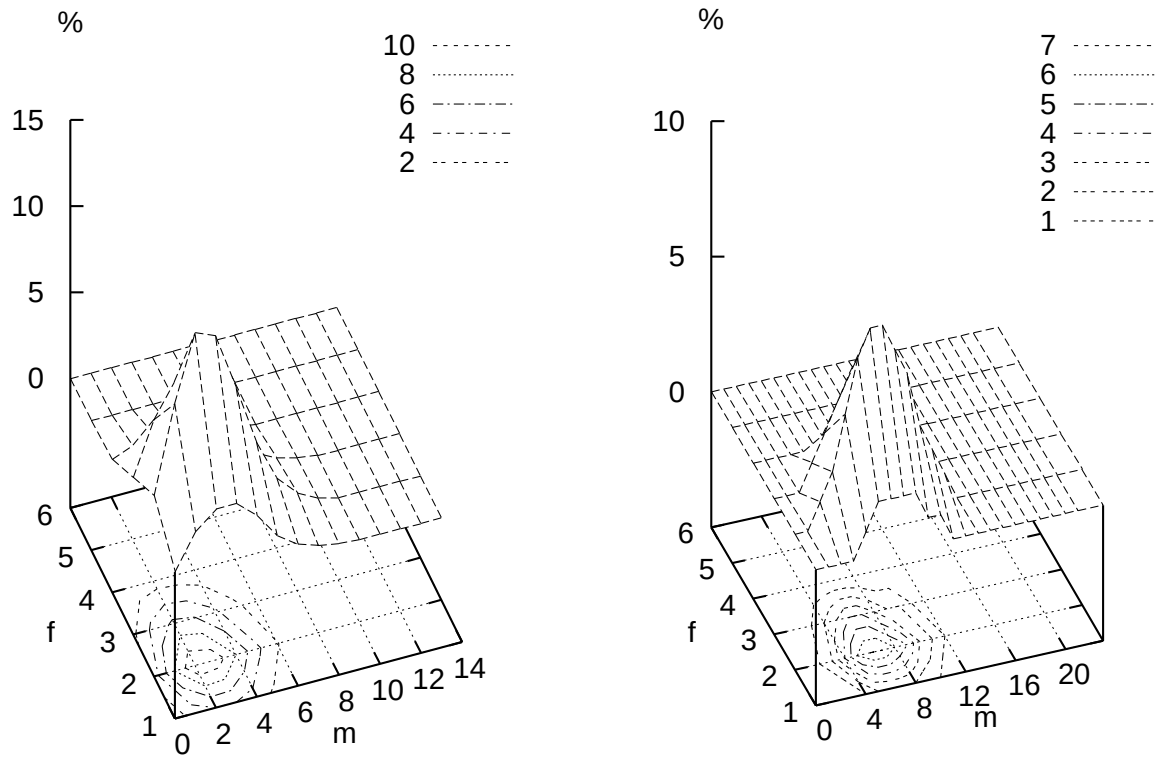


FIG. *6.22: Distribution jointe du nombre f de lignées fondatrices et du nombre m de mutations post-fondation, estimée sur* 100000 *simulations de jeux de données de taille $n = 100$ gènes, pour deux scénarios de fondation-explosion : à gauche, $r = 10000$, $a = 1$, $t_f = 0.001$, $\theta = 100$; à droite, $r = 100$, $a = 1$, $t_f = 0.05$, $\theta = 10$.*

Il serait intéressant d'explorer dans quelle gamme de valeurs du nombre de lignées fondatrices et du nombre de mutations post-fondation on reste capable de détecter une fondation-explosion. Un rapide essai avec un scénario de fondation récente donnant un mode de 9 lignées fondatrices a montré que le nombre de lignées ne change pas tellement le nombre d'allèles fondateurs. Ce sont surtout les allèles fréquents dans la population source qui sont échantillonnés à la fondation, éventuellement en plusieurs exemplaires. De fait, d'après un jeu de données de 5 loci (avec $n = 50$), il semble que l'on reste capable de détecter la fondation (non illustré).

6.2.4 De la loi *a posteriori* jointe aux lois marginales

Nous avons jusqu'ici donné essentiellement des représentations bivariées de la loi *a posteriori* des paramètres démographiques (seuls t_f et θ ont, à deux reprises, eu droit à des lois marginales). Peut-on inférer les valeurs d'un seul des paramètres de façon assez précise, pour des jeux de données d'un ou plusieurs loci et sans information *a priori* particulière sur les valeurs des autres paramètres ? L'échantillon monolocus de configuration F donne une loi *a posteriori*

assez propre (j'entends par là qu'il y a peu de points éparpillés hors des zones de forte densité). En le considérant, nous nous plaçons donc dans une situation favorable, pour l'estimation de paramètres. Supposons par exemple que l'on souhaite donner un intervalle de probabilité à posteriori 0.95 pour $\log_{10}(r)$ ($\mathrm{HPD}_{0.95}$, délimité par des points de même densité, c'est à dire situés sur une même courbe de niveau de la surface de densité *a posteriori*. La bande de forte densité *a posteriori* couvrant tout l'intervalle de densité *a priori* uniforme, l'intervalle $\mathrm{HPD}_{0.95}$ risque d'être très large. Effectivement, cet intervalle estimé à l'aide de LOCFIT est de la largeur du creneau de densité *a priori* uniforme (bornes -5 et 5). L'intervalle $\mathrm{HPD}_{0.95}$ ne nous informe donc guère sur le signe de la variation d'effectif entre N_1 et N_0. L'imprécision des lois *a posteriori* marginales pour a, t_f et θ est grande également (pour vous en rendre compte, projetez virtuellement les points d'un cadran bivarié sur la dimension pour laquelle vous voulez obtenir la loi marginale). En conclusion, un jeu de données monolocus peut contenir beaucoup d'information démographique, comme illustré sur les cadrans bivariés, en ce sens que les lois *a posteriori* sont concentrées sur un support étroit. Mais cette information ne permet pas d'inférer précisément les valeurs individuelles des paramètres. La situation est meilleure pour l'échantillon de 5 loci N_1 à N_5. Les représentations bivariées sont très informatives, et les vraies valeurs des paramètres tombent parfaitement dans les zones de forte densité *a posteriori*. 99% des points de l'échantillon *a posteriori* correspondent à des fondation-explosion (pour un poids *a priori* de 0.48). Cette précision se retrouve dans l'intervalle de plus forte densité *a posteriori* : on obtient $\mathrm{HPD}_{0.95} = [2.5, 7.4]$ pour $\log_{10}(r)$ qui était affecté d'une loi *a priori* uniforme sur $[-6; 8]$. Cet intervalle contient la vraie valeur $\log_{10}(r) = 4$. Des échantillons de petite taille (50 gènes) à 5 loci permettent donc dans ce cas d'inférer clairement que $r \gg 1$, même si la largeur de l'intervalle $\mathrm{HPD}_{0.95}$ correspond encore à 5 ordres de grandeur de variation pour r !

6.3 Violation des hypothèses du modèle démographique

Les hypothèses démographiques du modèle sont nombreuses, puisque l'on suppose une population close, pangamique, se reproduisant selon des modalités qui rendent pertinent l'usage du coalescent standard avec variations d'effectif. Les conséquences sur les inférences de la violation de chacune de ces hypothèses nécessite clarification. À ce jour, je n'ai exploré que les conséquences de la violation de l'hypothèse sur la forme de la courbe de variation d'effectif, à l'aide d'un modèle de fondation-explosion avec délai, décrit ci-après. Quelles que soient les données fournies à MSVAR, la méthode donne une loi *a posteriori*, qui bien sûr n'est pertinente que lorsque l'histoire de la population est raisonnablement bien décrite par le modèle. Nous verrons que les écarts à ce modèle peuvent avoir des conséquences sensiblement sur l'inférence, si l'on oublie que ce sont les paramètres du modèle que l'on infère, et non les paramètres populationnels.

6.3.1 Modèle démographique avec délais pour la genèse de données

On ajoute deux phases à la séquence démographique : un délai entre le crash et le début de la reprise exponentielle, et un plateau entre la fin de la croissance exponentielle et l'échantillonnage. Les paramètres naturels pour décrire le modèle de fondation-délai-explosion-plateau obtenu — appelé ci-après modèle avec délais— sont comme précédemment les trois effectif N_0, N_1 et N_2, mais cette fois trois dates t_f, t_e et t_k sont nécessaires au lieu d'une seule (*cf.* figure 6.23).

Exprimé relativement à la taille N_0, l'effectif $v(t) = N(t)/N_0$ est, avec $r = N_0/N_1$ et $a = N_0/N_2$ et pour des valeurs de t croissantes quand on remonte dans le passé :

$$v(t) = \begin{cases} 1 & \text{si } t_k \geq t, \\ r^{-\frac{t-t_k}{t_e-t_k}} & \text{si } t_e \geq t > t_f, \\ 1/r & \text{si } t_f \geq t > t_e, \\ 1/a & \text{si } t > t_f. \end{cases}$$

Cette modification permet des reprises démographiques beaucoup plus rapides pour une même valeur du paramètre $r = N_0/N_1$, si $t_e = t_f$. En effet, la croissance exponentielle est seulement répartie entre les dates t_f et t_k. La dérive génétique dans les premières générations s'en trouve réduite. Au contraire, s'il y a un délai avant la reprise démographique, la dérive a plus de temps pour éroder la diversité du groupe fondateur.

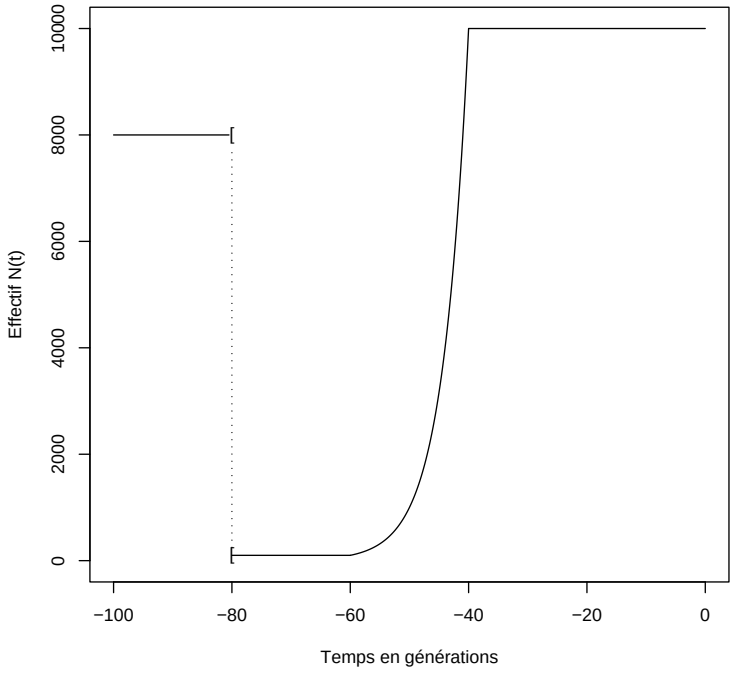


FIG. *6.23: Modèle démographique de fondation-explosion avec délai après la fondation et plateau entre la fin de la croissance exponentielle et l'échantillonnage. Une population de taille jusque-là constante* $N_2 = 8\,000$ *gènes tombe brutalement à* $N_1 = 100$ *gènes,* $10\,000 \times t_f = 80$ *générations avant d'être échantillonnée. La population reste à l'effectif* N_1 *pendant* 20 *générations, puis connaît à partir de* $10\,000 \times t_e = 60$ *générations une reprise démographique exponentielle pour se stabiliser à l'effectif de* $N_0 = 10\,000$ *gènes,* $10\,000 \times t_k = 40$ *générations avant l'échantillonnage.*

6.3.2 Simulation de coalescents pour le modèle avec délais

Comme explicité au paragraphe 4.2.2, la probabilité conditionnelle d'occurrence d'un événement de coalescence au temps t_{i+1}, sachant que la coalescence précédente a eu lieu au temps t_i est

$$p(c, t_{i+1}|t_i) = c(t_{i+1}) \ \exp\left(-\int_{t_i}^{t_{i+1}} c(t)dt\right)$$

où $c(t) = n_i(n_i - 1)/(2v(t))$ est le taux de coalescence instantané (on rappelle que n_i est le nombre de lignées pouvant coalescer).
La distribution cumulée correspondante $D(t_{i+1}|t_i)$ est

$$\int_{t_i}^{t_{i+1}} p(c, t|t_i)dt = 1 - \exp\left(-\int_{t_i}^{t_{i+1}} c(t)dt\right).$$

Pour le modèle de fondation-explosion avec délais, on obtient par intégration, et selon la date relative de t_i et t_{i+1} par rapport aux trois dates charnières t_k, t_e et t_f,

$$D(t_{i+1}|t_i) = \exp\left(-\frac{n_i(n_i-1)}{2}\,\mathcal{L}\right) \quad \text{où}$$

$$\mathcal{L} = \begin{cases} t_{i+1} - t_i & \text{si } t_k \geq t_{i+1} > t_i, \\ (t_k - t_i) + \dfrac{t_e - t_k}{\log r}\left(r^{\frac{t_{i+1}-t_k}{t_e-t_k}} - 1\right) & \text{si } t_e \geq t_{i+1} > t_k \geq t_i, \\ (t_k - t_i) + \dfrac{t_e - t_k}{\log r}(r-1) + r(t_{i+1} - t_e) & \text{si } t_f \geq t_{i+1} > t_e \geq t_k \geq t_i, \\ (t_k - t_i) + \dfrac{t_e - t_k}{\log r}(r-1) + r(t_f - t_e) + a(t_{i+1} - t_f) & \text{si } t_{i+1} > t_f \geq t_k \geq t_i, \\ \dfrac{t_e - t_k}{\log r}\left(r^{\frac{t_{i+1}-t_k}{t_e-t_k}} - r^{\frac{t_i-t_k}{t_e-t_k}}\right) & \text{si } t_e \geq t_{i+1} > t_i > t_k, \\ \dfrac{t_e - t_k}{\log r}\left(r - r^{\frac{t_i-t_k}{t_e-t_k}}\right) + r(t_{i+1} - t_e) & \text{si } t_f \geq t_{i+1} > t_e \geq t_i > t_k, \\ \dfrac{t_e - t_k}{\log r}\left(r - r^{\frac{t_i-t_k}{t_e-t_k}}\right) + r(t_f - t_e) + a(t_{i+1} - t_f) & \text{si } t_{i+1} > t_f > t_e > t_i > t_k, \\ r(t_{i+1} - t_i) & \text{si } t_f \geq t_{i+1} > t_i > t_e\ , \\ r(t_f - t_i) + a(t_{i+1} - t_f) & \text{si } t_{i+1} > t_f > t_i > t_e, \\ a(t_{i+1} - t_i) & \text{si } t_{i+1} > t_i > t_f. \end{cases}$$

Pour simuler les dates de coalescence successives dans l'histoire d'un échantillon, on résout en t_{i+1} l'égalité $D(t_{i+1}|t_i) = \mathcal{U}$ (ou de façon équivalente, $1 - D(t_{i+1}|t_i) = \mathcal{U}$), où $\mathcal{U}$ est une réalisation d'un tirage uniforme sur $[0;1]$, et on obtient :

$$
t_{i+1} \quad = \quad \begin{cases}
t_i - 2 \times \dfrac{\log \mathcal{U}}{n_i(n_i-1)} & \text{si } t_k \geq t_{i+1} > t_i, \\
t_k + \dfrac{t_e - t_k}{\log r} \times \log\left[1 + \dfrac{\log r}{t_e - t_k}\left(-2 \times \dfrac{\log \mathcal{U}}{n_i(n_i-1)} - (t_k - t_i)\right)\right] & \text{si } t_e \geq t_{i+1} > t_k \geq t_i, \\
t_e - 2 \times \dfrac{\log \mathcal{U}}{r\, n_i(n_i-1)} - \dfrac{t_k - t_i}{r} - \dfrac{(t_e - t_k)(r-1)}{r \log r} & \text{si } t_f \geq t_{i+1} > t_e \geq t_k \geq t_i, \\
t_f - 2 \times \dfrac{\log \mathcal{U}}{a\, n_i(n_i-1)} - \dfrac{t_k - t_i}{a} - \dfrac{r(t_f - t_e)}{a} - \dfrac{(r-1)(t_e - t_k)}{a \log r} & \text{si } t_{i+1} > t_f \geq t_k \geq t_i, \\
t_k + \dfrac{t_e - t_k}{\log r} \times \log\left[r^{\frac{t_i - t_k}{t_e - t_k}} - 2 \times \dfrac{\log \mathcal{U}}{n_i(n_i-1)} \times \dfrac{\log r}{t_e - t_k}\right] & \text{si } t_e \geq t_{i+1} > t_i > t_k, \\
t_e - 2 \times \dfrac{\log \mathcal{U}}{r\, n_i(n_i-1)} - \dfrac{t_e - t_k}{r \log r}\left(r - r^{\frac{t_i - t_k}{t_e - t_k}}\right) & \text{si } t_f \geq t_{i+1} > t_e \geq t_i > t_k, \\
t_f - 2 \times \dfrac{\log \mathcal{U}}{a\, n_i(n_i-1)} - \dfrac{t_e - t_k}{a \log r}\left(r - r^{\frac{t_i - t_k}{t_e - t_k}}\right) - \dfrac{r(t_f - t_e)}{a} & \text{si } t_{i+1} < t_f \geq t_e \geq t_i > t_k, \\
t_i - 2 \times \dfrac{\log \mathcal{U}}{r\, n_i(n_i-1)} & \text{si } t_f \geq t_{i+1} > t_i > t_e, \\
t_f - 2 \times \dfrac{\log \mathcal{U}}{a\, n_i(n_i-1)} - \dfrac{r(t_f - t_i)}{a} & \text{si } t_{i+1} > t_f \geq t_i > t_e. \\
t_i - 2 \times \dfrac{\log \mathcal{U}}{a\, n_i(n_i-1)} & \text{si } t_{i+1} > t_i > t_f.
\end{cases}
$$

6.3.3 Effet d'une stabilisation de l'effectif avant l'échantillonnage

Le modèle démographique avec délais a été utilisé pour générer 5 échantillons indépendants de taille $n = 50$, avec les valeurs des paramètres $r = 1\,000$, $a = 0.1$, $t_f = t_e = 0.2$, $t_k = 0.19$ et $\theta = 1$. Ces valeurs de paramètres seraient par exemple obtenues pour une population de taille stable $N_0 = 1\,000$ gènes depuis 190 générations, fondée il y a 200 générations par 10 gènes, à partir d'une population ancestrale de taille 10 000 gènes, avec un taux de mutation $\mu = 5.10^{-4}$ pour les marqueurs microsatellites échantillonnés. La phase de croissance exponentielle qui suit la fondation est donc très courte (5% du temps écoulé depuis la fondation).

Les échantillons simulés ont pour configurations $\{2, 48\}$, $\{5, 45\}$, $\{1, 47, 2\}$, $\{1, 0, 0, 0, 0, 2, 47\}$ et $\{9, 22, 4, 0, 2, 13\}$, et individuellement conduisent, pour une loi *a priori* du logarithme décimal de r, a, t_f et θ uniforme sur $[-5; 5]^4$, à des lois *a posteriori* respectivement des types (hybrides dans trois des cas) P_{se}, P_{se}, P_{e}, P_{df} et P_{d}.

La figure 6.24 illustre la loi *a posteriori* pour l'échantillon de 5 loci. Cette loi *a posteriori* est de type P_f, tirant vers P_d. L'échantillon de la loi *a posteriori* est plus diffus que pour les autres échantillons de 5 loci étudiés (qui eux avaient été simulés sous le modèle supposé dans MSVAR). Ce caractère diffus pourrait être un signe de la mauvaise adéquation entre les données et le modèle. *In fine* le jeu de données n'est que faiblement en faveur d'une fondation-explosion. Les valeurs de paramètres utilisées dans les simulations sont-elles au moins dans une zone de forte densité *a posteriori* ? Manifestement non d'après la représentation bivariée de r et t_f. Par exemple si r est ajusté à la vraie valeur, la zone de forte densité donne t_f de l'ordre de 10 fois trop récent. Je n'ai jamais rencontré de tel décalage pour des données simulées avec le modèle démographique supposé dans MSVAR. Ce résultat, qui mériterait d'être vérifié sur un

plus grand nombre de jeux de données, indique que la méthode doit être utilisée avec précaution pour interpréter des jeux de données de populations invasives ayant atteint un effectif stable bien avant l'échantillonnage. On risque alors d'inférer une date beaucoup trop récente pour le début de l'invasion.

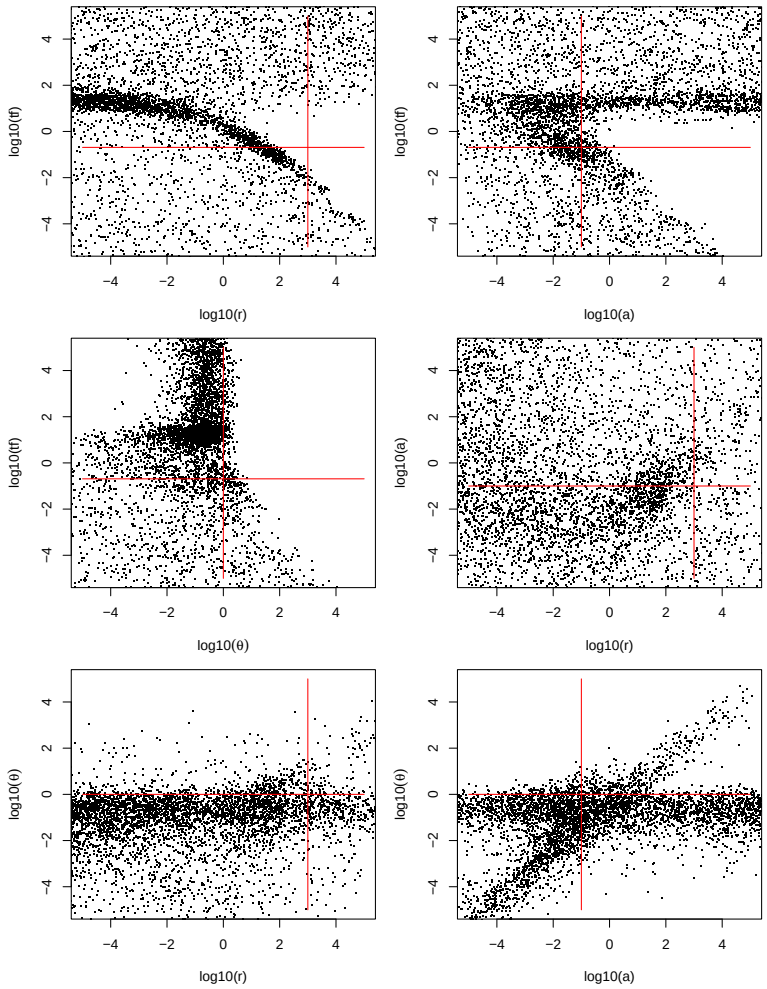


FIG. *6.24: Représentations bivariées de la loi* a posteriori *pour un échantillon de* 5 *loci simulé en supposant une fondation-explosion avec saturation rapide de l'effectif (configuration précisée dans le texte). Pour* $\log_{10}(\theta)$*, seules les valeurs obtenues pour l'échantillon* $\{2, 48\}$ *sont représentées. Les valeurs des paramètres utilisées pour la simulation des jeux de données sont indiquées en traits pleins pour les six représentations bivariées. Le point correspondant aux vraies valeurs des paramètres se trouve hors des zones de plus forte densité, conséquence de la violation des hypothèses sur l'histoire démographique.*

En conclusion, la méthode d'inférence donne la loi a posteriori *des paramètres du modèle sous-jacent : ce qu'elle permet de dire est que pour tel modèle et telles données, on obtient telle loi* a posteriori*. Il faut être extrêmement prudent dans l'interprétation de cette loi* a posteriori*. En particulier, il est erroné de transposer directement les valeurs des paramètres du modèle à des valeurs censées décrire l'histoire des populations.*

6.3.4 Effet d'un délai avant reprise démographique

L'effet d'un délai avant reprise démographique est de faire coalescer les lignées qui, sans ce délai, auraient été des lignées fondatrices candidates. L'effet du délai est donc d'augmenter la proportion de loci donnant une loi *a posteriori* de type P_e et de diminuer d'autant la proportion de loci pour lesquels la loi *a posteriori* est de type P_f ou P_d. Une reprise démographique lente empêche donc la détection de la phase de fondation. Seule la phase d'explosion peut rester détectable.

6.4 Loi a posteriori du paramètre θ, sous SMM

La reconsidération des lois *a posteriori* de types P_s, P_e, P_d et P_f permet de discuter des potentialités des données populationnelles pour l'estimation du taux de mutation SMM de marqueurs.

Lorsque la loi *a posteriori* est de type P_s, ce qui est fréquent en population stable, l'hypothèse de quasi équilibre mutation-dérive permet de déterminer θ de façon précise (figure 6.6, zone **A** sur le cadran **c**) en se basant sur la vraisemblance. Cela est dû au fait que θ est alors indépendant de t_f, r et a. Si la population est effectivement à l'équilibre mutation-dérive, le gain par rapport à une estimation de θ à partir de statistiques sommaires n'est pas visible, comme l'illustre M. Stephens (2001) [137]. En revanche, en cas d'événement démographique récent, suffisamment faible pour être indétectable par les approches fréquentistes (en ce sens qu'il ne conduit pas au rejet de l'hypothèse nulle d'équilibre démographique), l'estimation de θ à partir de statistiques sommaires conduira à une sous-estimation, alors que la méthode bayésienne devrait montrer un poids équivalent pour les zones **A** et **B** (figure 6.6), et donc rendre douteuse l'hypothèse d'équilibre mutation-dérive.

Si les gènes échantillonnés dérivent d'un unique gène fondateur, avec une généalogie caractéristique d'une croissance exponentielle, on a de fortes chances d'obtenir une loi *a posteriori* de type P_e. Dans ce cas, θ est très mal corrélé avec r et a, mais en revanche très bien corrélé avec la date t_f de début de l'explosion démographique (figure 6.9, cadrans **c**, **e** et **f**.

Dans le cas où la loi *a posteriori* est de type P_d, la corrélation entre t_f et θ s'estompe : pour t_f donné, on a seulement une borne supérieure sur θ (figure 6.12, cadran **c**). Une bonne corrélation entre θ et a apparaît : l'espacement entre les allèles fondateurs nous informe sur l'espacement des modes de la distribution dans la population ancestrale, et donc sur le taux de mutation.

Enfin, dans le cas où la loi *a posteriori* est de type P_f, on observe à la fois une corrélation de θ avec t_f et avec r (figure 6.15, cadrans **c** et **e**), ce qui est favorable à l'estimation des paramètres mutationnels, si l'on a des informations sur les paramètres démographiques.

En conclusion, les populations invasive depuis une date connue sont potentiellement très informatives sur le taux de mutation des marqueurs.

6.5 Interactions entre processus et modèle mutationnel

Les conséquences sur les inférences démographiques de mutations d'amplitude $a > 1$ —ou sauts— le long des généalogies de gènes ont été succintement explorées avec un nombre minimal de simulations —pour des raisons de limitation de la puissance de calcul. Pour cela, un scénario

de stabilité ($r = a = 1$, $\theta = 4$) a été choisi et trois modèles mutationnels : un modèle SMM, un modèle TPM_1 avec $p = 0.8$ et $j = 3$ (soit $m = 1.4$ et $s = 0.13$), et un modèle TPM_2 plus riche en sauts de grande amplitude, avec $p = 0.7$ et $j = 4$ (soit $m = 1.9$ et $s = 0.22$). Les configurations d'échantillons obtenues pour les 3 modèles, puis les inférences par MSVAR ont été comparées.

6.5.1 Processus mutationnel et configuration des échantillons

Pour les 3 modèles mutationnels, la distribution du nombre de mutations dans des généalogies simulées est la même, mais la distribution d'amplitude des mutations varie : on a exclusivement des mutations d'amplitude 1 pour le modèle SMM, et une proportion de mutations d'amplitude $a > 1$ de 0.13 et 0.22 pour les modèles TPM_1 et TPM_2 respectivement (figure 6.25, ligne supérieure). Sous TPM, les mutations produisent plus fréquemment des allèles inexistants : on a une moindre homoplasie. Cela se traduit par un plus grand nombre d'allèles (figure 6.25, ligne inférieure). Même à nombre d'allèles égal, on obtient sous TPM des distributions avec une plus grande variance de taille allélique. Par exemple, conditionnellement au nombre d'allèles 6, la moyenne de la variance empirique de taille allélique est 4.4, 8.1 et 14.2 pour les trois modèles, avec dans les trois cas le mode de la distribution $\simeq 2$ (non représenté). Cela correspond tout simplement à la non-représentation d'allèles de taille intermédiaire entre les extrêmes de l'échantillon : un processus TPM fait apparaître des "trous" dans les distributions de fréquences alléliques, qui deviennent disjointes.

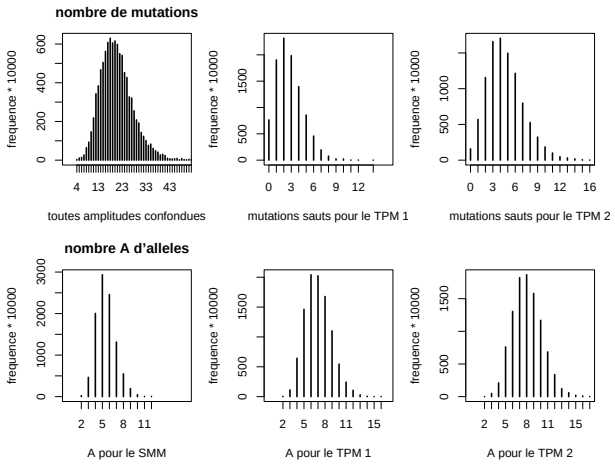


FIG. *6.25: Statistiques sommaires d'échantillons de taille $n = 100$ pour un scénario de stabilité démographique avec $\theta = 4$, et trois modèles mutationnels. Première ligne : distribution du nombre de mutations par généalogie —identique pour les 3 modèles mutationnels— et distribution du nombre de mutations "sauts" pour les modèles TPM_1 et TPM_2. Seconde ligne : distribution du nombre d'allèles pour les 3 modèles.*

Pour illustrer cet effet, on s'affranchit de la variabilité sur la topologie datée des généalogies de gènes, et sur la variabilité du nombre, de la position et du type des mutations. On simule dans ce but des configurations corrélées pour les 3 modèles. Une généalogie commune est tirée selon le coalescent standard, pour un échantillon de taille $n = 100$. Sur chaque branche de cette généalogie on ajoute des mutations, en nombre tiré selon une loi de Poisson de moyenne $\theta\delta t/2$, où δt est la longueur de la branche mesurée en unités de N_0 générations, et avec $\theta = 4$. La direction des mutations (perte ou gain de répétitions du motif) est décidée pour chaque mutation. Enfin, un réel dans $[0;1]$ est tiré uniformément pour chacune des mutations afin de déterminer sa place dans la distribution d'amplitude des trois modèle (par inversion de la fonction de répartition). Tout est donc commun à des triplets d'échantillons : la topologie datée, le nombre, la position, le sens des mutations. Seule change l'amplitude des mutations, et encore, cette amplitude correspond à une même valeur de la fonction de répartition pour les trois modèles. Le tableau 6.4 donne 5 triplets de configurations corrélées obtenues selon cette

procédure et montre que certains allèles intermédiaires ne sont pas représentés.

Intuitivement, un processus mutationnel TPM risque donc de mimer une fondation.

SMM	TPM_1	TPM_2
$\{6, 41, 4, 9, 23, 17\}$	$\{7, 49, 27, 6, 0, 0, 0, 11\}$	$\{1, 0, 0, 0, 6, 49, 27, 6, 0, 0, 0, 0, 0, 0, 11\}$
$\{4, 18, 75, 3\}$	$\{1, 0, 0, 1, 0, 0, 0, 3, 18, 74, 3\}$	$\{1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 3, 6, 12, 74, 3\}$
$\{1, 21, 20, 53, 5\}$	$\{1, 22, 19, 9, 1, 0, 0, 0, 44, 4\}$	$\{5, 3, 19, 17, 7, 1, 0, 0, 0, 0, 0, 42, 1, 2, 3\}$
$\{3, 7, 44, 37, 9\}$	$\{3, 2, 0, 0, 0, 18, 10, 1, 0, 5, 26, 27, 8\}$	$\{3, 2, 0, 0, 0, 0, 0, 0, 0, 18, 10, 1, 0, 0, 0, 0, 0, 5, 26, 27, 8\}$
$\{1, 9, 25, 22, 41, 2\}$	$\{1, 6, 3, 0, 2, 18, 7, 24, 4, 35\}$	$\{1, 6, 3, 0, 0, 2, 18, 6, 2, 0, 1, 20, 3, 15, 23\}$

TAB. *6.4: Triplets de configurations corrélées obtenus pour les modèles SMM, TPM_1 et TPM_2 (détails dans le texte).*

6.5.2 Processus TPM et inférence démographique sous SMM

Le premier triplet d'échantillons corrélés a été repris pour tester si un processus mutationnel TPM abuse MSVAR quand un SMM est supposé. On rappelle que les configurations sont :
$M_{SMM} = \{6, 41, 4, 9, 23, 17\}$ pour le processus mutationnel SMM,
$M_{TPM_1} = \{7, 49, 27, 6, 0, 0, 0, 11\}$ pour le TPM_1 avec $(m, s) = (1.4, 0.13)$,
$M_{TPM_2} = \{1, 0, 0, 0, 6, 49, 27, 6, 0, 0, 0, 0, 0, 0, 11\}$ pour le TPM_2 avec $(m, s) = (1.9, 0.22)$.
On a 24 mutations en tout dans la généalogie, dont 0, 3 et 6 mutations d'amplitude $a > 1$ respectivement pour les trois modèles. M_{SMM} est assez typique d'un échantillon généré en population stable, avec 2 modes de fréquences alléliques, et sans allèles manquants entre les deux classes extrêmes. M_{TPM_1} en diffère essentiellement par trois allèles manquants, et l'échantillon M_{TPM_2} est encore plus étalé, avec 3 et 6 allèles manquants consécutifs. Le processus mutationnel semble donc avoir généré un signal de déclin de population. Les figures 6.26 et 6.27 montrent qu'en réalité, seul le TPM_2 avec le plus de sauts a suffisamment affecté la configuration de l'échantillon pour modifier nettement la densité *a posteriori*, et produire un signal de déclin, avec ou sans explosion ultérieure : la loi a posteriori passe du type P_s à une forme intermédiaire entre P_d et P_f. Dans les trois cas, la loi *a priori* a été supposée uniforme pour le logarithme décimal de r, a, t_f, et θ sur $[-5, 5] \times [-5, 5] \times [-4, 4] \times [-4, 5]$.

Un processus mutationnel de type TPM semble donc pouvoir mimer un signal de fondation. Toutefois, sur l'exemple traité, ce TPM doit pour cela être fortement éloigné d'un SMM. Des conclusions plus définitives nécessitent un plus grand nombre de simulations.

6.5.3 Processus TPM supposé connu et inférences démographiques

Supposons connu le processus mutationnel —de type TPM— du marqueur. L'incorporation de cette donnée à la méthode d'inférence conduit-elle à une loi *a posteriori* de type P_s, pour des échantillons qui, en supposant un modèle SMM, ont un signal de déséquilibre démographique ? Pour cela, on reconsidère l'échantillon $C_{TPM_2} = \{1, 0, 0, 0, 6, 49, 27, 6, 0, 0, 0, 0, 0, 0, 11\}$ dont on rappelle qu'il a été simulé en supposant un TPM avec $p = 0.7$ et $j = 4$. On utilise MSVAR en supposant une loi *a priori* rectangulaire $[-5; 5]^4$ pour le logarithme décimal des paramètres r, a, θ et t_f, et —délit d'initié sur la nature du processus mutationnel— une masse de probabilité 1 pour $p = 0.7$ et pour $j = 4$. La figure 6.28 montre que l'on retrouve une loi *a posteriori* de type P_s.

Le fait de connaître précisément le processus mutationnel permet donc d'interpréter correctement le jeu de données en termes d'histoire démographique : on supprime le faux signal de déclin obtenu si l'on suppose un modèle SMM alors que le processus est TPM.

6.5.4 Processus et modèle TPM, et inférences démographiques

Supposons enfin que l'on ne connaisse pas exactement le processus mutationnel. On suppose par exemple *a priori* que le paramètre p de fréquence des "pas" a une distribution uniforme sur $[0;1]$, et que le logarithme décimal de la moyenne j d'amplitude des "sauts" a une distribution uniforme sur $[0;1]$. Cela correspond pour les paramètres plus intuitifs m et s à une loi *a priori* dont l'expression analytique peut être explicitée au moyen d'un changement de variable (encart page 141). Cette loi *a priori* favorise fortement les valeurs de paramètres proches d'un SMM (figure 6.29). On fait donc l'*a priori* —conforme aux données de pedigrees— que le processus mutationnel est un TPM qui avec forte probabilité est proche d'un SMM (peu de sauts et de faible amplitude). On s'intéresse à la loi *a posteriori* pour le même échantillon $C_{TPM_2} = \{1,0,0,0,6,49,27,6,0,0,0,0,0,0,11\}$ que précédemment. La loi *a posteriori* marginale pour les paramètres m et s est comparée à leur loi *a priori* (figure 6.29). Le jeu de données est en faveur d'un TPM assez marqué, avec une amplitude moyenne des sauts proche de la "vraie" valeur, mais une forte proportion de mutations d'amplitude $a > 1$. La loi *a posteriori* pour les paramètres démographiques est nettement de type P_s, alors que l'échantillon a un net signal de déséquilibre démographique, si l'on suppose un modèle SMM strict.

On supprime donc le faux signal de déséquilibre, même sans supposer les paramètres du TPM connu —et bien que l'on se trompe de valeurs pour ces paramètres (figure 6.30).

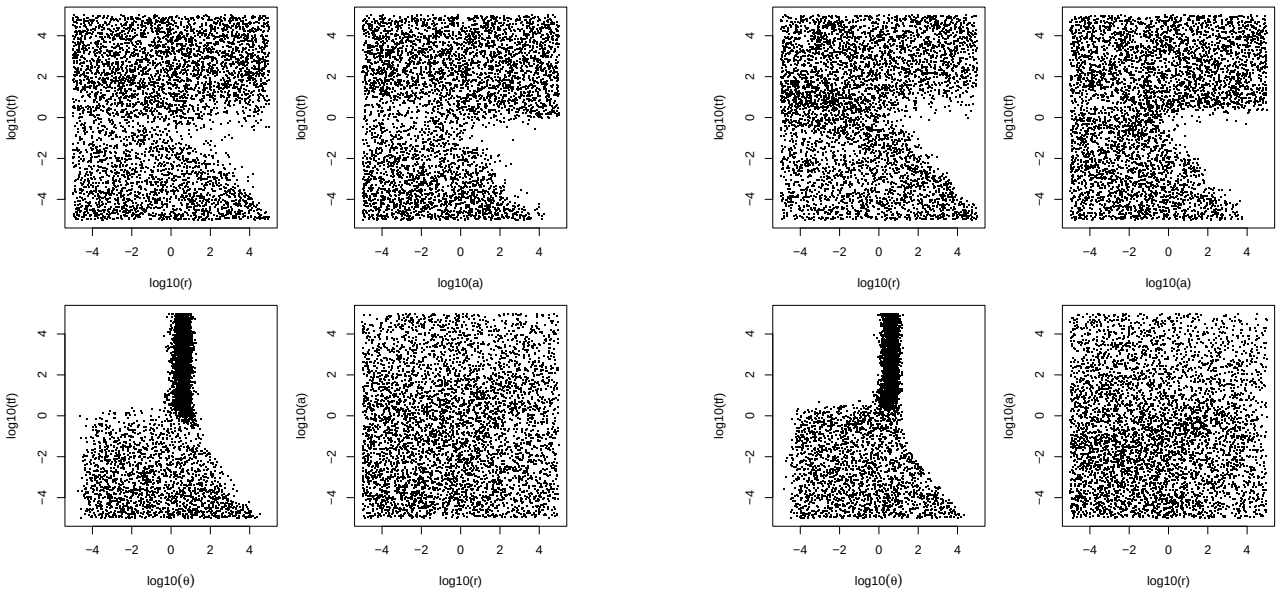


FIG. *6.26: Loi* a posteriori *de type P_s pour deux échantillons corrélés de configurations $\{6,41,4,9,23,17\}$ (deux colonnes de gauche) et $\{7,49,27,6,0,0,0,11\}$ (deux colonnes de droite) , simulés en supposant une stabilité démographique avec $\theta = 4$, respectivement sous SMM et sous TPM avec $(m,s) = (1.4, 0.13)$.*

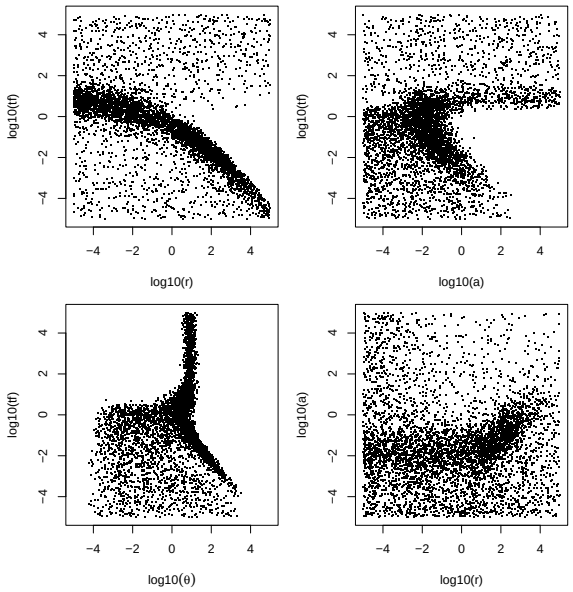


FIG. *6.27: Loi* a posteriori *intermédiaire entre les types P_d et P_f pour un échantillon de configuration* $\{1,0,0,0,6,49,27,6,0,0,0,0,0,0,11\}$ *simulé en supposant une stabilité démographique avec* $\theta = 4$*, sous TPM avec* $(m,s) = (1.9, 0.22)$

.

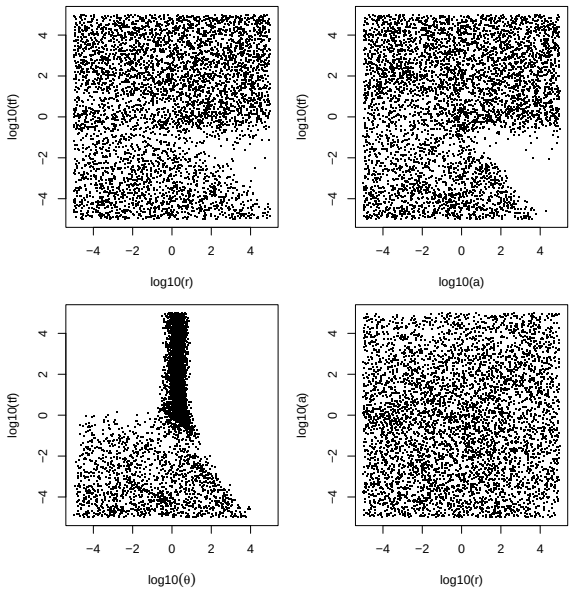


FIG. *6.28: Loi* a posteriori *nettement de type P_s pour un échantillon de configuration* $\{1,0,0,0,6,49,27,6,0,0,0,0,0,0,11\}$ *simulé en supposant une stabilité démographique avec* $\theta = 4$*, sous TPM avec* $p = 0.7$ *et* $j = 4$ *(soit* $(m,s) = (1.9, 0.22)$*). Pour les inférences, la loi* a priori *a été supposée uniforme sur* $[-5;5]^4$ *pour* $\log_{10}(r)$, $\log_{10}(a)$, $\log_{10}(t_f)$ *et* $\log_{10}(\theta)$*. Une masse de probabilité* 1 *a été supposée pour les paramètres du modèle TPM, aux valeurs* $p = 0.7$ *et* $j = 4$ *ayant servi à simuler le jeu de données.*

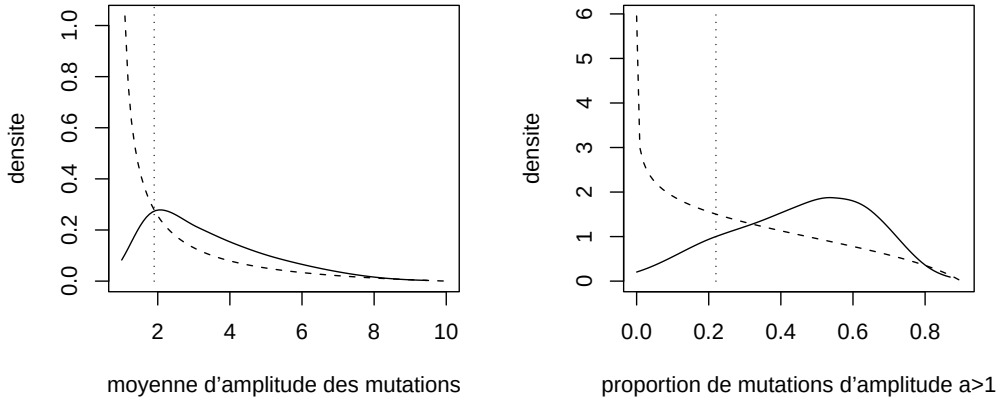


FIG. *6.29: Lois* a priori *(trait interrompu) et* a posteriori *(trait plein) comparées pour les paramètres m et s caractérisant un TPM. Les valeurs de ces paramètres ayant servi à simuler le jeu de données de configuration* $\{1,0,0,0,6,49,27,6,0,0,0,0,0,0,11\}$ *sont indiquées (barres verticales en pointillé).*

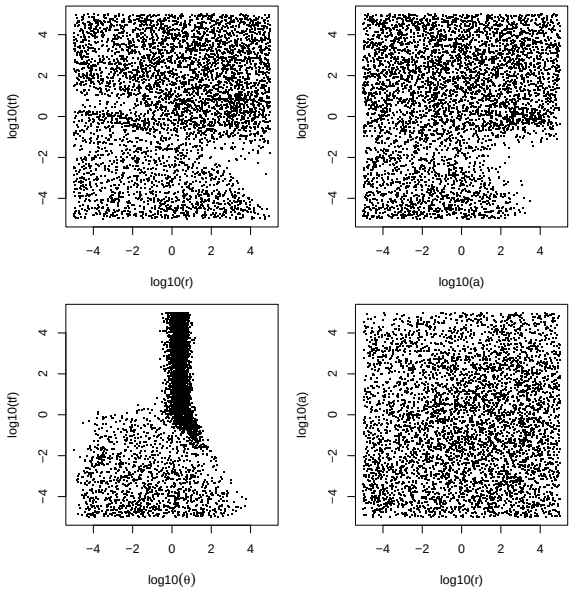


FIG. *6.30: Loi* a posteriori *nettement de type P_s pour un échantillon de configuration* $\{1,0,0,0,6,49,27,6,0,0,0,0,0,0,11\}$ *simulé en supposant une stabilité démographique avec $\theta = 4$, sous TPM avec $p = 0.7$ et $j = 4$ (soit $(m,s) = (1.9, 0.22)$). Pour les inférences, la loi* a priori *a été supposée uniforme sur* $[-5;5]^4$ *pour* $\log_{10}(r)$, $\log_{10}(a)$, $\log_{10}(t_f)$ *et* $log_{10}(\theta)$, *et uniforme sur* $[0;1]$ *pour les paramètres p et* $\log_{10}(j)$ *du TPM .*

Changement de variable de (p, j) vers (m, s) et lois a priori

Dans MSVAR, le modèle TPM est paramétré par p, la probabilité pour que l'amplitude d'une mutation soit calculée selon le modèle SMM, et par j, la moyenne de l'amplitude lorsqu'elle est tirée selon la phase géométrique du modèle. Cette paramétrisation est conforme à la définition du modèle TPM [31], et elle permet un calcul facile de la densité d'occurrence des événements. Malheureusement, le lien entre p, j et la distribution d'amplitude est difficile à établir de tête, à cause de la présence de mutations d'amplitude 1 dans la phase géométrique du modèle (encart page 81). Une paramétrisation plus intuitive considère la moyenne m d'amplitude des mutations et la proportion s de mutations d'amplitude $a > 1$:

$$(m, s) = (p + (1 - p)j, 1 - p + (p - 1)/j) .$$

On souhaite donc exprimer les inférences sur les paramètres du modèle TPM en fonction de (m, s), et pouvoir comparer la loi *a posteriori* de (m, s) à leur loi *a priori*. Ci-dessous, nous établissons cette loi *a priori* à partir de celle de (p, j) qui est utilisée dans MSVAR.
La densité *a priori* $g(m, s)$ s'exprime en fonction de la densité *a priori* $f(p, j)$ et de la valeur absolue du déterminant de la matrice des dérivées partielles de (p, j) comme

$$g(m, s) = f(p, j) \begin{vmatrix} \partial p/\partial m & \partial j/\partial m \\ \partial p/\partial s & \partial p/\partial s \end{vmatrix}$$

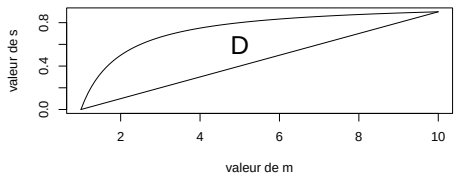
On s'est donné par définition $(m, s) = (p + (1 - p)j, 1 - p + (p - 1)/j)$, et on en tire par inversion $(p, j) = ((m - sm - 1)/(m - s - 1), (m - 1)/s)$ (bijection réciproque), ce qui permet de calculer les dérivées partielles puis le déterminant de la matrice et d'obtenir

$$g(m, s) = f(p, j)\frac{1 - m}{s(s - m + 1)}$$

Dans le cas particulier pour lequel la loi de p est uniforme sur $[0, 1]$ ainsi que la loi de $\log_{10}(j)$, la loi jointe *a priori* $f(p, j)$ s'exprime $f(p, j) = 1/(j\ \log(10))$ donc la loi jointe $g(m, s)$ est, sur un domaine $\mathcal{D}$,

$$g(m, s) = f(p, j)\frac{1 - m}{s(s - m + 1)} = \frac{1}{(m - s - 1)\ \log(10)}$$

Le domaine rectangulaire $(p, j) \in [0, 1] \times [1, 10]$ a pour image selon la bijection réciproque le sous-domaine $\mathcal{D}$ de $(m, s) \in [1, 10] \times [0, 9/10]$ dont les bornes sont définies par les conditions $s \leq 1 - 1/m$ et $m \leq 1 + 10s$:



La loi marginale de s est donc donnée par

$$\int_{\frac{1}{1-s}}^{1+10s} g(m, s)\, dm = \frac{2\log(3) + \log(s) - \log(\frac{s^2}{1-s})}{\log(2) + \log(5)}$$

et de même la loi marginale de m est

$$\int_{\frac{m-1}{10}}^{1-\frac{1}{m}} g(m, s)\, ds = \frac{2\log(3) - \log(2) - \log(5) + \log(m - 1) - \log(\frac{(m-1)^2}{m})}{\log(2) + \log(5)}$$

6.6 Conclusion sur l'exploration du modèle

6.6.1 Précision des inférences démographiques

L'exploration du modèle à partir de données simulées a permis de montrer qu'un unique marqueur microsatellite peut apporter un fort soutien au type de scénario démographique ayant servi à le simuler. Cela est particulièrement vrai lorsque ce scénario est une fondation-explosion. L'augmentation du nombre de loci indépendants permet de préciser les inférences, et avec seulement 50 gènes à 5 loci, simulés pour un scénario tranché ($r = 10\,000$, $a = 1$, $t_f = 0.05$, $\theta = 100$), on peut exclure les histoires paramétrisables qui ne sont pas des fondation-explosion récentes. Un plus grand nombre de loci permettrait certainement de réduire la largeur des intervalles HPD. Pour 10 loci et 50 gènes, l'espacement entre points échantillonnés nécessaire pour obtenir des estimations fiables de la loi *a posteriori* rend l'usage de la méthode assez pénible. Dans la pratique, je n'ai pas pu avec la faible puissance de calcul et la patience dont j'ai disposé (un unique PC équipé d'un processeur 1 GHz) mener à bout 5 runs indépendants de longueur suffisante pour traiter des jeux de données de 10 loci. Une amélioration de dernière minute de la méthode permet toutefois d'envisager son application sereinement, sur des jeux de données de taille environ 10 loci et 50 gènes : il s'agit d'une procédure de sauvegarde de l'état final de la MCMC, qui permet, en cas de non convergence, de repartir de cet état, éventuellement avec de nouvelles valeurs des paramètres, et en particulier du paramètre d'espacement entre points d'échantillonnage MCMC.

6.6.2 Intérêt des populations d'histoire documentée

Nous avons discuté de la corrélation entre paramètres du modèle, et expliqué que cette corrélation empêche d'inférer très précisément les paramètres considérés isolément. En revanche, dans le cas où l'on dispose d'information *a priori* —même avec une incertitude— sur certains paramètres, on peut déduire très précisément la valeur des paramètres qui lui sont corrélés. Ainsi, la connaissance de la date t_f d'une fondation permet d'inférer avec une faible incertitude la valeur du paranmètre d'explosion r (*cf.* figure 5.1 page 99). Cela est particulièrement intéressant pour estimer le paramètre synthétique θ : suite à une fondation-explosion, θ est bien corrélé (négativement) à t_f et à r, si bien que les populations ayant connu une fondation-explosion documentée peuvent permettre de l'estimer sur des données populationnelles.

6.6.3 Améliorations possibles de la méthode

Modèle de fondation-explosion avec délais

Nous avons constaté qu'un écart de l'histoire des variations d'effectif par rapport au modèle conduit à fausser les corrélations entre paramètres. C'est en particulier le cas lorsque l'effectif d'une population sature peu après la fondation au lieu de croître jusqu'à la date d'échantillonnage. Cela n'est pas étonnant : la méthode échantillonne des valeurs pour les paramètres du modèle quel que soit l'échantillon qu'on lui donne. La saturation d'effectif étant sans doute assez courante dans les cas d'application possibles, et les inférences étant lourdement faussées lorsqu'on l'ignore il faudrait paramétrer le plateau. Un délai avant la reprise démographique a sans doute également des effets forts. On pourrait donc intégrer le modèle de simulation de données avec délais à MSVAR).

Modèle mutationnel

Les deux simplifications majeures du modèle mutationnel qui pourraient être levées sont l'hypothèse d'une symétrie de la distribution des pertes et des gains de motif, et l'hypothèse d'une distribution d'amplitude indépendante de la taille. Avant d'implémenter ces possibilités, il serait intéressant de vérifier si les simplifications conduisent à des biais importants dans les inférences, sur des données simulées.

Utilisation de plusieurs échantillons

Les études populationnelles comportent souvent plusieurs échantillons prélevés à quelques années d'intervalle. Il serait tout à fait possible d'ajouter des lignées à coalescer aux différentes dates d'échantillonnage. Cela permettrait d'estimer la valeur de N_0 et non pas seulement le paramètre synthétique θ, comme cela est fait dans les méthodes temporelles d'estimation de la taille efficace.

6.6.4 Comparaison avec les méthodes alternatives

Dans un avenir proche, je souhaite comparer les méthodes actuellement disponibles pour la détection de déclins en marche d'escalier, de bottlenecks transitoires et de fondation-explosion. Ces méthodes sont : (i) les approches fréquentistes basées sur la comparaison de statistiques sommaires (*e.g.* BOTTLENECK [105], indice β de Kimmel [71]), (ii) les approches bayésiennes approchées basées sur de l'échantillonnage par acceptation-rejet [39, 7]. Un début de comparaison entre MSVAR et BOTTLENECK sera proposée au chapitre 7.

Troisième partie

Études de fondations dans des populations sauvages

Il est crucial de vérifier le fonctionnement des méthodes inférentielles, non seulement sur des données simulées, mais aussi en les mettant à l'épreuve sur des données réelles concernant des populations d'histoire connue. En effet, les populations naturelles sont susceptibles de violer certaines hypothèses sous-jacentes aux méthodes, et tester les méthodes sur des données réelles force à améliorer ces méthodes.

Deux histoires de fondation-explosion vont nous permettre de mettre en évidence les limites des hypothèses sous-jacentes à MSVAR, et de proposer des améliorations de la méthode. Ces améliorations seront nécessaires à l'utilisation de la méthode à de réelles fins d'inférence, dans des cas d'applications semblables.

Chapitre 7

Déséquilibre génétique et inférence mutationnelle : le cas de la population de chats harets (*Felis catus*) de Kerguelen

La population de chats harets de l'île principale de l'archipel des Kerguelen a connu une histoire de fondation-explosion documentée, décrite rapidement ci-dessous. Cette situation va nous permettre de vérifier le fonctionnement de MSVAR dans un cas réel, et de comparer la méthode probabiliste avec les méthodes fréquentistes alternatives (BOTTLENECK [24, 105]). On discutera de l'intérêt des informations historiques pour l'estimation des paramètres d'un modèle mutationnel pour les microsatellites marqueurs.

7.1 Contexte historique et biologique

Le chat *Felis catus* a été introduit volontairement par l'homme sur des îles sub-antarctiques, fréquemment dans un but de contrôle de populations invasives de rats. L'objectif était illusoire : le prédateur s'est en effet reporté sur des colonies d'oiseaux, principalement des espèces nichant dans des galeries souterraines, sans protection efficace contre la prédation par le chat [101, 69, 68]. Le rat a constitué une proie alternative, qui a permis au chat de survivre entre deux saisons de nichage [68]. Le chat a été éradiqué avec succès d'îles de petite taille, mais le contrôle de ses effectifs s'est avéré délicat sur les archipels isolés et de grande taille comme celui des Kerguelen. Des études démographiques [126] et génétiques sont menées sur la population de chats de l'île principale des Kerguelen (6600 km^2), afin de comprendre le fonctionnement de cette population et d'établir des mesures de contrôle efficaces.

L'histoire de l'introduction du chat sur Kerguelen est bien documentée, et correspond à une fondation localisée, suivie d'une explosion démographique avec extension de l'aire de répartition [29, 101, 102]. Plus précisément, la population actuelle dérive au plus de quatre chats fondateurs. Deux chats —en provenance de France— ont été introduits en 1951 à la station de recherche de Port-aux-Français [102]. Ils ont disparu de la station, mais on ne peut affirmer qu'ils n'ont pas donné de descendants harets. Un mâle en provenance de Madagascar et une femelle de Cap Town ont été de nouveau introduits à la station en 1956. Cette introduction a été suivie d'invasion, et la population de chats harets était loin d'être stabilisée à la fin des années 70 [29, 102]. On

est donc en présence d'une histoire de fondation-explosion récente historiquement documentée, qui correspond en première approximation au modèle démographique supposé dans MSVAR.

7.2 Applicabilité des hypothèses sous-jacentes à MSVAR

D. Pontier (Laboratoire Biométrie et Biologive Évolutive, Université Lyon-1) m'a fourni un jeu de données concernant 192 individus adultes capturés dans 4 localités étendues, et typés à 9 loci microsatellites comme décrit dans l'article annexé page 189. On se reportera à cet article pour une analyse génétique complète. Je discute ici, sur la base de cette analyse et des connaissances historiques, de la pertinence des hypothèses de MSVAR pour une application sur la population de chats de Kerguelen. Ces hypothèses concernent l'histoire démographique et le fonctionnement de la population, donc le processus de coalescence des gènes, et le processus mutationnel des marqueurs microsatellites.

7.2.1 Forme de la courbe démographique

MSVAR suppose qu'une population close, panmictique, à l'équilibre mutation-dérive, connaît à une date t_f dans le passé une variation brutale d'effectif, suivie d'une variation exponentielle jusqu'à la date d'échantillonnage. Discutons ces hypothèses successivement.

La population actuelle dérive sans doute de chats en provenance de deux —voire trois— aires géographiques continentales éloignées : l'Afrique du Sud, Madagascar et peut-être la France. On a donc eu un mélange à la fondation. Ce mélange peut être responsable d'une diversité du groupe fondateur supérieure à celle que l'on attendrait pour un groupe fondateur tiré d'une population panmictique. En particulier, les fondateurs risquent de porter des allèles plus distants en taille que ne sont distants les modes d'une distribution à l'équilibre mutation-dérive. Il faudra se rappeler au moment de l'interprétation des résultats que l'on infère la taille d'une population idéale source de tous les fondateurs (conformément à ce qui est modélisé).

La réduction brutale d'effectif modélise l'événement de fondation, qui est équivalent à l'échantillonnage de quelques individus dans une grande population.

Les données historiques indiquent que l'invasion a été rapide, et qu'elle s'est poursuivie pendant au moins deux décennies. L'hypothèse de reprise immédiate de la croissance est donc justifiée. L'hypothèse d'un maintien de la croissance jusqu'à l'échantillonnage est plus discutable, mais la population invasive a connu un régime de croissance pendant au moins la moitié du temps écoulé depuis l'introduction.

L'invasion de l'île principale des Kerguelen par le chat s'est accompagnée d'une extension de son aire de colonisation, et la population insulaire est nettement structurée. Cet aspect spatial est ignoré dans MSVAR, qui suppose que la population se comporte comme une seule unité.

7.2.2 Modèle généalogique du coalescent standard

MSVAR suppose que le processus de coalescence des gènes est bien décrit par le coalescent de Kingman avec variations d'effectif. Dans une population d'effectif constant, on suppose valable l'approximation continue du coalescent de Kingman dès lors que la taille d'échantillon ne dépasse pas la racine carrée de l'effectif de la population [81]. Lorsque la population est de petite taille, le processus de coalescence n'est plus indépendant d'un locus à un autre (on ne peut plus ignorer que les gènes sont portés par des individus). À un locus donné, le nombre de lignées ancestrales

d'un échantillon n'est pas contraint, dans le coalescent standard, à être inférieur au nombre de gènes présents dans la population, puisque le coalescent n'utilise ce nombre de gènes que pour renormaliser l'échelle des temps. On risque donc d'échantillonner des généalogies ayant un nombre de lignées fondatrices supérieur au nombre total de gènes fondateurs (illustration au paragraphe 7.3.2).

Il faudrait explorer les conséquences du petit effectif fondateur sur les inférences démographiques. Pour cela, on peut simuler des données selon le processus de coalescence discret —génération par génération— et tester si MSVAR infère correctement les valeurs des paramètres démographiques sur ces données.

7.2.3 Modèles mutationnels SMM et TPM

Nous avons montré dans la partie II que les données sur des marqueurs mutant selon un TPM peuvent, si elles sont interprétées en supposant un SMM, comporter un faux signal de déséquilibre démographique. Nous avons de plus montré que modéliser un TPM permet de supprimer ce faux-signal. L'un des objectifs du traitement des données de diversité du chat des Kerguelen est l'estimation des paramètres d'un modèle mutationnel pour les marqueurs. Cette estimation ne peut être précise que si le modèle démographique décrit correctement l'histoire démographique, et si le modèle mutationnel saisit les caractéristiques importantes du processus mutationnel.

Au vu des résultats de la partie II, il me semble important de paramétrer la saturation d'effectif dans la population, pour que les informations démographiques *a priori* soient transposables aux valeurs des paramètres du modèle démographique. C'est à cette condition que l'on peut aborder la question de l'inférence populationnelle des paramètres de mutation.

7.3 Signatures génétiques de la fondation-explosion

7.3.1 Configuration des données de typage

La figure 7.1 représente la configuration normalisée aux 9 loci typés, tous individus regroupés en un unique échantillon. Plusieurs loci montrent une distribution de comptages alléliques très disjointe (fca8, fca23, fca43, fca78, fca90), ce qui sous SMM ou sous un TPM léger est une signature de fondation (voir figure 2.10 page 55). Le tableau 7.1 indique les valeurs de statistiques sommaires calculées pour chacun des 9 loci.

7.3.2 Simulation de données en utilisant l'information historique

La méthode de simulation du coalescent décrite au paragraphe 5.1 a été utilisée pour produire des jeux de données cherchant à mimer celles de la population de chats de Kerguelen. Les valeurs des paramètres choisies sont le résultat d'une synthèse des données historiques et démographiques ([29, 101, 126]. Le taux de mutation choisi est $\mu = 10^{-3}$ (pour un modèle SMM ou un TPM avec $p = 0.8$ et $j = 2.0$), le nombre de génomes haploïdes respectivement pour les populations source, fondée et échantillonnée est $N_2 = 20\,000$, $N_1 = \{4; 8\}$ (hypothèses de 2 et 4 chats fondateurs), $N_0 = 20\,000$, et le nombre de générations depuis la fondation $t_a = 15$. Avec $\theta = 2N_0\mu$, $r = N_0/N_1$, $a = N_0/N_2$ et $t_f = t_a/N_0$, cela donnerait un jeux de paramètres démographiques $(\theta, r, a, t_f) = (40, 5\,000, 1, 0.00075)$ pour l'hypothèse basse de 2 fondateurs et

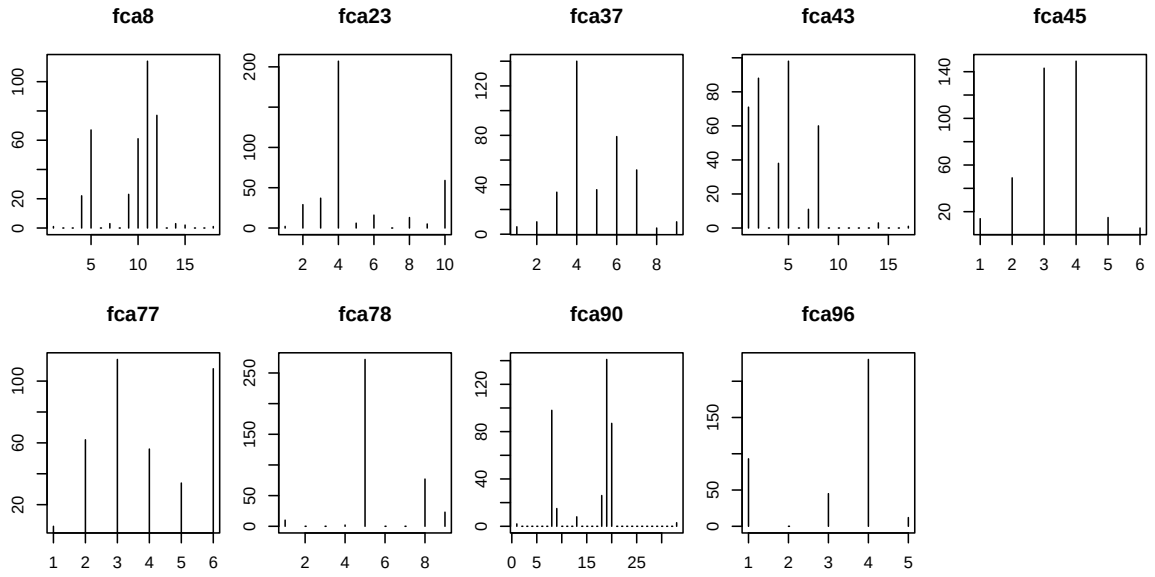


FIG. *7.1: Comptages alléliques à* 9 *loci microsatellites, pour un échantillon de* 192 *chats capturés sur l'île Kerguelen. Pour chaque histogramme : en abscisse, nombre de répétitions normalisé, en ordonnée, nombre de génomes haploïdes répertoriés pour chaque classe allélique.*

locus	n	A	V	P
fca8	374	11	8.21	0.20
fca23	374	9	6.26	0.35
fca37	372	9	2.57	0.23
fca43	370	8	7.09	0.20
fca45	376	6	0.88	0.32
fca77	380	6	2.33	0.23
fca78	384	5	2.66	0.55
fca90	380	8	29.41	0.26
fca96	380	4	1.68	0.44

TAB. *7.1: Statistiques sommaires calculées sur le jeu de données des chats de Kerguelen, pour chacun des* 9 *loci typés. Les* 192 *individus de l'échantillon ont été regroupés en une seule population, et n indique le nombre de génomes haploïdes typés avec succès pour chaque locus. A : nombre d'allèles dans l'échantillon. V : variance du nombre de répétitions. P : homozygotie calculée.*

de $(\theta, r, a, t_f) = (40, 2\,500, 1, 0.00075)$ pour l'hypothèse haute de 4 fondateurs. Près de 60% des coalescents simulés avec $r = 5000$ ont plus de 4 lignées non coalescées à t_f est , et le mode du nombre de fondateurs est 5, déjà supérieur au maximum souhaité de 4. Pour $r = 2500$, 50% des généalogies ont plus de 8 lignées fondatrices (avec un mode à 8 lignées). Il serait intéressant se savoir comment cela se traduit en terme de nombre d'allèles fondateurs.

Afin de corriger partiellement l'approximation continue d'un processus de coalescence discret en petite population, responsable de cet excès de fondateurs, M. Beaumont a suggéré de choisir pour r une valeur telle que la moyenne harmonique de la taille de population, sur la durée de la première génération, soit égale à 4 génomes haploïdes. Dans la pratique, cela revient à supposer un effectif plus faible au moment de la fondation, afin que la taille efficace sur la durée de la première génération soit la valeur cible (*e.g.* 4 individus). Le modèle de croissance exponentielle supposé dans MSVAR donne $N(t) = N_0 r^{-t/t_f}$, donc la moyenne harmonique m_h de $N(t)$ entre les temps t_a/N_0 et $(t_a - 1)/N_0$ (considérés à rebours du temps naturel) est

$$m_h = \left(- \int_{\frac{t_a}{N_0}}^{\frac{t_a-1}{N_0}} 1/N(t)\ dt \right)^{-1} = \frac{N_0 \log(r)}{t_a}\ \frac{r^{-(t_a-1)/t_a}}{r^{1/t_a-1}}.$$

Avec $t_a = 15$ générations et $N_0 = 20\,000$ génomes haploïdes, la valeur de r pour laquelle $m_h = 4$ est $r \simeq 6\,608$, et la valeur de r pour laquelle $m_h = 8$ est $r \simeq 4\,350$. Cela donne les jeux de paramètres L (hypothèse de 2 fondateurs) et H (4 fondateurs) donnés dans le tableau 7.2. Avec la correction, 27% des coalescents simulés sous le scénario L ont plus de 4 lignées fondatrices (ce sont pour les trois-quarts des coalescents avec 5 lignées fondatrices), et sous H, 1% des coalescents simulés ont plus de 8 lignées fondatrices. La correction harmonique réduit donc nettement la proportion de coalescents dans lesquels on a un excès de fondateurs, même si cette proportion reste forte pour le scenario L.

L'hypothèse d'une saturation de l'effectif à mi-parcours entre l'introduction et l'échantillonnage a été prise en compte en utilisant le modèle démographique avec délais (paragraphe 6.3.1), par les jeux de paramètres K_L et K_H (tableau 7.2). Une telle saturation de l'effectif est compatible avec les études démographiques de la population [126], et avec sa nette structuration. Enfin, un scénario de stabilité démographique S a été considéré pour comparaison, sous SMM uniquement (la diversité est déjà bien plus forte que dans le jeu de données réel, et un TPM empirerait cette situation). Les caractéristiques de données simulées en supposant ces trois scénarios sont résumées dans le tableau 7.2.

7.3.3 Vraisemblance des données pour des scénarios de fondation

Afin de comparer les différents scénarios, on a calculé pour chacun la probabilité de simulation d'un jeu de données de 9 loci ressemblant au jeu de données des Kerguelen. Plus précisément, le jeu de données est résumé par la distribution du nombre d'allèles, c'est à dire $D_A = (4, 5, 6, 6, 8, 8, 9, 9, 11)$ de l'homozygotie, ou par le logarithme népérien de la variance de taille allélique, $\log(V)$. L'homozygotie P a été arrondie à la valeur inférieure à 10^{-2} près ou 10^{-1} près, et on a supprimé les décimales du logarithme de V :
$D_{P:10^{-2}} = (0.20, 0.20, 0.23, 0.23, 0.26, 0.32, 0.35, 0.44, 0.55)$,
$D_{P:10^{-1}} = (0.2, 0.2, 0.2, 0.2, 0.2, 0.3, 0.3, 0.4, 0.5)$,
$D_{\log(V)} = (0, 0, 0, 0, 0, 1, 1, 2, 3)$.

Des distributions théoriques de ces statistiques pour les trois scénarios ont été estimées sur

modèle	θ	p	j	r	a	t_f	t_k	A	P
L^S	40	1	—	6608	1	0.00075	0	4.78	0.48
L^T	40	0.8	2.0	6608	1	0.00075	0	5.58	0.46
H^S	40	1	—	4350	1	0.00075	0	5.70	0.39
H^T	40	0.8	2.0	4350	1	0.00075	0	6.76	0.36
K_L^S	40	1	—	6608	1	0.00075	0.00035	6.84	0.33
K_L^T	40	0.8	2	6608	1	0.00075	0.00035	10.10	0.24
K_H^S	40	1	—	4350	1	0.00075	0.00035	7.92	0.27
S^S	40	1	—	1	1	—	—	16.08	0.11

TAB. *7.2: Jeux de valeurs des paramètres utilisées pour mimer l'histoire de la population de chats de Kerguelen. L et H supposent une courbe démographique telle que la taille de population croisse exponentiellement jusqu'à la date d'échantillonnage. L correspond à l'hypothèse de* 2 *individus fondateurs et H à celle de* 4 *individus fondateurs. K_L et K_H reprennent respectivement L et H et supposent une saturation de l'effectif à mi-temps entre la fondation et l'échantillonnage. L'exposant S ou T distingue l'hypothèse d'un processus mutationnel SMM ou TPM respectivement. S décrit une population stable ayant la même valeur de θ. La moyenne des statistiques A (nombre d'allèles), et P (homozygotie calculée) a été calculée sur* 100 000 *jeux de données simulés de taille $n = 380$.*

100 000 jeux de données simulés (de taille $n = 380$). La distribution continue de P a été discrétisée par arrondi à la valeur inférieure à 10^{-2} ou 10^{-1} près, et celle de $\log(V)$ à l'unité près. On appelle f_i la fréquence théorique de la classe i (pour l'une des trois statistiques). Les distributions D_A, D_P et $D_{\log(V)}$ sont considérées comme obtenues par $l = 9$ tirages indépendants (9 loci) selon les distributions théoriques, et les l_i sont le nombre de loci placés dans la classe i. On calcule alors la probabilité d'obtention de D_A, D_P et $D_{\log(V)}$ comme $\frac{l!}{\prod l_i!} \prod f_i^{l_i}$, où $\frac{l!}{\prod l_i!}$ doit être compris comme le nombre de permutations différenciables des loci (le nombre d'anagrammes de D_A, D_P et $D_{\log(V)}$). Le tableau 7.3 récapitule les résultats.

modèle	$p(D_A)$	$p(D_{P:10^{-2}})$	$p(D_{P:10^{-1}})$	$p(D_{\log(V)})$
L^S	$4.3\ 10^{-10}$	$2.9\ 10^{-14}$	$5.3\ 10^{-5}$	$4.6\ 10^{-7}$
L^T	$1.3\ 10^{-6}$	$3.6\ 10^{-13}$	$2.6\ 10^{-4}$	$1.7\ 10^{-10}$
H^S	$8.3\ 10^{-7}$	$4.1\ 10^{-11}$	$4.6\ 10^{-3}$	$6.7\ 10^{-8}$
H^T	$8.2\ 10^{-5}$	$1.8\ 10^{-10}$	$9.2\ 10^{-3}$	$3.3\ 10^{-13}$
K_L^S	$7.4\ 10^{-5}$	$3.8\ 10^{-10}$	$1.2\ 10^{-2}$	$9.8\ 10^{-9}$
K_L^T	$2.3\ 10^{-8}$	$1.6\ 10^{-11}$	$1.1\ 10^{-4}$	$2.7\ 10^{-20}$
K_H^S	$6.5\ 10^{-7}$	$2.0\ 10^{-10}$	$2.4\ 10^{-3}$	$2.6\ 10^{-10}$
S^S	$\simeq 0$	$\simeq 0$	$\simeq 0$	$5.7\ 10^{-25}$

TAB. *7.3: Probabilités $p(D_A)$ et $p(D_P)$ et $p(D_{\log(V)})$ des distributions D_A, D_P et $D_{\log(V)}$, pour les scénarios de fondation-explosion L, H et K et pour un scénario de stabilité S, définis par le tableau 7.2. Les probabilités $p(D_P)$ et $p(D_{\log(V)})$ dépendent du niveau de discrétisation de la statistique. Les résultats sont donnés pour un arrondi de P à 10^{-2} et à 10^{-1} près, et pour un arrondi de $\log(V)$ à l'entier près. Pour le scénario de stabilité S, certaines classes observées dans le jeu de données réel n'ont pas été atteintes une seule fois sur* 100000 *jeu de donnés simulés, d'où la valeur $\simeq 0$ pour les probabilités $p(D_A)$ et $p(D_P)$.*

Le scénario de stabilité S ne colle pas plus avec les observations génétiques qu'avec les données historiques. Parmi les scénarios de fondation-explosion comparés, ceux qui donnent

les plus fortes valeurs de $p(D)$ sont K_L^S et H^T. Le tableau 7.3 suggère deux remarques. La première est que *ce ne sont pas les mêmes scénarios qui donnent la plus forte vraisemblance aux observations des statistiques A, P et V*. En particulier, tous les scénarios supposant un TPM donnent en moyenne une trop forte variance de taille allélique, ce qui rend ces scénarios peu vraisemblables. À l'inverse, le TPM permet de simuler des jeux de données plus ressemblants avec les données réelles, pour ce qui est des statistiques A et P. La seconde remarque est que *la forme de la reprise démographique, le nombre de fondateurs et le modèle mutationnel interfèrent fortement pour déterminer la variabilité génétique de la population, ce qui rend difficile de les inférer*. Par exemple, selon que l'on considère un modèle TPM et une reprise démographique lente ou un modèle SMM et une reprise démographique rapide, les données en termes de A et de P ne sont pas en accord avec le même nombre de fondateurs (2 et 4 respectivement).

7.3.4 Détection d'un déséquilibre démographique par BOTTLENECK ?

Traitement du jeu de données réel

À l'aide du programme BOTTLENECK [24, 105], on a testé l'hypothèse d'un excès ou d'un déficit d'hétérozygotie calculée dans le jeu de données. Pour cela, l'hétérozygotie calculée a été comparée avec sa distribution attendue sous TPM (70% de SMM, variance 30 pour la paramétrisation utilisée dans le programme) conditionnellement au nombre d'allèles à chaque locus, et sous l'hypothèse d'un équilibre entre mutation et dérive. Le jeu de données complet (192 individus typés à 9 loci) a été traité, ainsi qu'un sous-jeu de données de 100 individus. Ces 100 chats ont tous été échantillonnés dans la même localité (Port-aux-Français) et avaient été assignés à un même cluster par la méthode d'assignation implémentée dans le programme STRUCTURE [107]. Cela évite de forts écarts à l'hypothèse de panmixie dans la population.

Ni le jeu de données complet, ni le sous-échantillon ne montrent d'écart significatif à l'équilibre mutation-dérive, selon un test du rang de Wilcoxon, au seuil 5%. Les P-values sont dans les deux cas de 0.08, pour l'hypothèse d'un excès d'hétérozygotie. On est donc proche de la significativité, et un jeu de données de plus grande taille permettrait peut-être de détecter un tel excès, signature d'un déclin.

Des observations similaires ont été faites sur un jeu de données provenant d'une population de wallabies introduite en Nouvelle-Zélande [83]. Dans ce cas comme dans celui des chats, la fondation-explosion est documentée sans grave erreur possible, et BOTTLENECK ne détecte pas de déséquilibre démographique, en tout cas sous TPM. Les auteurs de cette étude signalent que leurs distributions de fréquences alléliques présentent des lacunes (des allèles sont manquants entre les deux extrêmes). Nous avons vu que cela est assez caractéristique d'une fondation.

Une statistique simple à calculer sur les jeux de données, inutilisée dans BOTTLENECK, et qui devrait permettre de détecter les fondation-explosion, est la variance de taille allélique. L'index β de Kimmel utilise l'information de variance de taille allélique. Toutefois, cette variance est fortement dépendante du processus mutationnel, et un modèle TPM sans contraintes de taille donne facilement des distributions fantaisistes, surtout pour un modèle de fondation-explosion avec une grande taille de population ancestrale[1]. Il apparaît donc difficile de tenir compte de

[1]on remarquera que MSVAR ne suppose pas non plus de contrainte de taille pour le modèle TPM. Cette contrainte est en fait implicite, car les généalogies échantillonnées par MCMC sont toutes compatibles avec le jeu de données, et que des mutations saut trop divergentes devraient être compensées par un grand nombre de pas, ce qui diminue la vraisemblance des généalogies. Théoriquement, le fait —éventuel— que les mutations sauts soient préférentiellement dirigées vers la moyenne de taille allélique pourrait être un résultat de MSVAR

cette information dans les approches fréquentistes.

Traitement d'un jeu de données simulé plus important

Pour déterminer si la non-détection d'un déséquilibre est due à un trop petit nombre de loci, des jeux de données mimant celui des Kerguelen, mais de plus grande taille (20 loci, 192 individus) ont été simulés. On a pour cela supposé les deux scénarios qui maximisent $p(D_A)$ et $p(D_P)$ (car A et P sont les deux statistiques utilisées dans BOTTLENECK), c'est à dire les scénarios H^T et K_L^S. Pour le modèle H^T, qui suppose un TPM, les P-values sont sous SMM de 0.06 pour l'hypothèse d'un déficit d'hétérozygotie, et sous TPM de 0.04 pour l'hypothèse d'un excès d'hétérozygotie. Pour le modèle K_L^S, qui suppose un SMM, les P-values sont sous SMM de 0.001 pour l'hypothèse d'un déficit d'hétérozygotie, et sous TPM de 0.08 pour l'hypothèse d'un excès d'hétérozygotie. Il est troublant que les résultats dépendent si fortement du modèle mutationnel supposé dans BOTTLENECK, et surtout qu'il n'y ait pas de lien simple avec le modèle mutationnel effectivement utilisé pour simuler les données. Si l'on considère seulement les résultats de BOTTLENECK obtenus en supposant le bon modèle mutationnel, le jeu de données H^T permet de conclure à un excès d'hétérozygotie significatif, alors que le jeu de données K_L^S permet de conclure à un déficit d'hétérozygotie significatif.

Conclusion

Une histoire de fondation-explosion parait difficile à détecter à l'aide de BOTTLENECK, qui a été conçu essentiellement pour détecter des déclins, et non pas des bottlenecks transitoires. Nous allons chercher à déterminer si MSVAR, qui est basé sur la vraisemblance et peut paramétrer le déclin et l'explosion, distingue les signatures de la fondation et de l'explosion.

7.3.5 Détection d'un déséquilibre démographique par MSVAR

L'usage de MSVAR demande un temps et une puissance de calcul importants. En conséquence, pour une première exploration des données, on s'est limité à un sous-échantillon du jeu de données complet de 9 loci et 192 individus. Cet échantillon ne comporte que 4 loci et 50 gènes. Il n'a pas été sélectionné au hasard, mais de façon à limiter les écarts aux hypothèses de population close et panmictique sous-jacentes à MSVAR (*cf.* paragraphe suivant). Cela facilite l'interprétation des résultats.

Différents modèles démographiques ont été supposés, afin d'essayer de différencier les signatures de la fondation et de l'explosion démographique. (i) Un modèle de variation en marche d'escalier (déclin ou expansion, *cf.* paragraphe 7.3.6). (ii) Un modèle de variation exponentielle sans fondation initiale, identique à celui supposé dans la version initiale de MSVAR [6] (*cf.* paragraphe 7.3.7). (iii) Un modèle complet avec fondation et explosion (*cf.* paragraphe 7.3.8). Pour les trois types de familles de courbes démographiques, on a supposé soit un modèle mutationnel SMM, soit un modèle mutationnel TPM de paramètres fixes ($p = 0.8$, $j = 2$).

Choix d'un sous-échantillon

Pritchard, Stephens et Donnelly (2000) [108] ont décrit une méthode bayésienne —basée sur de l'échantillonnage MCMC— qui permet, à partir de génotypes à plusieurs loci non liés (données D), d'inférer un regroupement d'individus en populations, sans utiliser de regroupement a

priori. Cette méthode est inférée dans le programme STRUCTURE. Le modèle suppose que les loci sont à l'équilibre de Hardy-Weinberg et à l'équilibre de liaison, et la fission de l'échantillon complet en sous-échantillons est réalisée sur la base de l'écart à ces hypothèses. Les individus sont assignés de façon probabiliste à l'un des groupes, où à plusieurs si leurs génotypes indiquent qu'ils sont hybrides. On impose le nombre K de groupes à former, mais la méthode permet de comparer la vraisemblance des données pour différentes valeurs de K (et assiste donc l'utilisateur dans le chois de K).

On a imposé à la méthode de construire 4 groupes, et on a calculé la probabilité d'appartenance des individus à ces groupes, à partir de 1 000 000 itérations MCMC, et après 100 000 itérations non-reportées. Les groupes inférés correspondent bien aux 4 sites d'échantillonnage. Seulement 13 individus sur les 192 (10 sur 126 de Port-aux-Français, 1 sur 8 de Port-Couvreux et 1 sur 37 de Ratmanoff, *cf.* carte de Kerguelen dans l'article correspondant) ne sont pas assignés à leur groupe d'échantillonnage avec une probabilité supérieure à 0.5. Dans ce cas, ils sont identifiés comme migrants, et non pas comme descendants —hybrides— de migrants.

L'échantillon traité à l'aide de MSVAR correspond à 50 gènes aux loci fca77, fca78, fca90 et fca96, choisis par tirage uniforme sans remise, parmi les gènes des 116 individus assignés avec forte probabilité à la population de Port-aux-Français. La configuration normalisée de cet échantillon aux 4 loci est la suivante :
fca77 $= \{10, 13, 13, 2, 12\}$
fca78 $= \{41, 0, 0, 7, 2\}$
fca90 $= \{14, 4, 0, 0, 0, 1, 0, 0, 0, 0, 6, 21, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1\}$
fca96 $= \{11, 0, 8, 30, 1\}$

7.3.6 Histoire de variation d'effectif en marche d'escalier ?

On impose ici que l'histoire des variations d'effectif soit une marche d'escalier à une date t_f avant l'échantillonnage. Cela est permis dans MSVAR en fixant le paramètre de variation exponentielle $r = N_0/N_1$ à la valeur $r = 1$. Les paramètres θ, t_f et a sont libres de varier entre 10^-5 et 10^5, et on affecte une loi *a priori* uniforme sur $[-5; 5]$ à leur logarithme décimal. La figure 7.2 montre la bonne convergence, pour un espacement de 50 000 itérations entre positions échantillonnées, et illustre la loi *a posteriori* du paramètre a.

De façon intéressante (et assez inattendue), il est clairement inféré que la discontinuité démographique à t_f ne peut être importante. Cette discontinuité est décrite par le paramètre a. Si $log_{10}(a)$ vaut 0, on n'a un effectif constant, pour $log_{10}(a) > 0$, on a à la date t_f une expansion en marche d'escalier, et si $log_{10}(a) < 0$, on a un déclin en marche d'escalier. La loi *a posteriori* de a a son maximum pour $log_{10}(a) = 0$ —sous SMM aussi bien que sous TPM—, et l'intervalle HPD de probabilité 0.95 est à peu près $[-0.7, 0.7]$ (là aussi pour les deux modèles). On a donc le plus vraisemblablement une continuité de l'effectif à t_f, et avec probabilité 0.95, un rapport plus petit que 5 (en valeur absolue) entre la taille avant et après t_f.

On en conclut que les données ne soutiennent absolument pas une histoire de déclin (ou d'expansion) en marche d'escalier, mais pour le modèle démographique contraint par le choix de $r = 1$, une histoire de stabilité.

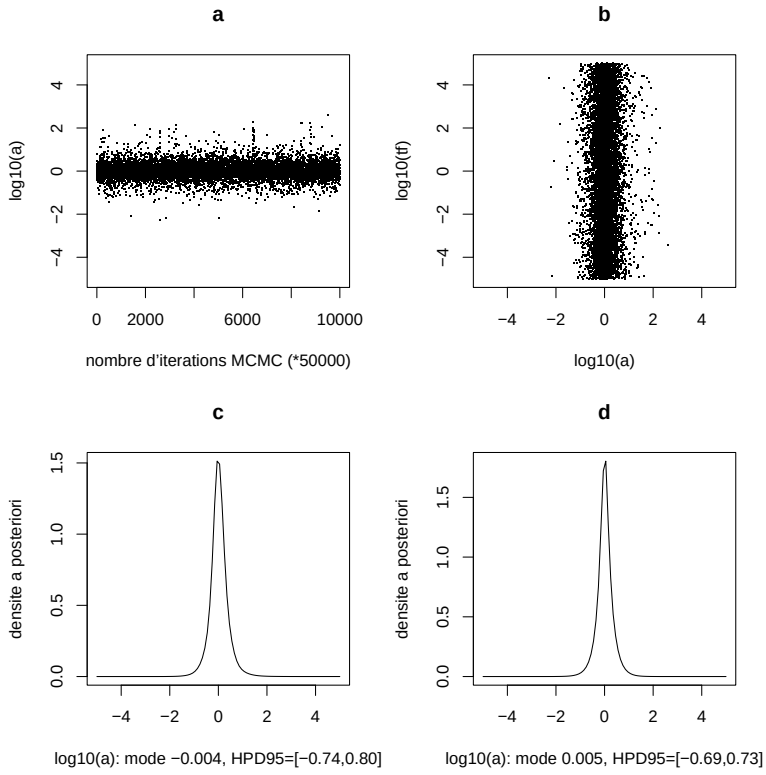


FIG. *7.2: Inférences démographiques pour un modèle en marche d'escalier défini par* $r = N_0/N_1 = 1$. **a** *: représentation de* a *en fonction du nombre d'itérations, pour le modèle SMM. L'autocorrélation entre points échantillonnés est faible pour un espacement de* $50\,000$ *: la convergence semble satisfaisante. On constate que* a *est distribué autour de la valeur* 1 *(*$\log_{10}(a) = 0$*).* **b** *: loi jointe* a posteriori *de* a *et* t_f*, pour le modèle SMM. Ces deux paramètres sont indépendants.* **c-d** *: loi marginale de* a *illustrée pour le modèle SMM (***c***) et le modèle TPM avec* $p = 0.8$ *et* $j = 2$ *(***d***). Le mode de la distribution marginale* a posteriori *est indiqué, ainsi que les bornes d'un intervalle de densité* a posteriori 0.95 *(*HPD_{95}*).*

7.3.7 Histoire de variation exponentielle sans fondation initiale ?

On suppose ici une histoire de variation d'effectif exponentielle depuis une date t_f avant l'échantillonnage, en continuité avec la taille de population ancestrale. Cela est permis dans MSVAR en imposant $a = r$ à chaque itération de la MCMC. Une loi *a priori* uniforme sur $[-5; 5]$ a été supposée pour $\log_{10}(\theta)$, $\log_{10}(r)$ et $\log_{10}(t_f)$.

La figure 7.3 montre la convergence insuffisante de la MCMC, pour un espacement de 50 000 itérations entre positions échantillonnées. Les interprétations sur la loi *a posteriori* des paramètres démographiques sont soumises à la confirmation des résultats avec une valeur d'espacement suffisamment grande pour assurer une meilleure convergence.

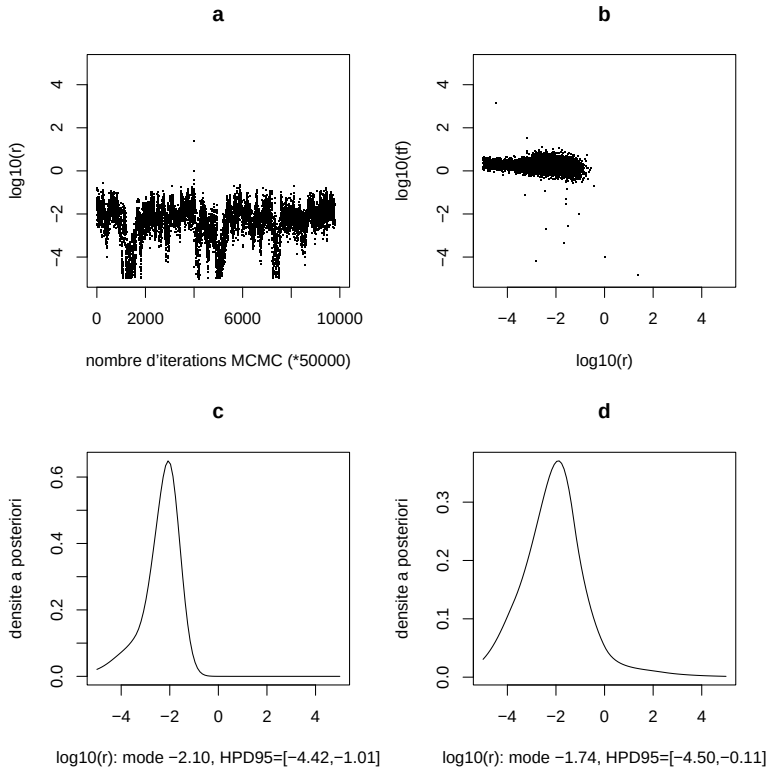


FIG. *7.3: Inférences démographiques pour un modèle de variation exponentielle de l'effectif défini par* $a = r$. **a** *: représentation de* r *en fonction du nombre d'itérations, pour le modèle SMM. La convergence n'est pas satisfaisante : la MCMC oscille entre deux régimes pour* r*. Néanmoins,* r *est en général inférieur à* 1. **b** *: loi jointe* a posteriori *de* r *et* t_f*, pour le modèle SMM. Ces deux paramètres sont indépendants.* **c-d** *: loi marginale de* r *illustrée pour le modèle SMM (***c***) et le modèle TPM avec* $p = 0.8$ *et* $j = 2$ *(***d***). Le mode de la distribution marginale* a posteriori *est indiqué, ainsi que les bornes d'un intervalle de densité* a posteriori 0.95 *(*HPD_{95}*).*

Lorsqu'un modèle de variation de taille exponentielle est supposé, la méthode probabiliste permet d'inférer un déclin (l'intervalle HPD_{95} pour r ne contient que des valeurs inférieures à 1, sous SMM aussi bien que sous TPM). Les valeurs inférées pour t_f sont quasiment indépendantes des valeurs inférées pour r, et sont proches de 1 (sous SMM aussi bien que sous TPM).

Les résultats obtenus pour le modèle en marche d'escalier et pour le modèle de variation exponentielle peuvent être interprétés comme suit. La fondation est un échantillonnage instantané

dans une distribution de fréquences alléliques caractéristique d'une grande population. La signature de cet échantillonnage —c'est à dire l'absence de certaines classes alléliques— est forte. La croissance exponentielle a des signatures bien plus subtiles, qui sont la présence de classes alléliques obtenues par mutation post-fondation. Cela peut expliquer que l'on infère un déclin lorsque l'on suppose une variation de taille exponentielle. De façon intéressante, l'échantillon est incompatible avec un déclin en marche d'escalier. Cela supposerait en effet une forte dérive dans la population fondée, supposée de faible effectif, et donc un nombre d'allèles bien plus faible que celui que l'on observe. Cette interprétation est insuffisante, car même pour t_f petit et/ou θ fort, a reste proche de 1. *Il y a sans doute trop d'allèles en faible fréquence pour une histoire de déclin en marche d'escalier : la signature de l'explosion doit empêcher de détecter le déclin.*

7.3.8 Histoire de fondation-explosion ?

Détecte-t-on les deux événements, si ils sont tous deux paramétrés ? On utilise ici le modèle démographique complet, paramétré par a, r, t_f et θ, et on suppose une loi *a priori* uniforme sur $[-5; 5]$ pour le logarithme décimal de chacun de ces paramètres.

La figure 7.4 montre la convergence insuffisante de la MCMC, pour un espacement de 50000 itérations entre positions échantillonnées (pour le modèle TPM). Les interprétations sur la loi *a posteriori* des paramètres démographiques sont soumises à la confirmation des résultats avec une valeur d'espacement suffisamment grande pour assurer une meilleure convergence. Du fait que la proportion relative de points dans les différentes zones est sujette à caution, je ne quantifierai pas ces proportions, et resterai très qualitative.

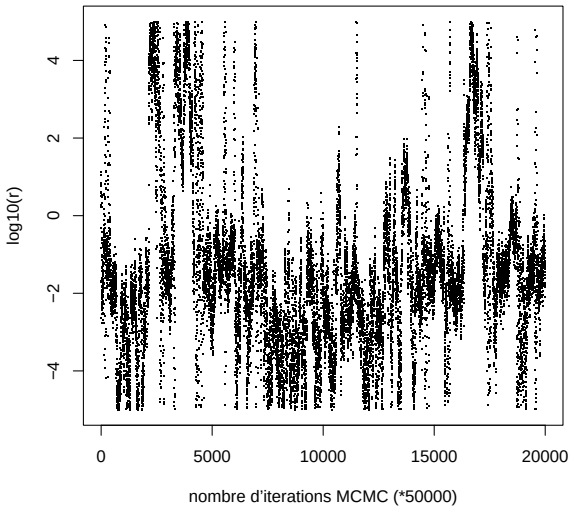


FIG. *7.4: Un espacement de* 50000 *ne permet pas d'atteindre un suffisamment faible niveau d'autocorrélation de la MCMC, pour la version multiparamétrée du modèle (illustration pour le modèle TPM). Toutefois, on verra que le type de la loi* a posteriori *est nettement P_d, ce qui ne changera vraisemblablement pas si l'on augmente l'espacement. Seule la proportion relative des points échantillonnés dans les différentes zones de plus forte densité de l'espace des paramètres devrait éventuellement changer.*

La figure 7.5 montre que la loi *a posteriori* obtenue pour l'échantillon de 50 gènes à 4 loci est nettement de type P_d, pour le modèle TPM (j'illustre celui-ci car la convergence est meilleure que pour le SMM, mais sous SMM aussi, la loi *a posteriori* est de type P_d). On peut pour comparaison se reporter à la figure 6.12 page 118). On aurait donc des signatures de déclin, suivi ou non d'explosion. Les valeurs des paramètres qui précédemment avaient servi à définir le scénario H^T (hypothèse de 4 fondateurs, sans saturation de l'effectif, et sous TPM), sont indiquées sur la figure 7.5. Ces valeurs sont effectivement dans des zones de forte de forte densité où à leur limite, mais on a bien d'autre scénarios vraisemblables.

Cela illustre un intérêt des approches bayésiennes, qui est qu'elles permettent de comparer un continuum de modèles. Dans une approche fréquentiste, on n'aurait sans doute même pas

cherché à savoir si des scénarios de pic-déclin donnent une forte vraisemblance aux données. On pourra répondre que l'on sait que la population a connu une fondation-explosion, et qu'il est inutile de rechercher des scénarios de pic-déclin par exemple. Ce choix, tout à fait justifiable, peut parfaitement être fait dans l'approche bayésienne, par le choix d'une loi a priori *à support dans le domaine des fondation-explosion.*

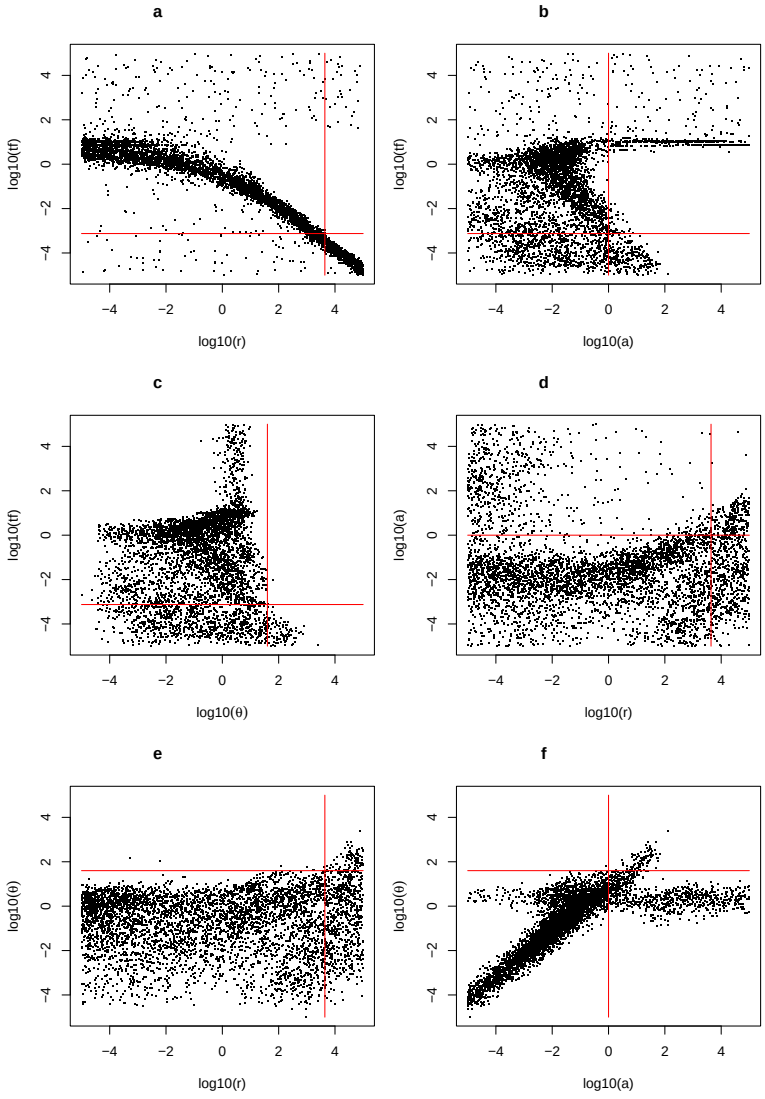


FIG. *7.5: Loi* a posteriori *de type P_d obtenue à partir d'un échantillon de* 50 *gènes et* 4 *loci, tiré du jeu de données complet sur les chats des Kerguelen. Le modèle mutationnel supposé est un TPM avec $p = 0.8$ et $j = 2.0$, et la loi* a priori *est uniforme sur $[-5; 5]$ pour le logarithme décimal de chacun des paramètres r, a, t_f et θ. L'échantillon MCMC représenté comporte* 5 000 *points. Les lignes rouges indiquent les valeurs des paramètres qui précédemment ont servi à définir le scénario H^T (hypothèse de* 4 *fondateurs, sans saturation de l'effectif, et sous TPM).*

S'il se confirme sur le jeu de données de 9 loci que la loi *a posteriori* est de type P_d, on ne sera malheureusement pas dans une situation favorable pour l'estimation du taux de mutation des marqueurs. En effet, même connaissant la date de la fondation (figure 7.5, cadran **c**), on aura seulement une borne supérieure pour le taux de mutation.

7.4 Conclusion

Les analyses préliminaires présentées dans ce chapitre permettent de proposer des améliorations de MSVAR, conditions de son utilisation à des fins inférentielles dans le cas de la population de chats des Kerguelen (et de beaucoup de populations invasives). Il me paraît important de paramétrer plus précisément la forme de la reprise démographique, par ajout de paramètres de délai avant reprise, et de saturation avant échantillonnage. En effet, même sans cette paramétrisation, la méthode permet d'inférer les paramètres du modèle. Mais la transposition des résultats en termes de processus démographique et mutationnel nécessite une bonne adéquation du modèle à la réalité.

Les analyses seront reprises, sur des jeux de données de plus grande taille, à l'aide d'une version en développement de MSVAR, modélisée hiérarchiquement. Cette version permet de manipuler directement les paramètres naturels (les effectifs, le nombre de générations depuis la fondation, le taux de mutation), au lieu des combinaisons de ces paramètres naturels utilisées dans la présente version. Elle permet en outre d'utiliser des lois *a priori* autres que rectangulaires. Si la loi *a posteriori* montre une forte corrélation entre les paramètres démographiques et les paramètres mutationnels, il sera possible de proposer une estimation du taux de mutation des marqueurs, voir des paramètres du modèle TPM.

Chapitre 8

Interactions entre une espèce invasive et une espèce autochtone : le cas de *Rattus norvegicus* et *Crocidura suaveolens* sur les îles bretonnes

Trois espèces du genre *Rattus* (*R. norvegicus*, *R. rattus* et *R. exulans*), commensales de l'homme, ont été introduites par lui dans plus de 80% des archipels du monde [3]. Les populations invasives de rats en milieu insulaire sont reconnues responsables de perturbations majeures des écosystèmes d'accueil (voir références citées par l'article [17], annexé page 187). Des études sont menées sur des populations insulaires de ces espèces, dans différentes provinces biogéographiques, afin de comprendre leur fonctionnement (thèse de J. Abdelkrim).

On se pose la question de savoir si les signatures génétiques du passé démographique des populations peuvent permettre de mettre en évidence des interactions entre les espèces introduites et des espèces autochtones (et plus généralement, entre deux espèces en interaction). Pour cela, on s'intéresse à l'espèce invasive *R. norvegicus* et à l'espèce *Crocidura suaveolens* qui sont en sympatrie sur certaines îles du plateau continental breton, et en allopatrie sur d'autres. Des études de structuration génétique sont menées à l'échelle régionale, et on cherchera à déterminer si les données génétiques dans les populations des deux espèces portent la trace d'un déclin de la musaraigne *C. suaveolens*, contemporain de l'invasion par *R. norvegicus*.

8.1 Hypothèses sur l'origine des populations

8.1.1 *Crocidura suaveolens*, isolée au moment de l'insularisation ?

Deux musaraignes du genre *Crocidura* (*C. russula* et *C. suaveolens*) occupent un grand nombre d'îles continentales des côtes de la Manche et de l'Atlantique, en Europe de l'ouest [25]. La répartition de ces deux espèces sur les îles du Ponant, égrenées le long des côtes françaises depuis le Cotentin jusqu'à l'estuaire de la Gironde, est très imbriquée. Ces deux espèces s'excluent mutuellement sur les îles, mais peuvent occuper des îles proches. La répartition et l'abondance de ces deux espèces sur le continent est très inégale. *C. russula* est abondante et largement répandue sur tout le littoral ouest européen du Portugal aux Pays-Bas. Elle est notamment abondante sur le littoral breton. *C. suaveolens* est beaucoup plus rare et présente une

distribution très sporadique le long des côtes atlantiques françaises. En Bretagne, elle n'est signalée que dans des pelotes de chouette effraie du Morbihan, en proportion extrêmement faible (moins de 1 pour 4000 proies).

J. F. Cosson *et al.* (1996) [25] discutent d'un scénario pouvant expliquer cette répartition. Toutes les îles du plateau continental ouest européen se sont insularisées à partir de 12 000 BP, à la fin du dernier épisode glaciaire. Les espèces vivant à l'époque dans ces régions —dont très vraisemblablement *C. suaveolens*— auraient été isolées par la montée des eaux pour former les populations insulaires actuelles. *C. russula*, problablement originaire de la péninsule ibérique, n'est arrivée sur les côtes de la Manche qu'autour de 6 000 BP et n'aurait pu atteindre que les îles dernièrement insularisées, où elle aurait pris l'avantage sur *C. suaveolens* comme elle l'a fait sur le continent. Des remaniements ultérieurs à l'insularisation pourraient expliquer la présence de *C. russula* sur des îles insularisées précocément. Elles seraient dues à des migrations spontanées ou à des introductions involontaires par l'homme.

Les populations insulaires de *C. suaveolens* seraient selon ce scénario des populations relictuelles isolées des populations continentales par la dernière transgression post-glaciaire.

8.1.2 *Rattus norvegicus*, une espèce introduite

Rattus norvegicus et *Rattus rattus* sont des espèces continentales originaires du Moyen-Orient. Commensales de l'homme, elles ont été dispersées efficacement par lui, et occupent actuellement plus de 80% des archipels du monde. *R. norvegicus*, le surmulot, est le plus fréquent sur les îles des côtes ouest européennes. En 1994, il était présent sur un grand nombre d'îles du Ponant, notamment en Bretagne. Le surmulot a un fort impact sur la faune insulaire, et en particulier sur l'avifaune à nidation sous-terraine. Cela a conduit à l'éradiquer d'un certain nombre d'îles, en particulier des îles réserves. Une telle mesure nécessite au préalable d'apprécier le caractère exogène de l'espèce, et les conséquences indésirables que sa disparition pourrait engendrer pour son écosystème d'accueil. Les plus anciennes traces du surmulot en Europe datent seulement du $16^{\text{ième}}$ siècle, et ce n'est qu'au $18^{\text{ième}}$ siècle qu'il a envahi massivement l'Europe de l'ouest. Cela est beaucoup plus tardif que l'insularisation des îles bretonnes. Le surmulot est donc un occupant récent de ces îles, très vraisemblablement introduit par l'homme (sauf peut-être pour les îles les plus proches du continent, qui peuvent avoir été colonisées à la nage). Les risques écologiques liés à l'éradication sont apparus dérisoires par rapport aux dégâts attribués au surmulot dans la littérature [103]. *R. norvegicus* a donc été éliminé en 1994 de l'archipel des Sept-îles (réserve ornithologique), et de l'archipel de Cancale. Puis en 1996 il a été éradiqué de Trielen et de l'île aux Chrétiens dans l'archipel de Molène. Les éradications ont été durables, sauf dans le cas de Cancale, extrêmement proche du continent. Le typage d'individus capturés cette année par l'ONC permettra certainement de déterminer si la population de surmulots provient d'individus immigrants ou de rescapés de l'opération d'éradication.

8.2 Structuration génétique à l'échelle régionale

L'article annexé page 187 présente les résultats d'une analyse de diversité génétique réalisée au cours de mon DEA, sur les îles de la mer d'Iroise (Finistère, France). Une étude du même type est en cours de réalisation, à l'échelle régionale (thèse de J. Abdelkrim). Ces études montrent une différenciation extrêmement forte entre les populations insulaires, à l'exception des populations d'îles reliées à marée-basse, qui peuvent être génétiquement proches. Certaines îles montrent une

diversité génétique très réduite par rapport à la diversité continentale. Cela peut s'interpréter par un petit nombre d'individus fondateurs à l'origine des populations, et/ou par une dérive génétique dans les populations insulaires qui sont isolées et de petite taille. Pour les populations qui ont été éradiquées, on a une estimation précise du nombre d'individus au moment de l'éradication. Cela doit permettre de différencier l'hypothèse d'une perte de diversité par dérive post-fondation, où par échantillonnage à la fondation.

8.3 Inférences sur une fondation-explosion

Je vais montrer ici des analyses tout à fait préliminaires, réalisées avec MSVAR sur un jeu de données de 48 surmulots (espèce *R. norvegicus*) capturés sur l'île aux chevaux. Cet ilôt localisé entre le continent et Belle-île présente une forte densité en surmulots. La population est très peu variable aux 8 loci typés. En effet, 5 loci sont monomorphes (configuration $\{48\}$), et 3 loci comportent deux allèles dans l'échantillon, selon les configurations $\{5, 0, 43\}$, $\{20, 0, 24\}$ et $\{45, 0, 3\}$. La loi *a posteriori* pour cet échantillon de 8 loci a été obtenue en supposant un modèle mutationnel SMM, et une loi *a priori* uniforme sur $[-5; 5]$ pour le logarithme décimal de r, a, t_f et θ (figure 8.1). L'espacement entre points d'échantillonnage est de 100 000, ce qui assure une convergence raisonnable (3 simulations indépendantes réalisées en partant de 3 valeurs initiales différentes du vecteur de paramètre ϕ ont donné des résultats semblables, même si à l'oeil, l'évolution des valeurs des paramètres en fonction du nombre d'itérations montre une nette autocorrélation de la MCMC).

La loi *a posteriori* est nettement de type P_d, et soutient assez fortement un scénario de fondation-explosion (73% des points de l'échantillon MCMC correspondent à des fondation-explosion (points représentés en noir sur la figure 8.1). Ce résultat est encourageant, d'autant qu'il semble valable également pour les autres îles traitées à ce jour (non-illustré pour des raisons de convergence insuffisante).

Des simulations et des analyses plus poussées seront reprises lorsque les données de diversité génétique régionales auront été traitées par les méthodes classiques d'analyse de la diversité génétique (thèse de J. Abdelkrim). Comme dans le cas du chat, il sera au préalable nécessaire de paramétrer la saturation de population, qui a sans doute eu lieu longtemps avant l'échantillonnage. En effet, si l'on extrapole aux populations naturelles le taux de croissance des populations de laboratoire, une dizaine de générations de reproduction suffit, à partir d'une femelle gestante, pour atteindre les effectifs observés au moment de l'éradication des populations (environ 200 individus par exemple sur l'île Trielen), alors que quelques centaines de générations ont pu s'écouler pour les îles les plus anciennement colonisées.

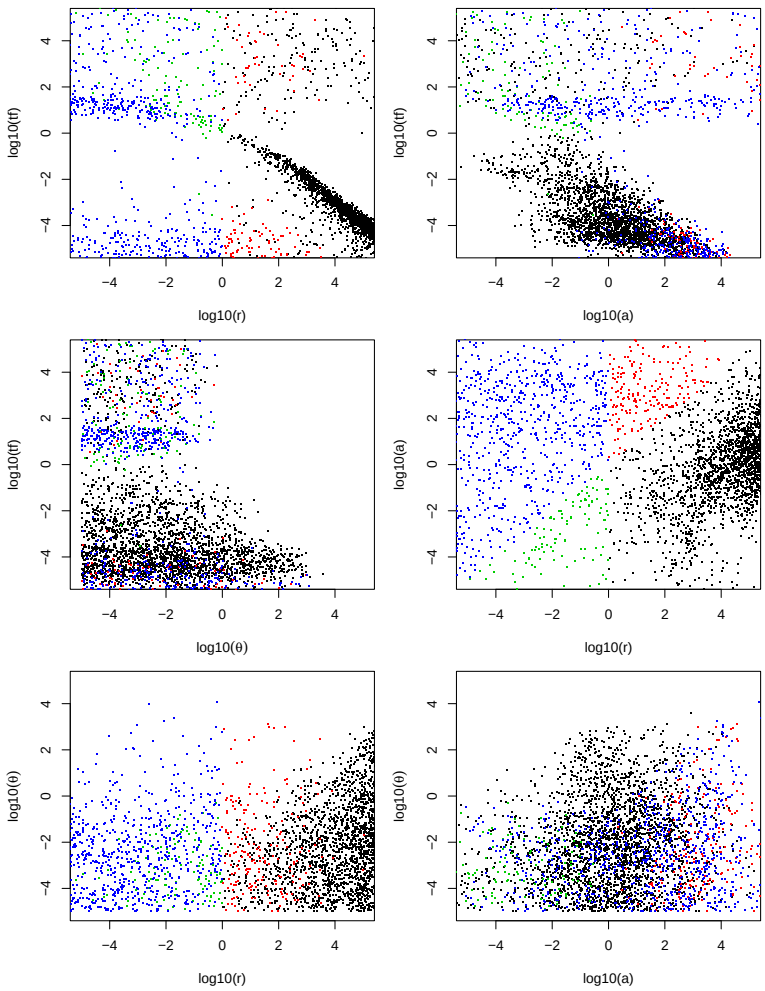


FIG. *8.1: Loi* a posteriori *de type P_d obtenue à partir d'un échantillon de* 48 *gènes et* 8 *loci, pour la populations de surmulots* R. norvegicus *de l'île aux Chevaux (Bretagne). Le modèle mutationnel supposé est un SMM, et la loi* a priori *est uniforme sur* $[-5;5]$ *pour le logarithme décimal de chacun des paramètres* r, a, t_f *et* θ. *L'échantillon MCMC représenté comporte* 5000 *points. Les données soutiennent assez fortement un scenario de fondation-explosion (*73% *des points de l'échantillon* a posteriori*).*

8.4 Interactions entre *C. suaveolens* et *R. norvegicus*

Les populations insulaires de la musaraigne *C. suaveolens* ont été échantillonnées dans les îles du plateau continental breton, en allopatrie ou en sympatrie avec le rat selon les îles. Des marqueurs microsatellites ont été mis au point sur l'espèce, certains par amplification de marqueurs définis chez l'espèce proche *C. russula* (3 marqueurs), d'autres par clonage chez *C. suaveolens* (5 marqueurs). Le typage des populations insulaires est en cours, et montre une diversité réduite (5 allèles au locus le plus variable), sur toutes les îles échantillonnées. Il n'est donc pas certain que les populations, sans doute de petite taille, aient conservé des signatures génétiques d'un éventuel déclin au moment de l'introduction du rat sur certaines îles. En effet, si le temps de coalescence des gènes est inférieur au temps écoulé depuis le déclin de la musaraigne, les généalogies de gènes ne remonteront pas jusqu'à ce déclin. La possibilité de vérifier la synchronisation entre le déclin de la musaraigne et l'introduction du rat est soumise à une coalescence des gènes antérieure au déclin pour la musaraigne.

Conclusion

Il sera nécessaire de paramétrer la saturation de l'effectif des populations, pour un usage de MSVAR à des fins inférentielles. On savait au début du développement de la méthode que les populations de rats avaient dû rapidement saturer. Pourquoi dans ces conditions ne pas avoir directement paramétré le plateau, d'autant que supposer un plateau est simple pour ce qui est du calcul de la vraisemblance (*cf.* modèle de fondation avec délais) ? Un modèle avec plateau n'a pas été choisi d'emblée parce qu'un problème se pose dans la gestion des lois *a priori*. En effet, les dates de la fondation (t_f), du début et de l'explosion (t_e) et de la saturation (t_k) sont contraintes à rester ordonnées selon l'inégalité $t_f \leq T_e \leq T_k \leq 0$. Les dates ne peuvent donc varier que dans l'intervalle délimité par les autres dates. La loi *a priori* d'une des dates est donc une fonction complexe, qui dépend de la loi *a priori* des autres dates, et non pas un créneau. Il sera plus facile de gérer cela avec la version hiérarchique —en développement— de la méthode, qui permet de choisir des lois *a priori* de formes autres que des créneaux.

Perspectives

Nous avons vu que l'application de MSVAR à des fin d'inférence dans des populations d'espèces invasives nécessite encore quelques améliorations, essentiellement la paramétrisation d'une saturation de l'effectif des populations. Mais le présent travail apporte également une amélioration importante du modèle mutationnel supposé : il permet de tenir compte de l'existence de mutations sauts pour les microsatellites, grâce à un modèle TPM. Il serait intéressant de vérifier si des inférences réalisées précedemment en supposant un SMM sont remises en cause si l'on suppose un TPM, modèle plus conforme à ce que l'on connait du processus mutationnel des marqueurs microsatellites.

Sur le plan méthodologique, nous nous sommes heurtés à des problèmes de convergence de MCMC. Ces problèmes sont intrinsèques à la méthode. En effet, l'usage de *l'algorithme* de Metropolis-Hastings nécessite d'effectuer de petits déplacements dans l'espace des généalogies et des paramètres. Pour des jeux de données de grande taille, l'exploration de l'espace des généalogies devient rapidement problématique.

Pourquoi ne pas utiliser, comme le suggèrent M. Stephens et P. Donnelly [138], un schéma d'Importance Sampling (IS) pour échantillonner les généalogies de gènes, et suppléer à la lenteur de l'algorithme de Metropolis-Hastings pour cette tâche ? On continuerait à utiliser l'algorithme de Metropolis-Hastings pour se déplacer dans l'espace des paramètres démographiques et mutationnels, et le *critère* de Metropolis-Hastings, pour accepter ou rejeter les déplacements dans l'espace des généalogies et des paramètres. Ce critère assurerait la convergence de la loi *a posteriori* obtenue par la méthode, vers la loi *a posteriori* exacte (ce qui n'est pas le cas pour les méthodes IS non couplées au critère de Metropolis-Hastings). Une méthode hybride de ce type a été proposée par M. Beaumont (article soumis, avec une application au problème spécifique de l'estimation de la taille efficace dans les populations). Elle serait généralisable au modèle étudié dans cette thèse, et permettrait de s'affranchir des problèmes de convergence, et donc de traiter des échantillons de plus grande taille.

Bibliographie

[1] Abrams, P. Evolution and the consequences of species introductions and deletions. *Ecology 77(5)* (1996), 1321–1328.

[2] Aris-Brosou, S., and Excoffier, L. The impact of population expansion and mutation rate heterogeneity on DNA sequence polymorphism. *Molecular Biology and Evolution 13(3)* (1996), 494–504.

[3] Atkinson, I. The spread of the commensal species of *Rattus* to oceanic islands and their effects on island avifauna. In *Conservation of island birds, International Council for Bird Preservation*, P. Moors, Ed., vol. 3. I.C.B.P Technical publication, Cambridge, UK, 1985, pp. 35–81.

[4] Austerlitz, F. Variations sur l'impact des processus démographiques sur la diversité génétique. Thèse de doctorat de l'Ecole Nationale du Génie Rural et des Eaux et Forêts, 1999.

[5] Barton, N. H. Founder effect speciation. In *Speciation and its consequences*, D. Otte and J. A. Endler, Eds. Sinauer, Sunderland, MA, 1989, pp. 229–256.

[6] Beaumont, M. A. Detecting population expansion and decline using microsatellites. *Genetics 153* (1999), 2013–2029.

[7] Beaumont, M. A. Recent developments in genetic data analysis : what can they tell us about human demographic history. Heredity, 2002.

[8] Beaumont, M. A., Zhang, W., and J, B. D. Approximate bayesian computation in population genetics. To appear in Genetics.

[9] Bell, G., and Jurka, J. The length distribution of perfect dimer repetitive DNA is consistent with its evolution by an unbiased single-step mutation process. *Journal of Molecular Evolution 44* (1997), 414–421.

[10] Berger, J. O., Liseo, B., and Wolpert, R. L. Integrated likelihood methods for eliminating nuisance parameters. *Statistical Sciences 14(1)* (1999), 1–28.

[11] Berry, R. Diversity and differenciation : the importance of island biology for general theory. *Oikos 41* (1983), 523–529.

[12] Berry, R. J. Darwin was astonished. *Biological Journal of the Linnean Society 21* (1984), 1–4.

[13] Berthier, P., Beaumont, M. A., Cornuet, J.-M., and Luikart, G. Likelihood-based estimation of the effective population size using temporal changes in allele frequencies : a genealogical approach. *Genetics 160* (2002), 741–751.

[14] Best, N. G., Cowles, M. K., and Vines, S. K. CODA Manual version 0.30. MRC Biostatistics Unit, Cambridge, 1995.

[15] Brooks, S. Markov Chain Monte Carlo method and its application. *The Statistician 47* (1998), 69–100.

[16] Brooks, S. P., and Gelman, A. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics 7* (1998), 434–455.

[17] Calmet, C., Pascal, M., and Samadi, S. Is it worth eradicating the invasive pest *Rattus norvegicus* from molène archipelago ? genetic structure as a decision making tool. *Biodiversity and Conservation 10(6)* (2001), 911–928.

[18] Carbone, I., and Kohn, M. A microbial population-species interface : nested cladistic and coalescent inference with multilocus data. *Molecular Ecology 10* (2001), 947–964.

[19] Carson, H. L. The genetics of speciation at the diploid level. *American Naturalist 109* (1975), 83–92.

[20] Carson, H. L. Speciation as a major reorganization of polygenic balances. In *Mecanisms of speciation*, C. Barigozzi, Ed. Liss, New-York, 1982, pp. 411–433.

[21] Carson, H. L., and Templeton, A. R. Genetic revolutions in relation to speciation phenomena : the founding of new populations. *Annual Review of Ecology and Systematics 15* (1984), 97–131.

[22] Cohen, J. The Earth is round ($p<0.5$). *American Psychologist 49* (1994), 997–1003.

[23] Cooper, G., Burroughs, N. J., Rand, D. A., Rubinsztein, D. C., and Amos, W. Markov Chain Monte Carlo analysis of human Y-chromosome microsatellites provides evidence of biased mutation. *Proceedings of the National Academy of Sciences of the United States of America 96(21)* (1999), 11916–11921.

[24] Cornuet, J.-M., and Luikart, G. Description and power analysis of two tests for detecting recent population bottlenecks. *Genetics 144* (1996), 2001–2014.

[25] Cosson, J.-F., Pascal, M., and Bioret, F. Origine et répartition des musaraignes du genre *Crocidura* dans les îles bretonnes. *Vie et Milieu, Life and Environment 46(3/4)* (1996), 233–244.

[26] Dallas, J. Estimation of microsatellite mutation rates in recombinant inbred strains of mouse. *Mammalian Genome 3* (1992), 452–456.

[27] Darwin, C. *On the origin of species by means of natural selection or the preservation of favoured races in the struggle for life.* John Murray, London, 1859.

[28] Deka, R., Guangyun, S., Smelser, D., Zhong, Y., Kimmel, M., and Chakraborty, R. Rate and directionality of mutations and effects of allele size constraints at anonymous, gene-associated and disease-causing trinucleotide loci. *Molecular Biology and Evolution 16(9)* (1999), 1166–1177.

[29] Derenne, P. Notes sur la biologie du chat haret de kerguelen. *Mammalia 40* (1976), 531–595.

[30] Di Rienzo, A., Donnelly, P., Toomajian, C., Sisk, B., Hill, A., Petzl-Erler, M. L., Haines, G. K., and Barch, D. H. Heterogeneity of microsatellite mutations within and between loci, and implications for human demographic histories. *Genetics 148* (1998), 1269–1284.

[31] Di Rienzo, A., Peterson, A. C., Garza, J. C., Valdes, A. M., Slatkin, M., and Freimer, N. B. Mutational processes of simple-sequence repeat loci in human populations. *Proceedings of the National Academy of Sciences of the United States of America 91* (1994), 3166–3170.

[32] Donnelly, P., and Simon, T. Coalescents and genealogical structure under neutrality. *Annual Review of Genetics 29* (1995), 401–421.

[33] Easteal, S. The history of introductions of *Bufo marinus* (Amphibia, Anoura) ; a natural experiment in evolution. *Biological Journal of the Linnean Society 16* (1981), 93–113.

[34] Eldredge, N., and Gould, S. J. Punctuated equilibria : an alternative to phyletic gradualism. In *Models in paleobiology*, T. J. M. Schopf, Ed. Freeman, Cooper, San Francisco, 1972, pp. 82–115.

[35] Ellegren, H., Primmer, C., and Sheldon, B. Microsatellite evolution : directionality or bias ? *Nature Genetics 11* (1995), 360–362.

[36] Ellengren, H. Heterogeneous mutation processes in human microsatellite DNA sequences. *Nature Genetics 24* (2000), 400–402.

[37] Estoup, A., and Cornuet, J.-M. Microsatellite evolution : inferences from molecular data. In *Microsatellites : evolution and applications*, D. B. Goldstein and C. Schlötterer, Eds. Oxford University Press, Oxford, 1999, pp. 50–65.

[38] Estoup, A., et al. Size homoplasy and mutational processes of interrupted microsatellites in two bee species, *Apis mellifera* and *Bombus terrestris* (apidae). *Molecular Biology and Evolution 12* (1995), 1074–1084.

[39] Estoup, A., Wilson, I. J., Sullivan, C., Cornuet, J.-M., and Moritz, C. Inferring population history from microsatellite and enzyme data in serially introduced cane toads *Bufo marinus*. *Genetics 159* (2001), 1671–1687.

[40] Ewens, W. *Mathematical population genetics*. Springer-Verlag, New-York, 1979.

[41] Fay, J., and Wu, C. A human population bottleneck can account for the discordance between patterns of mitochondrial versus nuclear DNA variation. *Molecular Biology and Evolution 16* (1999), 1003–1006.

[42] Fearnhead, P., and Peter, D. Estimating recombination rates from population genetic data. *Genetics 159* (2001), 1299–1318.

[43] Feldman, M. W., Bergman, A., Pollock, D., and Goldstein, D. B. Microsatellite genetic distance with range constraints : analytic description and problems of estimation. *Genetics 145* (1997), 207–216.

[44] Fritts, T. H., and Rodda, G. H. The role of introduced species in the degradation of island ecosystems : a case history of guam. *Annual Review of Ecology and Systematics 29* (1998), 113–140.

[45] Fu, Y.-X., and Li, W. Statistical test of neutrality of mutations. *Genetics 133* (1993), 693–709.

[46] Galtier, N., Depaulis, F., and Barton, N. H. Detecting bottlenecks and selective sweeps from DNA sequence polymorphism. *Genetics 155* (2000), 981–987.

[47] Garza, J. C., Slatkin, M., and Freimer, N. B. Microsatellite allele frequencies in human and chimpanzees, with implications for constraints on allele size. *Molecular Biology and Evolution 12(4)* (1995), 594–603.

[48] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. *Bayesian data analysis.* Chapman and Hall/CRC, 1995.

[49] Gelman, A., and Rubin, J. B. Inference from iterative simulation using multiple sequences. *Statistical Science 7* (1992), 457–472.

[50] Gilks, W. R., Richardson, S., and Spiegelhalter, D. J., Eds. *Markov Chain Monte Carlo in practice.* Chapman & Hall, London/New York, 1996.

[51] Goldstein, D. B., and Schlötterer, C., Eds. *Microsatellites : evolution and applications.* Oxford University Press, Oxford, 1999.

[52] Goldstein, D. B., Zhivotovsky, L. A., Nayar, K., Linares, A. R., Cavalli-Sforza, L. L., and Feldman, M. W. Statistical properties of the variation at linked microsatellite loci : Implications for the history of human Y chromosome. *Molecular Biology and Evolution 13(9)* (1996), 1213–1218.

[53] Gonser, R., Donnelly, P., Nicholson, G., and Di Rienzo, A. Microsatellite mutations and inferences about human demography. *Genetics 154* (2000), 1793–1807.

[54] Gould, S. J. Is a new and general theory of evolution emerging? *Paleobiology 6* (1980), 119–130.

[55] Griffith, R., and Tavaré, S. Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society of London. Series B 344* (1994), 403–410.

[56] Griffith, R., and Tavaré, S. Simulating probability distributions in the coalescent. *Theoretical Population Biology 46* (1994), 131–159.

[57] Griffiths, R. C. Lines of descent in the diffusion approximation of neutral Wright-Fisher models. *Theoretical Population Biology 17* (1980), 37–50.

[58] Grimmeth, S. *Probability and Random Processes.* Oxford, 1992.

[59] Hagen, R. L. In praise of the null hypothesis statistical test. *American Psychologist 52* (1997), 15–24.

[60] Harpending, H. C., Batzer, M. A., Gurven, M., Jorde, L. B., Rogers, A. R., and Sherry, S. T. Genetic traces of ancient demography. *Proceedings of the National Academy of Sciences of the United States of America 95* (1998), 1961–1967.

[61] Hartl, D. L., and Clark, A. G. *Principles of population genetics*, third edition ed. Sinauer, 1997.

[62] Hastings, W. K. Monte Carlo sampling methods using Markov Chains and their applications. *Biometrika 57(1)* (1970), 97–109.

[63] Huang, Q. Y., Xu, F. H., Shen, H., Deng, H. Y., Liu, Y. J., Li, J. L., Recker, R. R., and Deng, H. W. Mutation patterns at dinucleotide microsatellite loci in humans. *American Journal of Human Genetics 70(3)* (2002), 625–634.

[64] Hudson, R. R. Gene genealogies and the coalescent process. In *Oxford Surveys in Evolutionnary Biology*, vol. 7. Oxford University Press, 1990, pp. 1–44.

[65] Ihaka, R., and Gentleman, R. R : a language for data analysis and graphics. *Journal of Computational and Graphical Statistics 5(3)* (1996), 299–314.

[66] Jarne, P., and Lagoda, P. J. Microsatellites, from molecules to populations and back. *Trends in Ecology and Evolution 11* (1996), 424–429.

[67] Johnson, N. L., Kotz, S., and Balakrishnan, N. *Discrete multivariate distributions.* John Wiley & Sons, 1997.

[68] Johnstone, G. Threats to birds on subantarctic islands. In *Conservation of island birds. Case studies for the management of threatened islans species*, P. Moors, Ed., vol. 3. International Council for bird preservation, I.C.B.P Technical publication, Cambridge, UK, 1985, pp. 101–121.

[69] Jouventin, P., Stahl, J., H, W., and Mougin, J. The seabirds of the french subantarctic islands and adélie land, their status and conservation. In *Status and conservation of the world seabirds*, S. R. J.P. Croxall, P.G.H Evans, Ed., vol. 2. International Council for bird preservation, I.C.B.P Technical publication, Cambridge, UK, 1984, pp. 609–625.

[70] Kim, I., Phillips, C., Monjeau, J., Birney, E., Noacks, K., Pumo, D., Sikes, R., and Dole, J. Habitat islands, genetic diversity and gene flow in a patagonian rodent. *Molecular Ecology 7* (1998), 667–678.

[71] Kimmel, M., Chakraborty, R., King, J. P., Bamshad, M., Watkins, W. S., and Jorde, L. B. Signatures of population expansion in microsatellite repeat data. *Genetics 148* (1998), 1921–1930.

[72] Kimura, M. Solution of a process of random genetic drift with a continuous model. *PNAS 41* (1955), 144–150.

[73] Kimura, M., and Crow, J. F. The measurement of effective population numbers. *Evolution 17* (1963), 279–288.

[74] King, J. P., Kimmel, M., and Chakraborty, R. A power analysis of microsatellite based statistics for inferring past population growth. *Molecular Biology and Evolution 17(12)* (2000), 1859–1868.

[75] Kingman, J. The coalescent. *Stochastic Processes and their Applications 13* (1982), 235–248.

[76] Kingman, J. Exchangeability and the evolution of large populations. In *Exchangeability in probability and statistics*, G. Koch and F. Spizzichino, Eds. Amsterdam, North-Holland, 1982, pp. 97–112.

[77] Kingman, J. F. C. On the genealogy of large populations. *Journal of Applied Probability 19A* (1982), 27–43.

[78] Kingman, J. F. C. Origins of the coalescent : 1974-1982. *Genetics 156* (2000), 1461–1463.

[79] Knowles, L. Genealogical portraits of speciation in montane grasshoppers (genus *Melanoplus*) from the sky islands of the rocky mountains. *Proc. R. Soc. Lond. B 268* (2001), 319–324.

[80] Krogstad, F. Coding the Metropolis-Hastings algorithm. available at http ://students.washington.edu/ fkrogsta/bayes/stat538.pdf, 1999.

[81] Kuhner, M., Yamato, J., and Felsenstein, J. Maximum likelihood estimation of population growth rate based on the coalescent. *Genetics 149* (1998), 429–434.

[82] Lande, R. Genetics and demography in biological conservation. *Science 241* (1988), 1455–1460.

[83] Le Page, S. L., Livermore, R. A., Cooper, D. W., and Taylor, A. C. Genetic analysis of a documented population bottleneck : introduced Bennett's Wallabies (*Macrocarpus rufogriseus rufogriseus*) in New Zealand. *Molecular Ecology 9* (2000), 753–763.

[84] Loader, C. R. Local likelihood density estimation. *The Annals of Statistics 24* (1996), 1602–1618.

[85] Luikart, G., Allendorf, F., Cornuet, J.-M., and Sherwin, W. Distortion of allele frequency distributions provides a test for recent population bottlenecks. *Journal of Heredity 89(3)* (1998), 238–247.

[86] Luikart, G., and Cornuet, J.-M. Empirical evaluation of a test for identifying recently bottlenecked populations from allele frequency data. *Conservation Biology 12(1)* (1998), 228–237.

[87] Luikart, G., and England, P. R. Statistical analysis of microsatellite DNA data. *Trends in Ecology and Evolution 14* (1999), 253–256.

[88] Maruyama, T., and Fuerst, P. F. Population bottlenecks and non-equilibrium models in population genetics. I. allele numbers when populations evolve from zero variability. *Genetics 108* (1984), 745–763.

[89] Maruyama, T., and Fuerst, P. F. Population bottlenecks and non-equilibrium models in population genetics. II. *Genetics 111* (1985), 675–689.

[90] Maudet, C., Miller, C., Bassano, B., Breitenmoser-Würsten, C., Gauthier, D., Obexer-Ruff, G., Michallet, J., Taberlet, P., and Luikart, G. Microsatellite DNA and recent statistical methods in wildlife conservation management : applications in alpine ibex (*Capra ibex (ibex)*). *Molecular Ecology 11* (2002), 421–436.

[91] Mayr, E. Change of genetic environment and evolution. In *Evolution as a process*, J. S. Huxley, A. C. Hardy, and E. B. Ford, Eds. Allen & Uwin, London, 1954, pp. 156–180.

[92] Mayr, E. Review of *modes of speciation* by M.J.D. White. *Systematic Zoology 27* (1978), 478–482.

[93] Mayr, E. Processes of speciation in animals. In *Mecanisms of speciation*, C. Barigozzi, Ed. Liss, New-York, 1982, pp. 1–19.

[94] Milligan, B., Leebens-Mack, J., and Strand, A. Conservation genetics : beyond the maintenance of marker diversity. *Molecular Ecology 3* (1994), 423–435.

[95] Nauta, M. J., and Weissing, F. J. Constraints on allele size at microsatellite loci : implications for genetic differenciation. *Genetics 143* (1996), 1021–1032.

[96] Nei, M., Maruyama, T., and Chakraborty, R. The bottleneck effect and genetic variability in populations. *Evolution 29* (1975), 1–10.

[97] Nielsen, R. A likelihood approach to populations samples of microsatellite alleles. *Genetics 146* (1997), 711–716.

[98] Nielsen, R., and Wakeley, J. Distinguishing migration from isolation : a Markov Chain Monte Carlo approach. *Genetics 158* (2001), 885–896.

[99] Nordborg, M. Coalescent theory. In *Handbook of Statistical Genetics*, C. C. D.J. Balding, M. Bishop, Ed. Wiley, Chichester, 2001, pp. 179–212.

[100] Otte, D., and Endler, J. A., Eds. *Speciation and its consequences.* Sinauer, Sunderland, MA, 1989.

[101] Pascal, M. Structure et dynamique de la population de chats harets de l'archipel des kerguelen. *Mammalia 44(2)* (1980), 161–182.

[102] Pascal, M. Le chat haret (*Felis catus* l. 1758) aux îles kerguelen. *Arvicola 1* (1984), 31–35.

[103] Pascal, M., Siorat, F., Cosson, J.-F., and Burin des Roziers, H. Eradication de populations insulaires de surmulots : archipel des Sept-Iles, archipel de Cancale, Bretagne, France. *Vie et Milieu, Life and Environment 46* (1996), 267–283.

[104] Peter, D. The coalescent and microsatellite variability. In *Microsatellites : evolution and applications*, D. B. Goldstein and C. Schlötterer, Eds. Oxford University Press, Oxford, 1999, pp. 117–128.

[105] Piry, S., Luikart, G., and Cornuet, J.-M. BOTTLENECK : a computer program for detecting recent reductions in the effective population size using allele frequency data. *Journal of Heredity 90* (1999), 502–503.

[106] Press, W. H., Teukolsky, W. T., and Vetterling, B. P. *Numerical Recipes in C :*, second edition ed. Cambridge University Press, 1992.

[107] Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., and Feldman, M. W. Population growth of human Y chromosomes : a study of Y chromosome microsatellites. *Molecular Biology and Evolution 16* (1999), 1791–1798.

[108] Pritchard, J. K., Stephens, M., and Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics 155* (2000), 945–959.

[109] Rabinow, P. *Making PCR, a story of biotechnology.* University of Chicago Press, 1996.

[110] Raftery, A. E., and Lewis, S. M. Implementing MCMC. In *Markov Chain Monte Carlo in Practice*, S. R. W. R. Gilks and D. J. Spiegelhalter, Eds. Chapman & Hall, London, 1996, pp. 115–130.

[111] Raup, D. M. *Extinction. Bad genes or bad luck ?* W. W. Norton, 1991.

[112] Raymond, M., and Rousset, F. Genepop (ver. 1.2) : a population genetics software for exact test and ecumenicism. *Journal of Heredity 86* (1995), 248–249.

[113] Reich, D. E., Feldman, M. W., and Goldstein, D. B. Statistical properties of two tests that use multilocus data sets to detect population expansions. *Molecular Biology and Evolution 16(4)* (1999), 453–466.

[114] Renwick, A., Davison, L., Spratt, H., King, P., and Kimmel, M. DNA dinucleotide evolution in Humans : fitting theory to facts. *Genetics 159* (2001), 737–747.

[115] Ripley, B. D. *Stochastic simulation.* Wiley, New York, 1987.

[116] Robert, C. *Méthodes de Monte Carlo par Chaînes de Markov.* Economica, 1996.

[117] Rogers, A. R. Genetic evidence for a pleistocene population expansion. *Evolution 49(4)* (1995), 608–615.

[118] Rogers, A. R. Order emerging from chaos in human evolutionary genetics. *Proceedings of the National Academy of Sciences of the United States of America 98(3)* (2001), 779–780.

[119] Rogers, A. R., and Jorde, L. B. Genetic evidence on modern human origin. available on the web.

[120] Rubinstein, D., et al. Microsatellite evolution-evidence for directionality and variation in rate between species. *Nature Genetics 10* (1995), 337–343.

[121] Rundle, H. D., and Whitlock, M. C. Single founder-flush events and the evolution of reproductive isolation. *Evolution 52(6)* (1998), 1850–1855.

[122] Saccheri, I. J., Wilson, I. J., Nichols, R. A., Bruford, M. W., and Brakefiels, P. M. Inbreeding of bottlenecked butterfly populations : estimation using the likelihood of changes in marker allele frequencies. *Genetics 151* (1999), 1053–1063.

[123] Sakai, A. R., et al. The population biology of invasive species. *Annual Review of Ecology and Systematics 32* (2001), 305–332.

[124] Sakai, R. K., Scharf, S., Faloona, F., Mullis, K. B., Horn, G. T., Erlich, H. A., and N, A. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science 14(12)* (1996), 478–483.

[125] Saunders, I. W., Tavaré, S., and Watterson, G. A. On the genealogy of nested subsamples from haploid populations. *Advanced and Applied Probability 16* (1984), 471–491.

[126] Say, L., Gaillard, J., and Pontier, D. Spatio-temporal variation in cat population size in a sub-antarctic environment. *Polar Biology 25* (2002), 90–95.

[127] Schlötterer, C. Evolutionary dynamics of microsatellite DNA. *Chromosoma 109* (2000), 365–371.

[128] Schlötterer, C., Ritter, R., Harr, B., and Brem, G. High mutation rate of a long microsatellite allele in *Drosophila melanogaster* provides evidence of allele specific mutation rates. *Molecular Biology and Evolution 15* (1998), 1269–1274.

[129] Schlötterer, C., and Wiehe, T. Microsatellites, a neutral marker to infer selective sweeps. In *Microsatellites : evolution and applications*, D. B. Goldstein and C. Schlötterer, Eds. Oxford University Press, Oxford, 1999, pp. 238–248.

[130] Schug, M. D., Mackay, T. F., and Aquadro, C. F. Low mutation rates of microsatellite loci in *Drosophila melanogaster*. *Nature Genetics 15* (1997), 99–102.

[131] Shriver, M. D., Jin, L., Chakraborty, R., and Boerwinckle, E. VNTR allele frequency distributions under the stepwise mutation model : A computer simulation approach. *Genetics 134* (1993), 983–993.

[132] Sibert, A. Héritabilité non-génétique de la fécondité : effet sur le polymorphisme. Thèse de Doctorat du Museum National d'Histoire Naturelle, 2002.

[133] Slatkin, M. In defense of founder-flush theories of speciation. *American Naturalist 147(4)* (1996), 493–505.

[134] Software, N. R., Ed. *Numerical recipes in C : the art of scientific computing.* Cambridge University Press, 1992, ch. Chapter 7 : random numbers.

[135] Sokal, A. D. Methods in Statistical Mechanics, Foundations and New Algorithms. Lectures at the Carg'ese Summer School on *Functional Integration : Basics and Applications*, 1996.

[136] Stephens, M. Problems with computational methods in population genetics. Bulletin of the 52nd session of the International Statistical Institute.

[137] Stephens, M. Inference under the coalescent. In *Handbook of Statistical Genetics*, C. C. D.J. Balding, M. Bishop, Ed. Wiley, Chichester, 2001.

[138] Stephens, M., and Donnelly, P. Inference in molecular population genetics. *Journal of the Royal Statistical Society. Series B 62(4)* (2000), 605–655.

[139] Storz, J. F., and Beaumont, M. Testing for genetic evidence of population expansion and contraction in two indian fruit bat species (*Cynopterus sphinx* and *C. brachyotis*) : analysis of microsatellite variation using a hierarchical bayesian model. *Evolution 56(1)* (2002), 154–166.

[140] Tajima, F. Evolutionary relashionships of DNA sequences in finite populations. *Genetics 105* (1983), 437–460.

[141] Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics 123* (1989), 585–595.

[142] Tanner, M. A. *Tools for statistical inference : methods for the exploration of posterior distributions and likelihood functions*, third edition ed. Springer Series in Statistics. Springer, 1996.

[143] Tautz, D., and Renz, M. Simple sequences are ubiquitous repetitive components of Eucaryote genomes. *Nucl.Ac.Res. 12* (1984), 4127–4138.

[144] Tavaré, S. Line of descent and genealogical processes, and their applications in population genetics models. *Theoretical Population Biology 26* (1984), 119–164.

[145] Tavaré, S., Balding, D. J., Griffiths, R. C., and Donnelly, P. Inferring coalescence times from DNA sequence data. *Genetics 145* (1997), 505–518.

[146] Taylor, A., Sherwin, W., and Wayne, R. Genetic variation of microsatellite loci in a bottlenecked species : the northern hairy-nosed wombat *Lasiorhinus krefftii*. *Molecular Ecology 3* (1994), 277–290.

[147] Templeton, A. The theory of speciation *via* the founder principle. *Genetics 94* (1980), 1011–1038.

[148] Thuillet, A.-C., Bru, D., David, J.-L., Roumet, P., Santoni, S., Sourdille, P., and Bataillon, T. Direct estimation of mutation rate for 10 microsatellite loci in durum wheat, *Triticum turgidum* (L.) Thell. spp durum desf. *Molecular Biology and Evolution 19(1)* (2002), 122–125.

[149] Udupa, S. M., and Baum, M. High mutation rate and mutational bias at $(TTA)_n$ microsatellite loci in chickpea (*Cicer arietinum* l.). *Molecular Genetics and Genomics 265* (2001), 1097–1103.

[150] Valdes, A. M., Slatkin, M., and Freimer, N. Allele frequencies at microsatellite loci : the stepwise mutation model revisited. *Genetics 133* (1993), 737–749.

[151] Watterson, G. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* (1975), 119–164.

[152] Weiss, G., and von Haeseler, A. Inference of population history using a likelihood approach. *Genetics 149* (1998), 1539–1546.

[153] Williamson, M. *Biological invasions*. Chapman and Hall, London, 1997.

[154] Wilson, I. J., and Balding, D. J. Genealogical inference from microsatellite data. *Genetics 150* (1998), 499–510.

[155] Wilson, I. J., Weale, M. E., and Balding, D. J. Inferences from DNA data : population histories, evolutionary processes and forensic match probabilities.

[156] Wright, S. Evolution in mendelian populations. *Genetics 16* (1931), 97–159.

[157] Xu, X., M, P., and Z, F. The direction of microsatellite mutations is dependent upon allele length. *Nature Genetics 24* (2000), 396–399.

[158] Young, D. *The Discovery of Evolution.* Cambridge University Press, Cambridge, 1992.

[159] Zhivotovsky, L. A., Feldman, M. W., and Grishechkin, S. A. Biased mutations and microsatellite variation. *Molecular Biology and Evolution 14(9)* (1997), 926–933.

Table des figures

Liste des tableaux

Annexes

Annexe A :

Claire CALMET and Mark BEAUMONT (in prep.). Inferring population history from microsatellite data : interactions between a founder-flush demographic model and two-phase mutational model.

Annexe B :

Claire CALMET, Michel PASCAL and Sarah SAMADI (2001). Is it worth eradicating the invasive pest *Rattus norvegicus* from Molène archipelago ? Genetic structure as a decision making tool. Biodiversity and Conservation *10* : 911-928.

Annexe C :

Dominique PONTIER, Claire CALMET, Ludovic SAY and François BONHOMME. Documented introduction of the feral cat (*Felis catus* L.) on a Sub-Antarctic island : genetic signatures of the founder-flush and methodological implications.

Annexe A :

Claire CALMET and Mark BEAUMONT (in prep.). Inferring population history from microsatellite data : interactions between a founder-flush demographic model and two-phase mutational model.

Sera soumis à Genetics. La présente version n'engage que le premier auteur.

Annexe B :

Claire CALMET, Michel PASCAL and Sarah SAMADI (2001). Is it worth eradicating the invasive pest *Rattus norvegicus* from Molène archipelago ? Genetic structure as a decision making tool. Biodiversity and Conservation *10* : 911-928.

Annexe C :

Dominique PONTIER, Claire CALMET, Ludovic SAY and François BONHOMME. Documented introduction of the feral cat (*Felis catus* L.) on a Sub-Antarctic island : genetic signatures of the founder-flush and methodological implications.

En révision pour Molecular Ecology.