Inférences démographiques à partir de données génomiques sous des modèles spatialisés.

Thimothée Virgoulay

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

En Biologie des populations et écologie

École doctorale : Biodiversité, Agriculture, Alimentation, Environnement, Terre, Eau - Montpellier

Unité de recherche : UMR Institut des Sciences de l'Evolution de Montpellier (ISEM)
UMR Centre de Biologie pour la Gestion des Populations (CBGP)

INFERENCES DEMOGRAPHIQUES A PARTIR DE DONNEES GENOMIQUES SOUS DES MODELES SPATIALISES

Présentée par Thimothée VIRGOULAY
Le 23 mai 2022

Sous la direction de François ROUSSET

Devant le jury composé de

Guillaume ACHAZ, Professeur d'université, Université de Paris	Rapporteur
Frédéric AUSTERLITZ, Directeur de recherche, CNRS	Examinateur
Raphaël LEBLOIS, Chargé de recherche, INRAE	Co-encadrant
Jean-Michel MARIN, Professeur d'université, Université de Montpellier	Co-encadrant
Joëlle RONFORT, Directrice de recherche, INRAE	**Presidente de jury**
François ROUSSET, Directeur de recherche, CNRS	Directeur de thèse
Bertrand SERVIN, Directeur de recherche, INRAE	Rapporteur

UNIVERSITÉ DE MONTPELLIER

Résumé

Cette thèse a pour but de développer un ensemble de concepts et d'outils inférentiels adaptés à l'analyse du polymorphisme génétique neutre dans une classe bien spécifique de modèles stochastiques de génétique des populations : les modèles démographiques spatialisés de populations où la dispersion des individus entre générations est limitée dans l'espace (modèles d'isolement par la distance). Cette analyse repose sur la combinaison de modèles stochastiques de l'évolution des populations (la coalescence en génération par génération) et d'une méthode d'inférence statistique par simulation ("the summary-likelihood method"). Dans un premier temps, les divers outils furent mis au point et testés indépendamment avant d'être intégrés ensemble afin de créer un canevas d'inférence reliant des outils adaptés à nos modèles. Dans un second temps, nous avons utilisé ce canevas afin d'étudier les performances et les bénéfices de ces nouvelles méthodes d'inférences par simulation dans ce contexte d'isolement par la distance, notamment l'influence de la prise en compte des patrons de déséquilibre de liaison sur l'inférence des paramètres de densité et de dispersion.

Abstract

This thesis aims at presenting a set of concepts and inferential tools adapted to the analysis of neutral genetic polymorphism in a specific class of stochastic population genetics models : spatialized demographic models of populations where the dispersal of individuals between generations is limited in space (isolation by distance models). Such analysis is based on the combination of stochastic models of population evolution (generation by generation coalescence) and of a statistical inference method by simulation ("the summary-likelihood method"). In a first step various tools were developed and tested independently before being integrated together to create an inferential framework adapted to our models. In a second step, we used this framework to study the performance and benefits of these new inference methods and the influence of considering linkage disequilibrium patterns on parameter inference under distance isolation models.

Remerciements

"[...] the "pageant" of evolution as a staggeringly improbable series of events, sensible enough in retrospect and subject to rigorous explanation, but utterly unpredictable and quite unrepeatable."

Wonderful Life, S.J. Gould

Pour leur temps, leur pédagogie, leur infinie patience et leur capacité surhumaine à me porter à bout de bras, je tiens à remercier mes encadrants Raphaël Leblois, François Rousset et même si je cite : "nous avons moins interagi " Jean-Michel Marin.

Pour tout le reste il y a tous les autres, les collègues sans qui rien n'aurait été possible. Je remercie avant tout la Patate qui se reconnaîtra (meilleure co-bureau même le roi la demande), ma copilote de thèse (je lui dédis mon plus beau chant du panda), une grenobloise de trop bon conseil (et de trop bon esprit) mais aussi tous les précaires du CBGP qui m'ont permis de ne pas oublier d'apprécier cette vie.

Les Poneys sont sans conteste une part importante de l'expérience et je me demande bien dans quel canal de discorde j'aurais fini par errer sans eux. Merci d'avoir toujours été là, de m'avoir supporté et d'être un élément structurant de mon lore. Vous êtes OP (niv+1 pour avoir passé vos soirées à corriger mes fautes d'orthographe).

À mon Père et à ma Soeur qui m'ont épaulé dans les moments difficiles et dont l'amour est un baume à mes désespérances.

À ma Mère qui dit-on aurait été très fière de moi.

À Cerise qui m'a accompagné jusque-là (je n'y serais jamais arrivé sans toi).

À toi qui même si tu l'ignores m'as permis de tenir jusqu'au bout.

Je vous aime.

Table des matières

Table des figures

Liste des tableaux

1 Introduction

1.1 Estimation des paramètres démographiques

La connaissance des paramètres démographiques des populations naturelles (telles que les tailles et densités de populations, taux et distances de dispersion, la structure d'âge, etc.), ainsi que de leurs changements passés (i.e. l'histoire démographique des populations), est cruciale dans de nombreux domaines de l'écologie et de la biologie évolutive car ce sont des paramètres clés dans le fonctionnement passé, actuel et futur des populations. En biologie de la conservation par exemple, les tailles de populations associées aux taux de dispersion entre ces populations peuvent aider à caractériser l'état démographique actuel des populations (répartition des effectifs de l'espèce et connectivité entre populations ; Caughley, 1994 ; Hanski, 1998 ; Heller and Zavaleta, 2009). De plus, leurs variations en fonction de l'hétérogénéité du paysage peuvent aider à caractériser les milieux plus ou moins favorables à l'installation ou aux déplacement des individus. Enfin, les changements récents de certains de ces paramètres peuvent nous informer sur les risques

d'extinction des populations en lien avec des baisses d'effectifs ou de connectivité (Norris, 2004). Un autre exemple évident est en biologie des invasions, où la caractérisation de la distribution des distances de dispersion peut apporter de l'information sur la vitesse d'expansion d'une population envahissante (Caswell et al., 2003 ; Hastings et al., 2005).

En plus de ces exemples relativement triviaux concernant le fonctionnement purement démographique passé et actuel des populations, la caractérisation des paramètres démographiques des populations a une portée beaucoup plus large puisqu'ils contrôlent, en interaction avec les paramètres sélectifs, l'évolution de divers caractères et traits d'histoire de vie des individus. Dans le contexte de l'adaptation locale des organismes à un environnement hétérogène par exemple, les paramètres démographiques peuvent renseigner sur le potentiel adaptatif de chaque population à (des modifications de) son environnement. En effet, les tailles de populations et les taux de dispersion déterminent (1) la dérive locale, et donc le niveau (et son évolution dans le temps) de la diversité génétique locale sur laquelle la sélection va s'exercer ; et (2) le niveau de flux de gènes, et donc la répartition des allèles délétères ou avantageux entre populations dans les différents environnements(Whitlock, 2002 ; Lopez et al., 2009). En écologie comportementale, les taux d'extinction locale et de dispersion contrôlent le degré de compétition locale entre individus et les processus de sélection de parentèle, qui peuvent influencer l'évolution de nombreux traits d'histoire de vie (e.g. Lehmann et al., 2006). En cas de changement environnemental, la vitesse d'expansion d'un allèle adaptatif va fortement dépendre de la distribution des distances de dispersion et des tailles/densités locales des populations. Enfin, les clines de fréquences allélique donnent des exemples classiques illustrant comment l'étendue de l'adaptation locale ou l'hybridation entre taxons peut dépendre de ces taux de dispersion (Endler, 1977). En effet, la vitesse d'atteinte de l'équilibre du cline et sa forme dépendent du rapport entre les pressions de sélection et les caractéristiques de la distribution de dispersion. Un bon exemple de cette dynamiques de cline peut être observé chez le phalène du bouleau (*B. betularia*) dans la région du Sud-Ouest de l'Angleterre (Saccheri et al., 2008). La connaissance des paramètres démographiques apporte donc non seulement de l'information sur le fonctionnement démographique des populations mais aussi sur le potentiel de réponse démographique et adaptative de ces populations à des changements environnementaux.

Dans le contexte actuel des changements globaux (climat, utilisation des sols, fragmentation des habitats naturels, pollutions, invasions biologiques), il est clairement nécessaire de mettre en place des mesures ciblées de protection ou de gestion des populations menacées pour préserver de la biodiversité face aux pressions anthropiques directes ou indirectes. De même, en agriculture, il parait primordial de développer des approches plus écologiques pour la gestion des organismes d'intérêt agronomique (les ravageurs et auxiliaires de cultures, leurs vecteurs et antagonistes) afin de réduire l'utilisation de produits phytosanitaires. On peut citer l'exemple de la lutte biologique mais aussi la manipulation des équilibres écologiques entre populations de différentes espèces afin de défavoriser les ravageurs et leurs vecteurs tout en favorisant leurs antagonistes et les auxiliaires (Lewis et al., 1997). Or, de telles approches nécessitent de mieux comprendre les dynamiques populationnelles à différentes échelles (de la population à l'espèce, de la parcelle au bassin de production, ...) et demandent donc une connaissance fine (1) du statut démo-génétique de ces populations (diversité génétique, consanguinité, potentiel adaptatif, taille efficace, structuration génétique) et (2) de leur fonctionnement démographique local, actuel et sur un passé récent, en lien avec leur environnement et ses changements récents.

Toutes ces raisons, des plus académiques aux plus appliquées, justifient une attention particulière pour la caractérisation précise des paramètres démographiques des populations et de leurs variations passées, notamment à de petites échelles spatio-temporelles, et une attention particulière pour la quantification des processus de dispersion localisée. Ces paramètres démographiques peuvent théoriquement être estimés par des suivis démographiques des populations en utilisant des techniques de capture-marquage-recapture (CMR) ou de suivis d'abondance par piégeage/comptage (e.g. Peron et al., 2010). Cependant (1) la CMR, en plus d'être extrêmement coûteuse et compliquée à mettre en place, n'est guère applicable à certains organismes difficiles à capturer et/ou marquer (caractérisés par, e.g., de très faibles ou très fortes densités, une très petite taille, une forte mortalité ; Osborne et al., 2002) ; (2) les suivis d'abondance, plus faciles à mettre en place, ne donnent aucune indication sur la dispersion. Les différents processus démographiques contribuant à façonner la répartition de la diversité génétique au sein des populations, il est possible d'inférer les processus démographiques à partir de l'analyse de la diversité génétique neutre (Goldstein and Harvey, 1999). Une alternative aux études

démographiques consiste donc à utiliser la génétique des populations pour avoir un accès "indirect" (Slatkin, 1985) aux paramètres démographiques d'intérêt à partir de données génétiques.

1.2 Estimation à partir de données génétiques

Depuis l'apparition des premières données de génotypage (allozymes dans les années 60), la génétique des populations a développé des méthodes d'estimation de paramètres démographiques à partir de données génétiques. Longtemps limitées à l'analyse de petits jeux de données peu informatifs sous des modèles démo-génétiques très simples, ces développements ont explosé à partir des années 1990-2000 du fait de l'augmentation de la puissance de calcul informatique et de l'informativité des données génétiques grâce aux nouvelles techniques de séquençage (Luikart et al., 2003 ; Marchi et al., 2021). Si certaines de ces méthodes ont pu donner des résultats raisonnables et parfois même impressionnants comme l'inférence détaillée des variations passées de la taille d'une population à partir d'un seul génome diploïde (Li and Durbin, 2011), elles ont encore de nombreuses limitations.

En particulier, les modèles de génétique des populations sont souvent fondés sur de nombreuses approximations, parmi lesquelles les approximations des processus en (sous-) populations panmictiques de grande taille ($N \to \infty$) avec des petits taux d'évènements (e.g., taux de mutation $\mu \to 0$, taux de migration $m \to 0$). Ces approximations (dites approximations de diffusion ou du n-coalescent) sont notamment caractérisées par le fait que le paramètres de ces modèles sont les produits de tailles de populations avec les paramètres de taux d'évènements (e.g. $2N\mu \to \theta$, $2Nm \to \gamma$). Même si ces approximations sont pertinentes pour un grand nombre d'études et de situations biologiques, notamment lorsqu'elles sont utilisées pour étudier des processus évolutifs à grandes échelles spatiales et temporelles, elles ne sont potentiellement pas valides pour l'analyse des processus évolutifs à petites échelles. Et surtout, elles excluent de pouvoir estimer séparément les paramètres N, μ et m qui apparaissent dans ces produits. Un objectif majeur de ce travail est de se donner les moyens d'estimer séparément ceux-ci, et d'évaluer la qualité des estimations ainsi obtenues.

De plus, la majorité des modèles utilisés ne prend pas en compte la structuration des populations (e.g. les méthodes retraçant l'histoire des tailles de populations, comme celle de Li and Durbin, 2011) alors qu'il est maintenant bien établi que négliger la structuration des populations peut fortement biaiser les inférences des variations passées des tailles de populations (Leblois et al., 2006, 2014 ; Mazet et al., 2016 ; Chikhi et al., 2010, 2018). Et celles qui en tiennent compte considèrent le plus souvent des modèles ne décrivant pas de manière adéquate les processus de dispersion localisée dans l'espace, selon lesquels la probabilité de dispersion décroît avec la distance géographique, processus que l'on retrouve chez une majorité d'espèces (Endler, 1977). Enfin, à petites échelles spatiales et pour de nombreuses espèces, la définition (et la notion même) de (sous-)populations panmictiques est problématique car les individus ne sont pas regroupés en unités panmictiques mais plutôt répartis de manière continue dans l'habitat. Une catégorie de modèles démogénétiques, dit d'"isolement par la distance" (Wright, 1943), que nous détaillerons par la suite, considère une dispersion potentiellement localisée dans l'espace, et une répartition des individus pouvant être en (sous-)populations panmictiques ou complètement homogène sur un habitat continu. Ces modèles d'"isolement par la distance", bien qu'ils fassent aussi de nombreuses hypothèses que nous discuterons par la suite, semblent les plus pertinents pour estimer, à partir de données génétiques, les paramètres démographiques des populations, dont la dispersion, à petites échelles spatio-temporelles.

Enfin, l'ajustement des modèles est souvent difficile. En particulier, la vraisemblance des données n'est généralement pas calculable analytiquement, sauf dans quelques cas particuliers d'intérêt limité (tels que le modèle en allèles infini de Ewens, 1972 ; ou le modèle en îles infini), fort éloignés de scénarios d'intérêt tels que des changements récents de tailles de populations. Cette difficulté se rencontre déjà pour l'analyse d'informations génétiques locus par locus, mais *a fortiori* pour des données contenant une information sur les associations statistiques en loci (déséquilibres de liaison). Or, des travaux comme ceux de Li and Durbin ont montré que les patrons de déséquilibre de liaison contiennent potentiellement de l'information sur les variations ancestrales des tailles de populations. D'autres travaux comme ceux de Al-Asadi et al. (2019) ont montré que les patrons de déséquilibre de liaison contiennent potentiellement de l'information sur la dispersion.

La prise en compte progressive de ces difficultés a mené au développement de différentes méthodes, que l'on peut introduire dans un ordre historique. Les méthodes fondées sur le calcul du F_{ST} de Wright où de quantités analogues sont très classiques. Elles peuvent être formalisées en termes de calculs sur des paires de (copies de) gènes et perdent donc l'information potentiellement contenue dans les associations statistiques entre triplets, etc. information qui peut être retenue par les méthodes de vraisemblance. Un ensemble de méthodes alternatives vise à estimer la vraisemblance (1) de l'échantillon total par différents algorithmes de simulation basés sur la théorie de coalescence (e.g. Hey and Nielsen, 2007 ; Rousset et al., 2018) ; ou (2) du spectre de fréquences alléliques (AFS, pour *allel frequency spectrum*), un ensemble de statistiques résumant "efficacement" l'échantillon génétique ; e.g. Gutenkunst et al., 2009 ; Excoffier et al., 2013). D'autres méthodes ont été développées dans le cadre de modèles Bayésiens hiérarchiques dans lesquels les distributions prior des paramètres sont issues d'approximations diverses spécifiques des modèles démo-génétiques sous-jacents (e.g. dans le cadre de modèles spatialisés Petkova et al., 2016). L'efficacité de toutes ces méthodes dépend du développement (souvent laborieux) d'algorithmes adaptés aux processus populationnels considérés. Enfin, des méthodes d'inférence par simulation des processus, très développées dans un schéma d'inférence Bayésienne depuis le début des années 2000, consistent à simuler des réalisations des processus pour différentes valeurs de paramètres, et à retenir les paramètres (typiquement sous forme d'une distribution à *posteriori*) qui paraissent le mieux produire des simulations qui se rapprochent des données observées (méthodes dites ABC pour *Approximate Bayesian Computation* ; Beaumont et al., 2002 ; Beaumont, 2010). Plus récemment, une méthode alternative d'inférence par simulation, fondée sur l'utilisation de la densité jointe des statistiques et des paramètres pour inférer la surface de vraisemblance des paramètres sachant les statistiques résumantes observées sur l'échantillon, a été développée (méthode dite "Summary Likelihood" ; Rousset et al., 2017). Toutes ces méthodes d'inférence par simulation permettent l'estimation de paramètres sous des scénarios moins contraints que les méthodes précédentes, à partir du moment où ceux-ci peuvent être simulés rapidement. Ces méthodes d'inférence par simulation se fondent sur la réduction des données en un certain nombre de descripteurs (les *statistiques résumantes*). Une partie de l'information statistique peut être perdue lors du calcul de ces statistiques, de façon analogue à la perte d'information potentielle quand on calcule des F_{ST} au lieu d'effectuer une analyse par maximum de vraisemblance. De ce fait, une étape importante dans d'utilisation de ces méthodes d'inférence par simulation

est de trouver un ensemble de statistiques résumantes minimisant la perte d'information présente dans l'échantillon complet sur les paramètres d'intérêt des modèles considérés. L'objectif global de cette thèse est de tester l'intérêt de ces méthodes d'inférence par simulation dans le cadre de l'inférence de paramètres démographiques tels que les tailles et densités de population ainsi que les caractéristiques de dispersion, à de petites échelles spatiales et temporelles ; notamment, d'évaluer quels paramètres, ou combinaisons de paramètres, peuvent être estimés avec ces méthodes et avec quelle précision. Un des objectifs plus spécifiques de cette thèse est d'utiliser l'information contenue dans les déséquilibres de liaison, et en particulier d'explorer la possibilité que cette information permette d'estimer séparément les paramètres de taille/densité de population et les autres paramètres d'intérêt d'un modèle spatialisé, et non simplement le produit de telles quantités. Ceci suppose de pouvoir simuler des données génétiques (1) avec déséquilibre de liaison (i.e. prendre en compte la recombinaison pour simuler des locus liés), et (2) sans faire dans le programme de simulation d'approximation supposant de grandes tailles de population et des faibles probabilités d'évènements par génération.

Un premier objectif concret de cette thèse a donc été de développer un simulateur de données multilocus (incluant le cas de données génomiques sur des grands fragments de chromosomes) sous des modèles *exacts* de coalescence en présence de dispersion localisée. [1] En effet les simulateurs existants ne présentaient pas toutes ces propriétés. Par exemple le simulateur précédemment utilisé pour l'analyse de la dispersion localisée (`IBDsim`, Leblois et al., 2009) simulait indépendamment chaque locus et ne permettait donc pas de prendre en compte la liaison entre eux. D'autres simulateurs appliquent souvent les approximations en population large, sauf exception ne prenant pas en compte la dispersion localisée.

Une fois ce simulateur développé, le deuxième objectif concret de cette thèse est de l'utiliser pour l'estimation de paramètres démographiques à partir de données génétiques incluant l'information de déséquilibre de liaison, et donc de le coupler avec des méthodes d'inférence par simulation.

1. Dans ce travail, on distingue le "coalescent" (le processus aléatoire décrit par Kingman, fondé sur des approximations en population large) et les processus de coalescence (le fait que des lignées génétiques se rejoignent en des lignées ancestrales communes). On parlera de coalescence exacte quand on ne fera pas d'approximation en population large. Ces modèles restent évidemment inexacts par potentiellement toutes leurs autres hypothèses.

1.3 Inférences par simulation

La première de ces méthodes (ABC-RF ; Pudlo et al., 2016 ; Raynal et al., 2018) est une méthode ABC reposant sur l'utilisation de la méthode d'apprentissage automatique dite "forêts aléatoires" (*random forests*, dénomination imagée basée sur l'utilisation d'un ensemble d'arbres de décision produits par des simulations avec une composante aléatoire ; Breiman, 2001). Cette méthode d'apprentissage permet de construire de façon automatique et simple d'utilisation, ici sur la base d'un tableau de données d'apprentissage dont chaque ligne i contient la valeur θ_i d'un paramètre du processus et un ensemble de statistiques résumantes $\mathcal{S}_i$ calculées sur une réalisation du processus sachant θ_i et d'autres paramètres simulés, des estimateurs non-paramétriques d'une fonction de régression donnant l'espérance du paramètre θ sachant les valeurs des statistiques résumantes. Appliquée à un jeu de données, l'estimateur déduit des données d'apprentissage estime donc l'espérance de la distribution a *posteriori* du paramètre sachant les statistiques résumantes. La méthode ABC-RF produit plus généralement une estimation de la distribution marginale a *posteriori*, pour chaque paramètre séparément.

La deuxième de ces méthodes, dite *summary likelihood*, est assez proche d'ABC-RF au premier abord : elle peut partir des mêmes simulations que les méthodes ABC. Elle peut utiliser diverses méthodes pour estimer de même une fonction de régression non paramétrique, et les forêts aléatoires y sont la méthode utilisée par défaut dans ce but. Le nombre minimal de statistiques nécessaires pour l'estimation de p paramètres étant p statistiques, une première étape de la méthode consiste à réduire un grand nombre de statistiques résumantes calculées sur un jeu de données en p statistiques résumantes *projetées* conçues pour conserver le maximum d'information pour chaque paramètre. Idéalement, pour conserver un maximum d'informations, ces statistiques résumantes devraient être des fonctions monotones de l'estimateur par maximum de vraisemblance, mais on s'intéresse aux cas ou de tels estimateurs ne sont pas simplement calculables. La deuxième étape de la méthode de *summary likelihood* diverge de l'ABC en général : plutôt que des distributions a *posteriori*, elle vise à inférer une surface de vraisemblance des p paramètres, sachant les p statistiques projetées observées sur l'échantillon. Des techniques classiques pour l'inférence statistique peuvent être appliquées une fois cette fonction de vraisem-

blance disponible. La vraisemblance est conçue conjointement pour tous les paramètres, et en cela elle se distingue plus spécifiquement des distributions marginales postérieures produites par l'ABC-RF.

La méthode de *summary likelihood* a été développée dans l'objectif d'inférences ayant des propriétés "fréquentistes", à savoir le contrôle du taux d'erreur des conclusions pour toute valeur de paramètre (suivant l'exemple canonique des intervalles de confiance ; Neyman, 1935). Le passage par la construction d'une surface de vraisemblance vise à exploiter les relativement bonnes propriétés fréquentistes des méthodes de vraisemblance (éventuellement corrigées par des méthodes de bootstrap). Du point de vue de ces propriétés, la différence entre intervalles de confiance produits par *summary likelihood* et intervalles de crédibilité déduits de distributions a *posteriori* estimées par une méthode ABC a été illustrée par Rousset et al. (2017), en utilisant une première implémentation de la méthode de *summary likelihood*, qui construisait une surface de vraisemblance de façon moins efficace que celle utilisée dans ce travail. La méthode, telle qu'elle sera utilisée ici, est implémentée dans la librairie R `Infusion` (Rousset, 2022) mais non décrite dans une publication. Ce travail de thèse constitue sa première application en dehors de quelques exemples-jouet.

Pour contrôler l'ensemble du processus d'inférence de façon efficace, il est apparu nécessaire de programmer également une librairie de calcul de statistiques résumantes, qui s'est donc rajoutée comme étape intermédiaire de ce travail. Au-delà d'un ensemble de statistiques classiques de diversité et de structuration des populations, cette librairie (`GSumStat`) vise à calculer de façon efficace des descripteurs de la structure génétique à très fine échelle et des descripteurs des déséquilibres de liaison en fonction simultanément de la distance géographique entre individus et de la distance physique sur un chromosome. Enfin, Une librairie R dédiée, `gspace2infr`, a été implémentée pour automatiser le couplage des différentes composantes de l'inférence, automatiser les calculs décrits dans les chapitres suivants, et faciliter l'évaluation de cette nouvelle méthode d'inférence sur des données simulées. Schématiquement, cette librairie R `gspace2infr` (1) appelle le simulateur `GSpace`, dont les sources sont incluses dans `gspace2infr`, pour lancer un grand nombre de simulations de jeux de données ; (2) lance le calcul des statistiques résumantes par la librairie `GSumStat`, dont les sources sont également incluses dans `gspace2infr`, sur ces jeux de données simulés ; et (3) appelle des fonctions de la librairie `Infusion` pour l'inférence à partir du tableau de référence produit lors des deux étapes précédentes. La

procédure d'inférence est itérative, ce qui implique que `GSpace` et `GSumStat` sont automatiquement ré-executés au cours de plusieurs itérations, pour de nouvelles valeurs de paramètres déterminées par `Infusion`.

Les chapitres suivants de cette thèse présentent donc ces étapes successives :

Les propriétés du programme `GSpace` ;

Les propriétés de la librairie `GSumStat` ;

Un exemple d'utilisation de `gspace2infr` ;

L'évaluation de la performance de la méthode de *summary likelihood* dans le contexte de l'inférence des paramètres de dispersion, de taille et de densité d'une population structurée en isolement par la distance, selon que les déséquilibres de liaison sont pris en compte ou non.

2

Le programme de simulation `GSpace`

Comme nous l'avons vu précédemment, l'inférence par simulation permet de contourner les limites des méthodes statistiques basées sur des approximations analytiques. Cependant, les logiciels permettant la simulation de données génomiques (prenant en compte le déséquilibre de liaison) dans le cadre de modèles en populations structurées sans les approximations découlant de l'hypothèse d'une grande taille de population sont rares ou lents.

A cette fin fut créé `GSpace`, un programme de simulation des données génomiques neutres prenant en compte la recombinaison sous un large éventail de modèles démographiques. Il simule un échantillon d'individus haploïdes ou diploïdes suivant un cycle de vie haplo-diploïde standard. Il est basé sur un processus de coalescence en génération par génération couplé à un algorithme de recombinaison. De plus, il permet de gérer de manière simple le processus de dispersion dans l'espace. Ce chapitre a pour vocation d'expliciter

et de détailler ces différents points. Les spécificités les plus notables de GSpacesont le fait de combiner une prise en compte des effets de petite taille de population, de la liaison génétique entre marqueurs, et des processus de dispersion localisée dans l'espace.

2.1 Cycle de vie et modèles démographiques

2.1.1 Cycles de vie haplo-diploïde avec recombinaison

Tous les organismes passent par une séquence de changements se répétant de manière cyclique à travers les générations. On parle de cycle de vie (Bell and Koufopanou, 1991). Dans le cas des organismes sexués ce cycle est dit "haplo-diploïde" puisqu'il consiste en l'alternance d'une phase haploïde à n chromosomes et d'une phase diploïde à $2n$ chromosomes (à l'exception notable de certaines algues rouges tel que *Antithamnionella* sp. pour lesquels il existe une phase supplémentaire). Ce cycle de vie peut être, en termes de temps passé dans chaque phase, majoritairement haploïde (comme chez de nombreuses algues, e.g *Spirogyra* sp.), majoritairement diploïde (comme chez la plupart des vertébrés e.g *Homo sapiens*) ou un équilibre entre les deux (e.g *Laminaria digitata*, de nombreux insectes sociaux).

Dans le cadre de cette thèse nous nous sommes intéressés à l'impact de la prise en compte de la recombinaison sur les capacités de nos méthodes à inférer des paramètres démographiques. Nous définirons la recombinaison comme étant le processus permettant l'échange de matériel génétique entre les locus de deux chromosomes homologues lors de la prophase de la méiose (voir figure 2.1).

2.1.2 Modèle de Wright-Fisher

En génétique des populations il existe un certain nombre de modèles démographiques classiquement utilisés. L'un des premiers modèles est celui de Wright (1931) et Fisher (1930) que je vais ici rapidement expliciter à titre d'exemple.

Dans ce modèle, on considère une population de taille constante et finie de N individus haploïdes, à générations non chevauchantes, panmictique et isolée. De plus, tous les individus ont, en espérance, le même succès reproducteur (i.e. de 1 descendant par individu

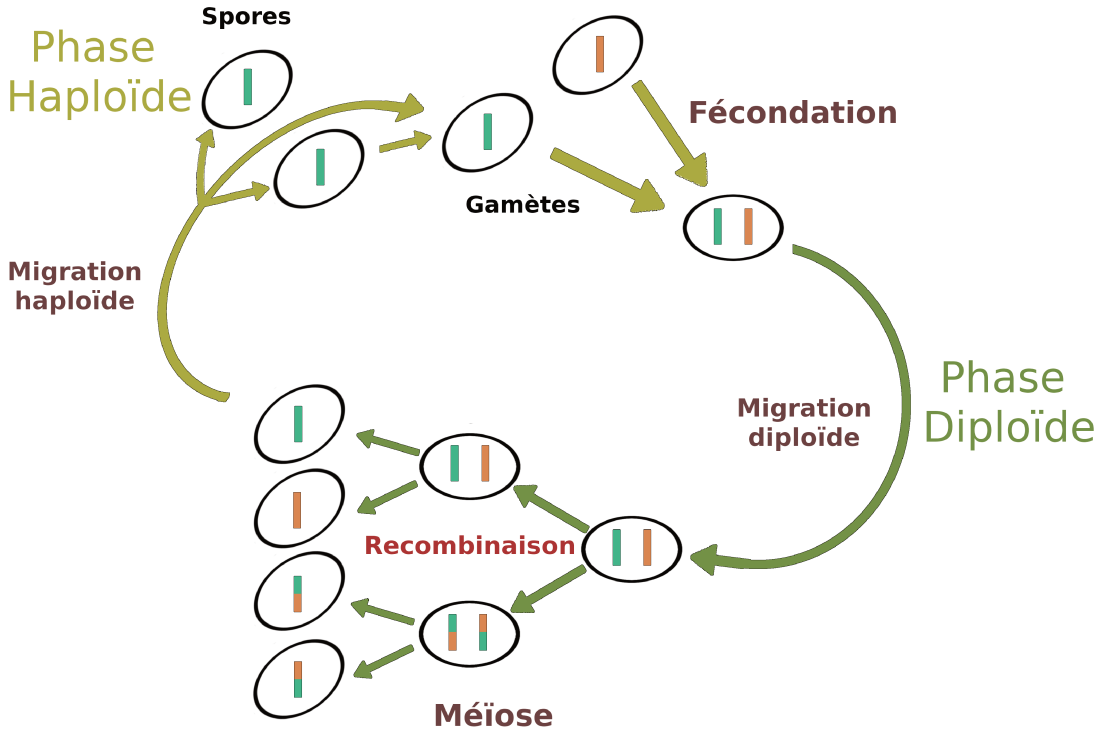


Figure 2.1 – Cycle de vie "haplo-diploïde" représentant le passage des chromosomes au cours des différentes phases possibles d'un organisme. Ici un chromosome de la paire de chromosomes maternels (en vert) et de la pair de chromosome paternels (en orange) échangent du matériel génétique durant la méiose.

puisque la taille de la population est constante). Le nombre d'individus étant fini, il existe une probabilité non nulle qu'un individu ne puisse transmettre ses gènes à la génération suivante. En suivant Kimura (1968), si l'on considère le nombre K_i de copies d'un allèle neutre et sa fréquence $p_i = K_i/N$, le nombre de copies de cet allèle dans la population à la génération suivante (K_j) suit une loi binomiale de paramètres N et p_i. K_i peut donc varier de 0 à N en une seule génération, avec $\mathrm{E}[K_j] = Np_i = K_i$ et $\mathrm{Var}[K_j] = Np_i(1-p_i)$ (voir aussi le chapitre 3 de Wakeley 2009).

Si l'on ne considère pas la mutation alors la perte ou la fixation de cet allèle n'est dépendante que de la dérive.

Ainsi, si la fréquence d'un allèle à la génération 0 est p_0 alors la diversité génétique de Nei (Nei 1973) attendue dans la population est $H_{e0} = 2p_0(1-p_0)$. Considérons l'espérance de cette probabilité à la génération 1 :

$$
\begin{aligned}
E[H_{e1}] &= E[2p_1(1-p_1)] \qquad (2.1)\\
&= 2\left(E\left[\frac{K_1}{N}\right] - E\left[\left(\frac{K_1}{N}\right)^2\right]\right) = 2(E[K_1]\frac{1}{N} - E[(K_1)^2]\frac{1}{N^2})\\
&= 2(E[K_1]\frac{1}{N^2} - (E[K_1]^2 + Var[K_1])\frac{1}{N^2})\\
&= 2(K_0\frac{1}{N} - K_0^2\frac{1}{N^2} + Np_0(1-p_0)\frac{1}{N^2})\\
&= 2(p_0 - p_0^2 + p_0(1-p_0)\frac{1}{N}) = 2(p_0(1-p_0)(1-\frac{1}{N}) = H_{e0}(1-\frac{1}{N})
\end{aligned}
$$

On peut ainsi estimer que la diversité génétique dans ce modèle diminue d'un facteur $1/N$ à chaque génération, et l'on retrouve ici le résultat classique que la dérive est fonction de la taille de la population.

Bien que peu coûteux à simuler et simple à analyser, le modèle Wright-Fisher ne tient pas compte du fait que dans la nature, la dispersion spatiale au sein d'une population est le plus souvent limitée (Wright 1943 ; Endler 1977 ; Rousset 1997). Ainsi les populations sont le plus souvent structurées dans l'espace et non pas panmictiques.

2.1.3 Modèles de populations structurées

Les flux de gènes/la migration entre sous-populations (on parle de dèmes) panmictiques peuvent être modélisés au sein d'un certains nombres de modèle.

Modèle en île

Selon le modèle en île (Wright 1931), la population globale est structurée en n dèmes de tailles N_{ind}, échangeant des migrants à un taux m (taux de dispersion uniforme sur l'habitat). Un des résultats classiques issu de ce modèle est que le taux de différentiation moyen entre les dèmes $F_{\mathrm{ST}} = \frac{1}{1+2Nm}$ avec N le nombre de copies de gènes dans le dème. Cependant les probabilités de migration étant identiques entre toutes les paires de dèmes (isotropie), ce modèle ne permet pas d'étudier l'impact de la dispersion quand celle-ci est limitée aux dèmes les plus proches.

Stepping stone

Le modèle en stepping stone (Kimura and Weiss 1964) est un modèle où les dèmes sont placés sur une grille régulière et n'échangent des migrants qu'avec leurs voisins immédiats. C'est un modèle qui prend en compte la dispersion limitée dans l'espace (dans le cas d'un stepping stone non circulaire) et donc l'aspect spatial. Bien que dans ce modèle la dispersion soit limitée, il ne permet pas de simuler la survenue d'événements de dispersion à longue distance qui, bien que rares, sont des évènements impactant fortement l'organisation de la diversité au sein des populations (Endler 1977, Excoffier and Ray 2008).

Isolement par la distance (IBD)

Le modèle en isolement par la distance (Wright 1943) est un modèle permettant de décrire la dispersion des populations naturelles de manière plus pertinente que les précédents. L'idée est de modéliser la dispersion non plus uniquement par un taux de migration mais aussi par une distribution donnant la probabilité qu'un descendant disperse (et se reproduise) dans un dème en fonction de la distance entre ce dème et le dème de son(ses) parent(s).

Ce modèle a pour autre intérêt de contenir les modèles précédents. En effet, si la distribution de dispersion est uniforme sur l'ensemble de l'habitat, on modélise une dispersion équivalente à celle d'un modèle en île. Si cette dispersion est limitée aux dèmes adjacents, on modélise une dispersion équivalente à celle d'un modèle en stepping stone. Bien sûr,

la distribution de dispersion peut avoir d'autres formes (voir figure 2.10) permettant de modéliser à la fois de nombreux évènements de dispersion à courte distance et de rares évènements de dispersion longue distance, comme il est observé chez de nombreuses espèces (pour plus de détails voir la section 2.3.3).

Une difficulté des modèles IBD est qu'ils sont lents à simuler. Pour alléger ce fardeau calculatoire il est intéressant de se pencher sur un processus permettant de simuler efficacement (en temps de calcul) la variation génétique neutre : le coalescent.

2.2 Théorie de la coalescence et graphe ancestral de recombinaison

2.2.1 Principe de la théorie de la coalescence

Dans un cadre neutre, il est possible de séparer les processus mutationnels et les processus démographiques. Par définition, le nombre de descendants d'un gène [1] ne dépend pas de son état allélique. Les mutations n'affectant pas la généalogie, cette dernière ne dépend que des paramètres démographiques et reproductifs (e.g. l'auto-fécondation).

En se basant sur les réflexions de Malécot [2] pour un échantillon d'individus de la population, il est possible de retrouver leurs séries de copies de gènes ancestrales (nommées lignées ancestrales dans la suite de cette thèse) en remontant le temps. Dans ce cadre lorsque deux lignées ancestrales sont issues du même gène parental on dit qu'elles coalescent. De coalescence en coalescence, il est ainsi possible de retrouver l'ancêtre commun le plus récent à toutes ces lignées (MRCA pour *most recent commun ancestor*, voir figure 2.2).

1. Dans ce document, on suit une terminologie selon laquelle un "gène", ou une "copie de gène" est l'information portée par un chromosome donné à un locus donné. Un individu diploïde peut ainsi transmettre deux gènes distincts à chaque locus à ses descendants, appartenant ou non à la même classe allèlique.

2. "Le lien de filiation qui unit l'un [des] individus à l'un des ancêtres [...] est une chaîne d'ascendance. Deux chaînes d'ascendance aboutissant à un même ancêtre forment une chaîne de parenté entre les deux individus pris chacun sur l'une des chaînes d'ascendance exclusivement." Malécot 1966

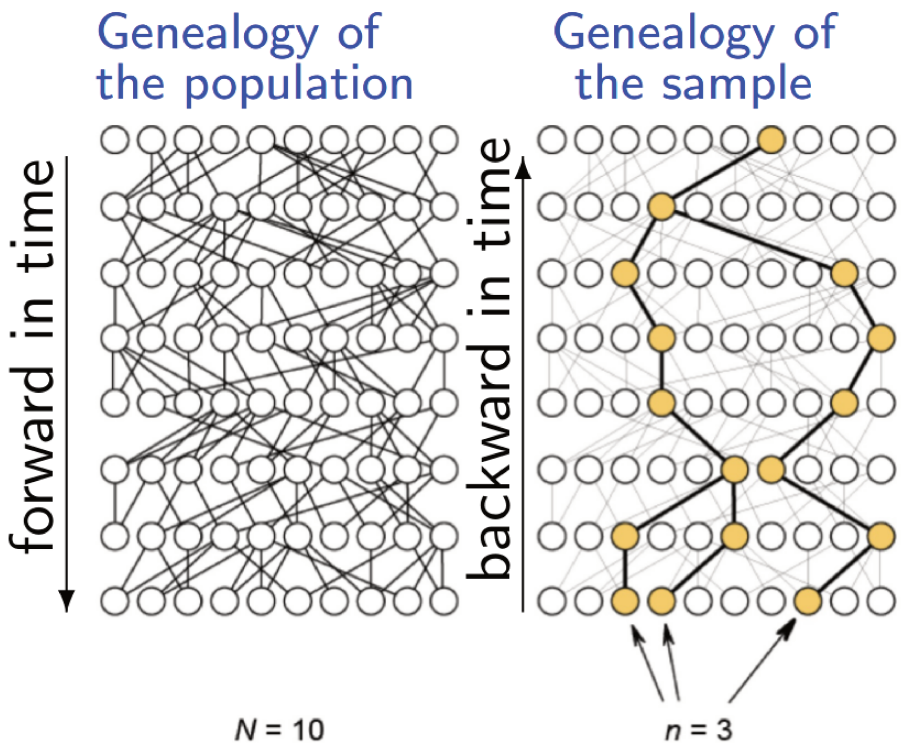


Figure 2.2 – Représentation des lignées de gènes au sein d'une population théorique de 10 gènes évoluant dans un modèle de Wright-Fisher. A gauche en avançant dans le temps et à droite en remontant dans le temps. Pour l'échantillon de 3 gènes (couleur jaune), à la dernière génération il est possible de remonter le long de leurs lignées ancestrales respectives afin de retrouver leurs MRCA.

2.2.2 Le coalescent de Kingman

Le coalescent de Kingman (Kingman 1982) est un processus décrivant la création de l'arbre de lignées ancestrales d'un ensemble de gènes remontant jusqu'à leur ancêtre commun le plus récent (voir figure 2.3). Ce processus consiste à traquer les lignées ancestrales en remontant le temps. Lorsque deux lignées sont copies d'un même ancêtre, on continue le processus en ne considérant qu'une seule lignée ancestrale. Ainsi un échantillon de k lignées comprendra $k-1$ évènements de coalescence. Les évènements de coalescence ont lieu à un temps $0 < T_i < \infty$ pour tout $0 < i < k$. La coalescence reconstitue, pour chaque lignée, un arbre généalogique, dit arbre de gène, sur lequel il est possible d'appliquer des mutations (voir section 2.3.4). Du fait de la dérive, la coalescence est un processus stochastique. Cependant, Kingman et d'autres auteurs (par exemple Möhle, 1998) ont montré qu'il s'agissait du processus ancestral limite pour une variété de modèles (dont Wright-Fisher).

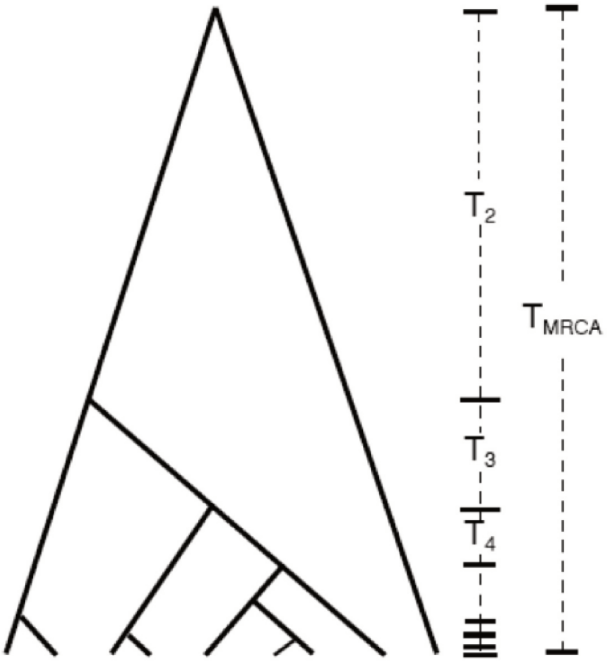


Figure 2.3 – Arbre de coalescence issu d'un coalescent de Kingman. Les temps auxquels ont lieu un évènement de coalescence entre deux lignées ancestrales (T_x) se distribuent en suivant une loi exponentielle.

Dans le cadre d'un modèle Wright-Fisher il est simple de calculer un certain nombre d'attendus sur l'arbre de coalescence. Par exemple, vu que la probabilité que 2 lignées ancestrales coalescent dans une population haploïde de taille N à la génération précédente est $P(G_2 = 1) = 1/N$, la probabilité qu'elles coalescent en g générations peut s'écrire :

$$P(G_2 = g) = \frac{1}{N}\left(1 - \frac{1}{N}\right)^{g-1} \tag{2.2}$$

c'est-à-dire une loi géométrique. À condition que $\frac{1}{N} \ll 1$ cette équation s'approxime en :

$$P(G_2 = g) \simeq \frac{1}{N}\left(e^{-\frac{1}{N}}\right)^{g-1} = \frac{1}{N}e^{-\frac{g}{N}}e^{\frac{1}{N}} = \frac{1}{N}e^{-\frac{g}{N}} \tag{2.3}$$

La distribution de G_2 suit ainsi une loi exponentielle de paramètre $\lambda = 1/N$ et d'espérance N. Ainsi en moyenne N générations sont nécessaires à deux lignées pour coalescer.

De la même manière, sachant que la probabilité que 2 lignées parmi k lignées coalescent à la génération précédente est $P(G_k = 1) = \binom{k}{2}\frac{1}{N}$ la probabilité qu'elles coalescent en g générations suit une loi géométrique qui peut s'approximer par une loi exponentielle :

$$P(G_k = g) \simeq \binom{k}{2}\frac{1}{N}e^{\binom{k}{2}-\frac{g}{N}} \tag{2.4}$$

Toutes les formes des arbres étant équiprobables il est possible d'affirmer que cette distribution exponentielle caractérise complètement le coalescent de Kingman. En faisant l'hypothèse qu'il y a $k-1$ évènements de coalescence de paires de gènes le long de l'arbre (c'est-à-dire aucun évènement de coalescence multiple), la distribution de chaque temps de coalescence successif pour $j = k, \ldots, 2$ suit ainsi une loi exponentielle de paramètre $\lambda = \frac{j(j-1)}{2N}$ et d'espérance $\frac{2N}{j(j-1)}$. le temps de coalescence de l'échantillon complet est :

$$\begin{aligned} E[T_{MRCA}] &\simeq \sum_{j=2}^{k} E[G_j] = \sum_{j=2}^{k} \frac{2N}{j(j-1)} \\ &= 2N \sum_{j=2}^{k} \left(\frac{1}{j-1} - \frac{1}{j}\right) = 2N(1 - \frac{1}{k}) \end{aligned} \tag{2.5}$$

Ainsi l'ensemble de l'histoire génétique de l'échantillon de taille k considéré au sein d'une population de taille N peut-être simulé en $k-1$ évènements tirés dans des lois exponentielles successives de paramètre $\lambda = \frac{j(j-1)}{2N}$ pour j décroissant de k à 2.

Le coalescent de Kingman peut être étendu à des modèles avec migration, avec des populations de tailles variables, *etc.* Cependant, l'utilisation du coalescent est limitée par une hypothèse forte : les probabilités d'évènements de coalescence, de recombinaison et de migration dans le modèle doivent rester faibles.

2.2.3 La coalescence génération par génération

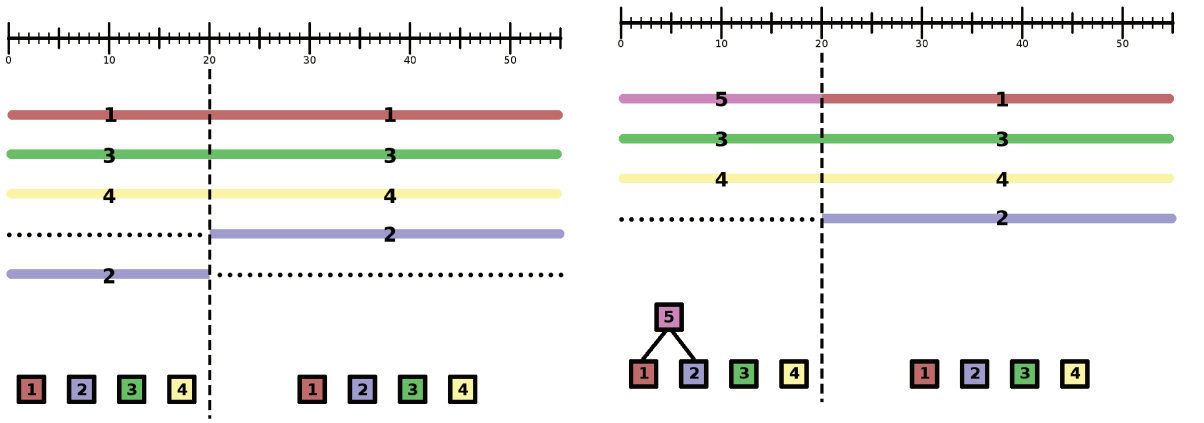
Dans le coalescent de Kingman ces faibles probabilités de coalescence peuvent se traduire par l'obligation que $\forall k, \frac{k}{N} \ll 1$ afin que la probabilité d'un évènement de coalescence entre plus de 2 lignées ancestrales soit négligeable. Ainsi si $N \to 1$ (ou si $k \to N$), il est évident que le coalescent n'est plus valide pour simuler sans biais les arbres de gène de l'échantillon.

Le processus de coalescence peut être simulé sans biais en ignorant l'hypothèse du faible taux d'évènement si l'on considère un processus itératif de génération en génération, autorisant les coalescences multiples à chaque génération jusqu'à ce que les $k-1$ lignées se soient éteintes.

Ainsi, en autorisant que plus de deux lignées coalescent en une unique lignée ancestrale à une génération donnée, on accélère la coalescence globale des lignées ancestrales, avec par exemple pour effet un $E[T_{MRCA}] < 2N(1-\frac{1}{k})$ (Fu 2006).

2.2.4 Graphe ancestral de recombinaison (ARG)

Considérer la recombinaison dans un processus de coalescence n'est pas trivial. En effet, là où en remontant dans le temps la coalescence va "rassembler" les lignées ancestrales (réduisant donc leurs nombres), la recombinaison va les "séparer", créant deux arbres ancestraux de coalescence pour les lignées, séparés par chaque nouvel évènement de recombinaison (voir figure 2.4).

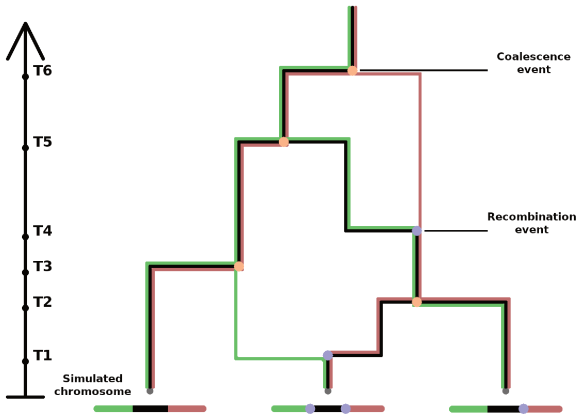


(a) Représentation d'un évènement de recombinaison ayant lieu à la 20ème position de recombinaison possible du chromosome portant la lignée ancestrale 2. Les points noirs représentent les lignées ancestrales non retrouvées dans l'échantillon. Cet évènement oblige à considérer deux arbres de gènes partiellement indépendants, un pour le segment non recombinant $[0; 20[$ et l'autre pour le segment $[20; 50[$.

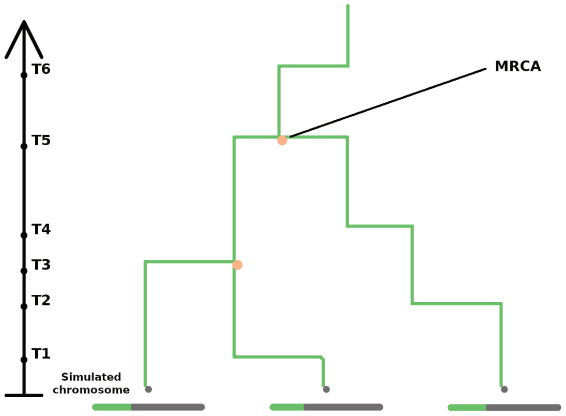
(b) Représentation d'un évènement de coalescence entre la lignée ancestrale 2 du segment $[0; 20[$ et la lignée ancestrale 1 du segment $[0; 50[$. Les deux lignées 1 et 2 coalescent en une lignée ancestrale commune du segment $[0; 20[$. Il n'y a aucun évènement de coalescence à enregistrer entre la lignée suivie 1 et la lignée du segment $[20; 50[$, non retrouvée dans l'échantillon. Cet évènement n'impacte donc que l'arbre de gène du segment $[0; 20[$.

Figure 2.4 – Représentation d'un évènement de recombinaison suivi d'un évènement de coalescence.

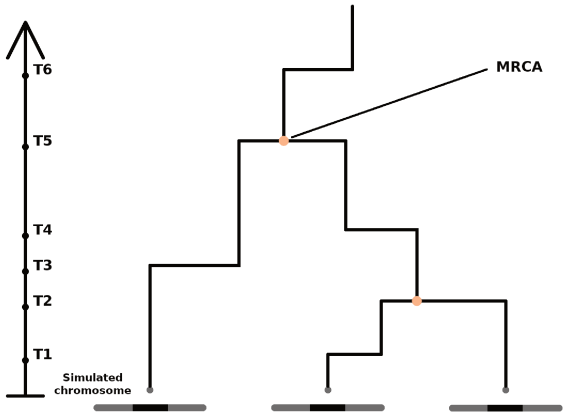
Cette approche mène à reconstruire non pas un arbre par gène mais un ensemble d'arbres, appelé graphe ancestral de recombinaison (voir figure 2.5a, Hudson 1983). Chacun des arbres du graphe ancestral de recombinaison correspond à l'arbre de coalescence d'une partie du génome n'ayant jamais recombiné (ou "segment non recombinant", voir figure 2.5b, 2.5c, 2.5d).



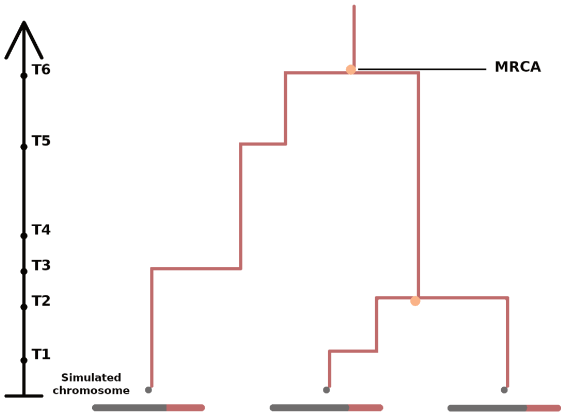
(a) Exemple de graphe ancestral de recombinaison pour un échantillon de 3 individus. Les points mauves représentent les évènements de recombinaison et les points jaunes les évènements de coalescence.



(b) Arbre de gène porté par le segment non recombinant le plus à gauche des chromosomes considérés dans l'échantillon.



(c) Arbre de gène porté par le segment non recombinant central des chromosomes considérés dans l'échantillon.



(d) Arbre de gène porté par le segment non recombinant le plus à droite des chromosomes considérés dans l'échantillon.

Figure 2.5 – Graphe ancestral de recombinaison et ensemble des arbres de gène le composant. On peut remarquer que la topologie des arbres de gène portés par deux segments non recombinants voisins n'est pas indépendante. Elle ne diffère qu'à partir du moment (en remontant dans le temps) où un évènement de recombinaison va séparer les lignées ancestrales portées par les deux segments.

2.3 Implémentation dans GSpace

Pour simuler précisément les effets de petite taille de dème, GSpace combine dans une implémentation en C++ moderne certaines parties de l'algorithme modifié de Hudson (Hudson 1983) pour la recombinaison et la coalescence, précédemment implémentées dans msprime (Kelleher et al. 2016), avec des caractéristiques d'un précédent simulateur génération par génération, IBDsim (Leblois et al. 2009). Il se distingue d'IBDsim par le fait que ce dernier simule des marqueurs indépendamment sans recombinaison et pas des individus. Cette dernière hypothèse peut être pratiquement identique au fait de supposer une recombinaison libre entre marqueurs, mais n'y est pas totalement équivalente, du fait que même des marqueurs non liés vont avoir des arbres de coalescence non indépendants. En effet, les marqueurs non liés vont tous être conditionnels à l'arbre généalogique réalisé des individus de la population en fonction des évènements individuels réalisés de migration et de reproduction qui ne sont pas indépendants pour chaque marqueur. Les algorithmes de GSpace prennent en compte la liaison entre marqueurs, mais aussi cette dépendance plus subtile entre marqueurs en recombinaison libre, induisant elle aussi des déséquilibres de liaison (voir figure 2.6).

2.3.1 Algorithme de Hudson

L'algorithme de Hudson implémenté dans ms (premier simulateur sous coalescent de Kingman prenant en compte la recombinaison, voir Hudson 2002) peut se résumer ainsi : en remontant dans le temps, le programme va tirer successivement des évènements de recombinaison et de coalescence (dont les temps sont tirés dans des distributions exponentielles, voir section 2.2.2) jusqu'à ce que les ancêtres communs de toutes les lignées génomiques aient été trouvés. Cependant, faire simuler un nombre importants de marqueurs par ms est assez lent et utilise beaucoup d'espace mémoire.

La problématique de simuler de manière efficace en temps et en espace mémoire un graphe ancestral de recombinaison pour des données génomiques et d'extraire les arbres de gène de chaque partie non recombinante du graphe, fut résolue par Kelleher et al. 2016 dans l'implémentation de msprime. Cependant msprime simule selon le coalescent de Kingman, rendant ainsi ce programme impropre à l'utilisation que nous avons définie ci-dessus.

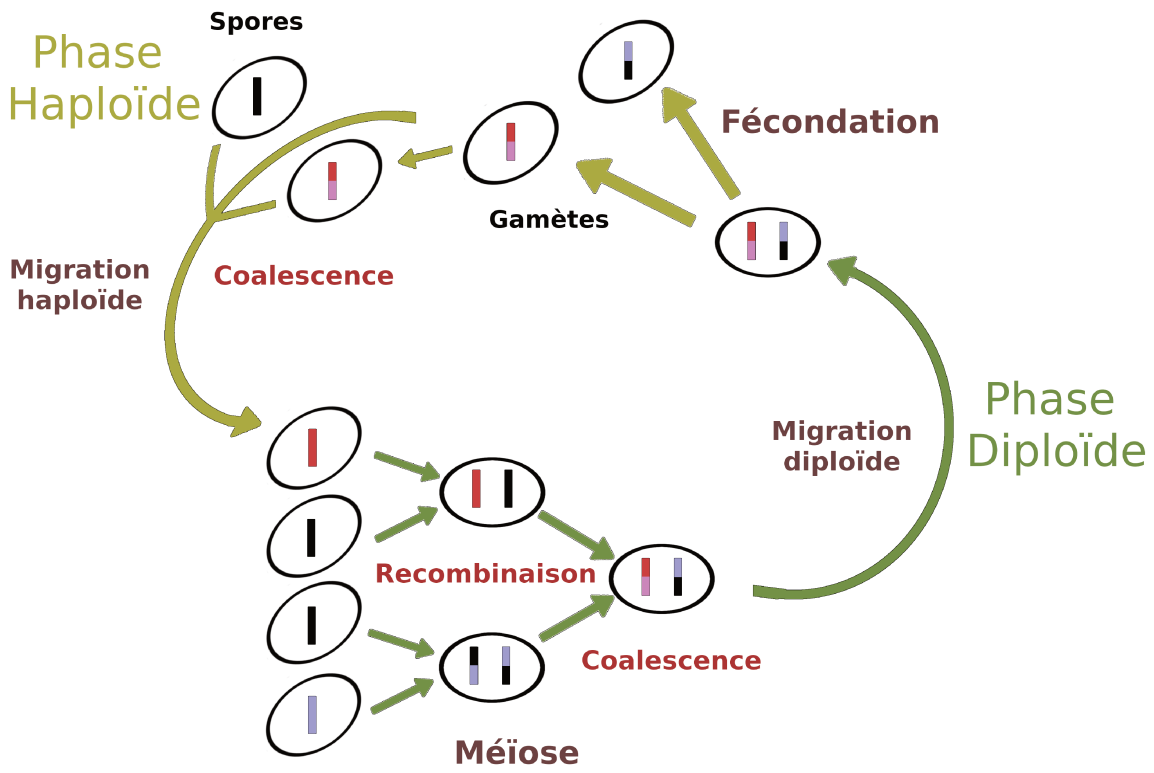


Figure 2.6 – Cycle de vie "haplo-diploïde" représentant, en remontant le temps, la recombinaison et la coalescence des lignées génétiques au cours des différentes phases possibles d'un organisme. Les lignées ancestrales échantillonnées sont représentés en rouge et bleu. Les lignées noires représentent les lignées non échantillonnées. La lignée issue de la coalescence des lignées rouge et bleu est représentée en fuchsia.

Au cours de ma thèse, une extension en génération par génération nommé DTWF (pour *Discrete Time Wright-Fisher*) fut implémentée dans `msprime` (Nelson et al. 2020). Cependant nous verrons à la section 2.3.4 et à la section 2.4 que cette version n'est pas identique à `GSpace`.

Le choix fut fait d'adapter l'algorithme modifié de Hudson (décrit précisément dans l'algorithme 1 ci dessous) afin de le rendre applicable dans le cadre d'un processus de coalescence génération par génération.

Algorithm 1 Algorithme de coalescence et de recombinaison de `msprime` dans le cadre d'un modèle de Wright-Fisher

procedure MS(nombre de lignées n'ayant pas encore trouvé leur MRCA (n_l), longueur totale des n_l lignées (L_t), taux de recombinaison entre site par génération (ρ))
$t = 0$
  while $n_l \neq 0$ **do**
    $\lambda_r = \rho.L_t$, $\lambda = \lambda_r + n_l(n_l - 1)$, et $t = t + random\_exponentielle(\lambda)$.
    Tirer un événement entre coalescence (COA) et recombinaison (REC)
    if $random\_uniform_r(0,1)$ [3] $< \lambda_r/\lambda$ **then**
      $site = random\_uniform_e(1, L)$ [4]
      Faire $REC(site)$
    else
      (x, y) =tirage uniforme parmi les $n_l(n_l - 1)/2$ paires distinctes de lignées
      Faire $COA(x, y, t)$
procedure REC(site)
  Séparer la lignée x en deux nouvelles lignées au niveau de *site*
  $n_l = n_l + 1$
procedure COA(x, y, t)
  Enregistrer un évènement de coalescence à t entre la lignée x et la lignée y en une nouvelle lignée z
  $n_l = n_l - 2 + 1$
  $L_t = L_t - Longueur_x - Longueur_y + Longueur_z$

4. Tirage uniforme d'un nombre réel entre 0 et 1.
4. Tirage uniforme d'un nombre entier entre 1 et L.

2.3.2 Algorithme de coalescence en génération par génération

Dans un processus de coalescent de Kingman, la probabilité que deux évènements de coalescence et/ou de recombinaison arrivent au même moment sur une ou plusieurs lignée(s) ancestrale(s) est négligeable. Cette propriété se traduit dans l'algorithme de Hudson (voir algorithme 1) par le fait qu'un seul évènement est considéré à chaque étape de l'algorithme.

Dans le cadre d'un processus de coalescence génération par génération, plusieurs évènements de recombinaison et/ou de coalescence peuvent advenir simultanément. Prendre en compte ceci dans l'algorithme en génération par génération nécessite de transformer profondément l'algorithme 1 (voir algorithme 2). Par exemple, considérer tous les évènements de recombinaison au sein d'une même génération amène à devoir gérer les potentiels multiples évènements de recombinaison ayant lieu sur un même chromosome.

Multiples évènements de recombinaison sur un même chromosome

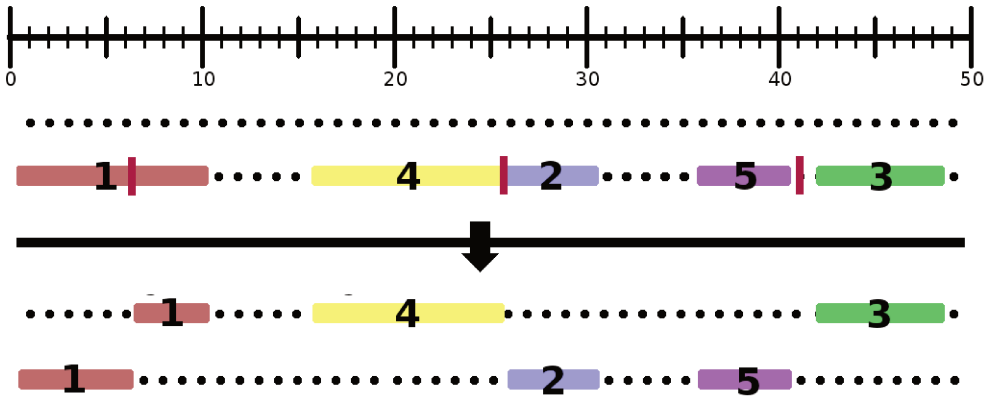


Figure 2.7 – Représentation de multiples évènements de recombinaison sur un chromosome entre les segments $[0;6[$, $[6;16[$, $[16;41[$ et $[41;50[$. Les lignées 1, 4 et 3 se retrouvent portées par les segments $[6;16[$ et $[41;50[$ du chromosome homologue.

En accord avec les hypothèses du coalescent, l'algorithme de Hudson simule un unique évènement de recombinaison à chaque étape. Les probabilités que deux évènements ainsi simulés se succèdent sur un même chromosome sur une même génération étant nulles (dans le cadre du coalescent de Kingman) il est possible de ne simuler que la séparation

Algorithm 2 Algorithme de coalescence et recombinaison de GSpace dans le cadre d'un modèle de Wright-Fisher en génération par génération

procedure GS(nombre de lignées n'ayant pas encore trouvé leur MRCA (n_l), longueur totale des n_l lignées (L_t), taux de recombinaison entre site par génération (ρ))

$g = 0$

while $n_l \neq 0$ **do**

$nbr\_rec = random\_exponentielle(\rho.L_t)$

while $nbr\_rec \neq 0$ **do**

$site = random\_uniform_e(1, L_t)$

Faire $REC(site)$

$nbr\_rec = nbr\_rec - 1$

Pour chaque lignée (l_n) assigner un parent (P_z) parmi les N parents possibles.

for x allant de $0; n_l$ **do**

$z = random\_uniform_e(1, N)$

$P_z \leftarrow l_n$

Coalescer les lignées (l_n) assignées à un même parent (P_z).

for z allant de $[0; N[$ **do**

while $|P_z| > 1$ **do**

$x = l_1, y = l_2$

Faire $l_1 = COA(x, y, g)$

$g = g + 1$

des lignées ancestrales en remontant dans le temps (voir figure 2.4a), en ignorant le liaison possibles entre lignées sur un chromosome homologue. En effet, en cas de recombinaison multiple, deux (ou plus) lignées ancestrales portées par deux (ou plus) segments non recombinants du chromosome recombiné peuvent être originaire de deux (ou plus) segments non recombinants du chromosome homologue (voir figure 2.7). L'histoire ancestrale de ces lignées portées par des segments non recombinants n'est pas indépendante.

L'algorithme en génération par génération se doit de prendre en compte explicitement le chromosome homologue afin de gérer les évènements multiples de recombinaison. Cette prise en compte passe par la représentation explicite du chromosome homologue dans l'algorithme en génération par génération même si ce chromosome ne porte pas de lignées ancestrales (chromosome fantôme) lors de la phase diploïde.

Cette représentation en paire de chromosomes homologues lors de la phase diploïde du cycle de vie amène à devoir gérer la prise en compte explicite des deux chromosomes parentaux lors de la reproduction sexuée, ce qui n'était pas nécessaire dans les algorithmes basés sur le coalescent du fait des évènements de recombinaison uniques.

Prise en compte des chromosomes paternels et maternels lors de la reproduction sexuée

Pour des organismes recombinants, les lignées portées par des chromosomes homologues au sein d'un même individu proviennent des deux chromosomes parentaux différents. En coalescence génération par génération cet aspect nécessite d'assigner aléatoirement chacun des chromosomes homologues d'un individu à un chromosome parental différent (voir figure 2.8). La probabilité de coalescence des lignées ancestrales devient ici conditionnelle à l'origine parentale du chromosome qui les portent. Cela revient a choisir d'abord le parent ancêtre avec une probabilité de $1/N_{ind}$, puis le chromosome ancestral chez ce parent avec une probabilité de $1/2$, pour chaque évènement de coalescence, au lieu de choisir directement le chromosome ancestral avec une probabilité de $1/2N_{ind}$ sous le coalescent de Kingman.

La figure 2.8 nous permet d'observer que, dans le cadre de prise en compte de la fécondation en coalescence, retrouver l'individu parental commun à deux gamètes n'entraîne pas obligatoirement la coalescence des lignées ancestrales qu'ils portent. En effet, ici, la

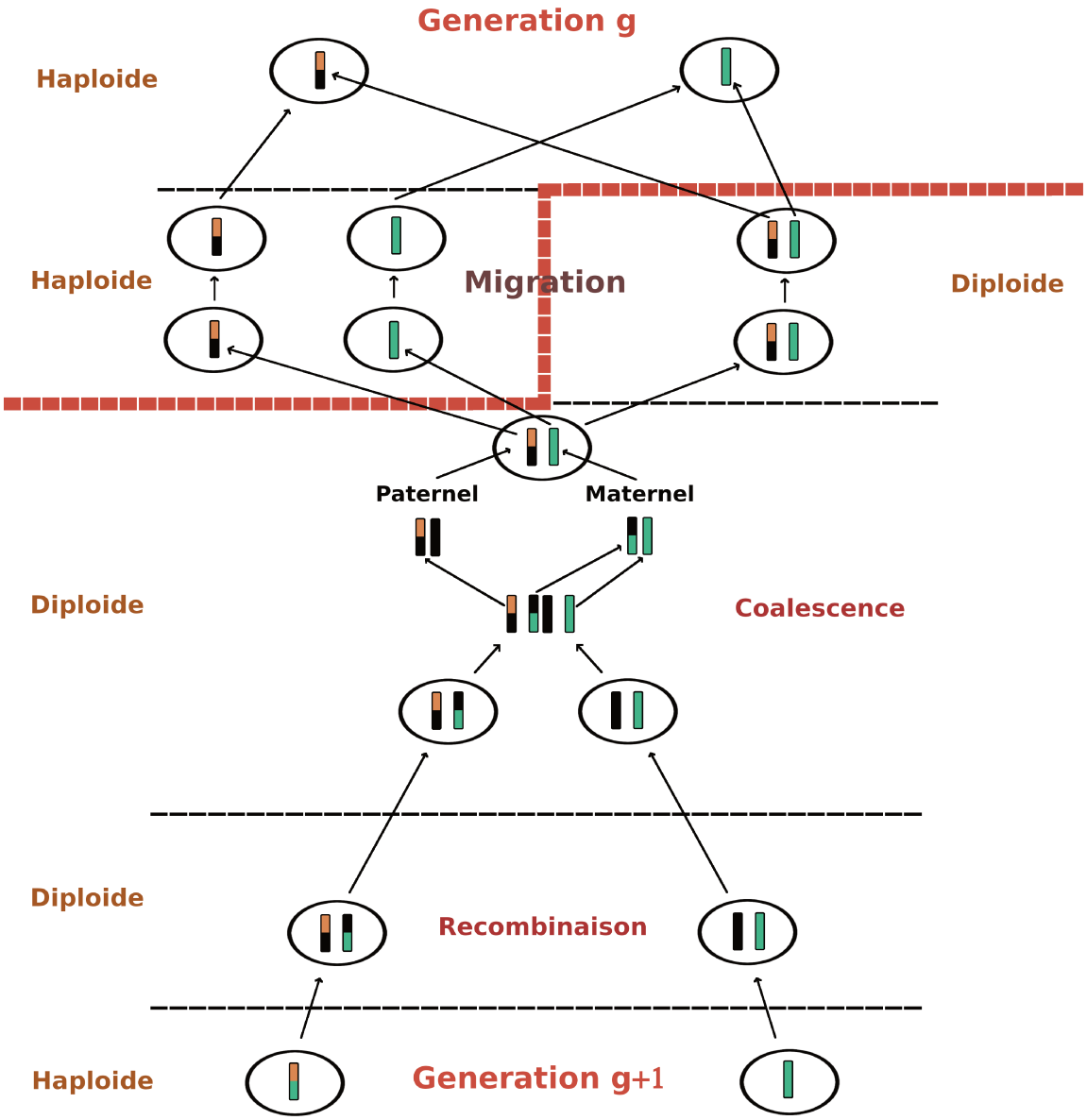


Figure 2.8 – Représentation aplanie du cycle de la figure 2.6. Les lignées ancestrales échantillonnées sont représentées en verte (lignée portée par le chromosome maternel) et orange (lignée portée par le chromosome paternel).

lignée verte et la lignée orange, portées par le même chromosome se retrouvent séparées en raison d'un évènement de recombinaison. Par la suite les chromosomes portant la lignée verte se retrouvent assignées au chromosome maternel permettant à ces dernières de coalescer. Le chromosome portant la lignée orange étant d'origine paternel il peut par la suite migrer de manière indépendante ou avec les lignées vertes coalescées selon que la migration est haploïde (gamétique) ou diploïde (individu juvénile).

Un aspect intéressant émergeant de la prise en compte des chromosomes homologues est qu'il est possible de faire migrer ces derniers de manière non indépendant au moment de la phase diploïde du cycle, permettant de modéliser de façon plus pertinente la dispersion des juvéniles diploïde, qui est plus courante que la migration gamétique chez la majorité des espèces (animales) à phase diploïde majoritaire.

2.3.3 Gestion de la migration sur un réseau de dèmes

L'algorithme de migration utilisé dans `GSpace` est issu de `IBDsim` et peut être résumé de la manière suivante : à chaque génération g, les coordonnées du parent de chaque individu porteur de lignée(s) ancestrales sont tirées aléatoirement dans la distribution de dispersion "arrière" (*backward dispersal distribution*) donnant la probabilité de la position d'un parent, étant donné la position du descendant sur le réseau.

Cette distribution est calculée de deux façons différentes selon que la population est isotrope ou anisotrope. Dans le cadre d'une population isotrope, tous les dèmes ont la même taille, les distributions de dispersion sont identiques pour chaque dème et il n'y a pas d'effets de bord de l'habitat (habitat en cercle ou en tore, voir figure 2.9). Dans ce cas, la distribution de dispersion arrière est identique à la distribution de dispersion avant (*forward distribution*) qui décrit la probabilité qu'un descendant se reproduise à une certaine distance de l'endroit où ses parents se sont reproduits.

Dans le cadre d'une population anisotrope (dèmes de tailles différentes, distribution non identique pour chaque dème, ou juste un habitat avec des effets de bords) les distributions arrières sont potentiellement différentes en chaque noeud du réseau, et sont calculées en fonction de la distribution avant en tenant compte de la taille relative des dèmes "atteignable" par dispersion à partir de de ce noeud du réseau. En effet, les dèmes occupés par de plus grand nombre d'individus contribuent plus au nombre de migrants

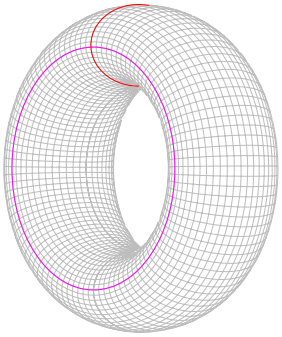


Figure 2.9 – Représentation graphique d'une grille en tore.

arrivant dans un dème, donc, pour tout deme d_1, la probabilité p_{d_1,d_2} qu'une lignée soit migrante de d_2, en fonction de $f_{d_1-d_2}$ la probabilité de dispersion avant entre les deux dèmes est égale à

$$p_{d_1,d_2} = \frac{N_{d_2} f_{d_1-d_2}}{\sum N_{d_x} f_{d_1-d_x}} \tag{2.6}$$

où la somme est calculée sur tous les demes d_x à une distance maximale de dispersion par rapport à d_1.

Ce type de calcul permet donc non seulement de prendre en compte les variations de la taille des dèmes dans un espace hétérogène mais aussi les effets de bord. En effet, l'habitat peut être supposé linéaire (représentant par exemple une rivière ou un trait de côte), ou bidimensionnel avec des effets de bord. Dans ce second cas, on suppose par simplicité que la dispersion se produit indépendamment dans chaque dimension (c'est-à-dire selon deux axes orthogonaux), ce qui n'est pas forcément très réaliste. En particulier, les distributions ainsi obtenues peuvent s'éloigner fortement de distributions symétriques par rotation. Vouloir approcher un telle symétrie sur un réseau discret mènerait à des complications mais pourrait être une piste intéressante à explorer dans l'avenir.

Les distributions de dispersion biologiquement réalistes ont souvent des fortes probabilités à des distances proches, associées à une probabilité de dispersion décroissant lentement à longue distance (on parle de kurtosis élevé, Endler 1977, Kot et al. 1996). Cependant, les distributions de probabilité discrètes couramment utilisées ne sont pas les plus appropriées pour représenter cela car elles impliquent qu'un kurtosis élevé ne peut être obtenu qu'en supposant une faible probabilité de dispersion, c'est-à-dire que la plupart des descendants se reproduisent exactement là où leurs parents se sont reproduits (Rousset, 2000). Par conséquent, sont incluses deux familles de distributions de dispersion

qui permettent une kurtosis et des taux de migration élevés : la distribution de Pareto discrétisée [5] ; et la distribution de Sichel [6] qui permet de modéliser une grande diversité de distributions de dispersion (voir figure 2.10), et dont l'intérêt pour la modélisation de la dispersion est décrit par Chesson and Lee 2005. Des distributions uniformes, gaussienne discrétisée, et géométrique sont également implémentées.

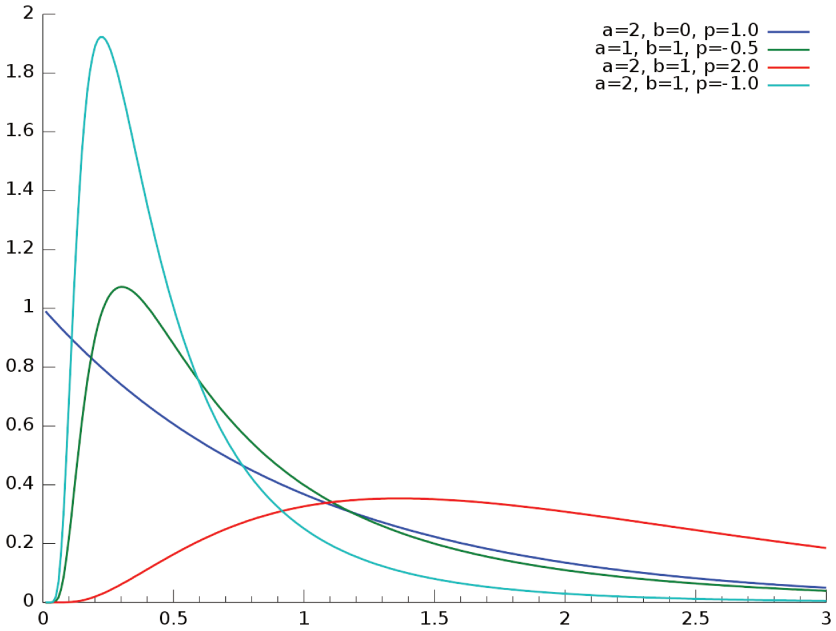


Figure 2.10 – Forme de la distribution inverse Gaussienne généralisée en fonction de la valeur de ses 3 paramètres a, b, p.

2.3.4 Gestion des mutations

Dans le cadre de GSpace la mutation est appliquée après la construction du graphe ancestral de recombinaison en se basant sur le principe de dépendance partielle des topologies des arbres entre eux (les arbres de gène portés par les segments de part et d'autre d'un événement de recombinaison ne différent potentiellement, en remontant dans le temps, qu'à partir de cet évènement de recombinaison, voir figure 2.5). Il est ainsi possible d'optimiser

5. aussi connue comme loi Zeta, de paramètre s, définie pour les entiers positifs $k \geq 1$, et de fonction de masse $\mathbb{P}(X = k) = 1/(k^s \sum_{n=1}^{\infty} \frac{1}{n^s})$

6. si $X|\lambda \sim Poisson(\lambda)$ et si $\lambda|... \sim Gaussienne\ Inverse\ Generalisee(...)$ alors $X|...$ suit une distribution de Sichel telle que $P(X|...) = \int P(X|\lambda)P(\lambda)d\lambda$

grandement la place et le temps investis dans la construction des arbres des segments non recombinants en les construisant successivement pour chaque segment non recombinant le long des chromosomes (comme l'a bien montré Kelleher et al. 2016).

Chaque arbre pouvant être descendu indépendamment il suffit par la suite de parcourir les arbres en les descendant et en leur appliquant un modèle mutationnel. GSpace considérant un nombre fini de sites discrets de mutation (i.e. les nucléotides pour des données de type séquences ADN ou des locus pour des données de type allélique), contrairement au modèle à nombre de site infini (ISM pour *infinite site model*), un grand nombre de modèles mutationnels est applicable aussi bien pour des données de type micro-satellite que des données de séquence (Jukes-Cantor, *etc*). Par contre GSpace ne peut simuler le modèle en sites infinis, qui représente une infinité de points de recombinaison possibles sur un intervalle continu comme implémenté dans ms et msprime.

2.4 Comparaisons avec d'autres simulateurs

GSpace (version 1.8) est, au moment où est écrite cette thèse, le seul simulateur de coalescence génération par génération spécifiquement conçu pour traiter la recombinaison et permettant de manipuler facilement (et de manière optimisée en termes de temps de calcul) des modèles de migration complexes.

Il ne peut donc pas être comparé en termes de temps de calcul à d'autres simulateurs en ce qui concerne des modèles d'IBD complexes, mais il a tout de même été comparé aux algorithmes de cinq autres simulateurs dans le cas de modèles plus simples : le coalescent de Kingman implémenté dans msprime 1.1.1 ('msprime') et FastSimcoal2 2.7.5 ('fsc2') ; un algorithme en génération par génération implémenté dans msprime 1.1.1 ('DTWF') ; et un algorithme de simulation en avançant dans le temps SLiM 3.7.1 ('Slim', voir table 2.1).

Comme attendu, les simulateurs basés sur les approximations du coalescent de Kingman sont les plus rapides dans le modèle en Wright-Fisher. Cependant, dans le modèle en île (cas B et C) on constate de façon remarquable que GSpace peut être plus rapide bien qu'il ne fasse pas ces approximations. De plus, l'écart dans le cas le plus défavorable à GSpace n'est pas considérable et pourrait sûrement être réduit. Par ailleurs, GSpace est toujours plus rapide que les autres simulateurs génération par génération.

	Simulateurs				
Cas	msprime	fsc2	GSpace	DTWF	Slim
A	1.031	11.457	16.393	45.966	112.829
B	95.918	25.770	17.476	164.619	110.527
C	0.409	0.038	0.090	9.901	49.422

Table 2.1 – Comparaison des temps de calcul entre GSpace et d'autres simulateurs sous trois modèles démographiques différents. Temps d'exécution moyen en secondes sur 100 (10 pour SLiM) réplicats pour la simulation d'un échantillon de 1000 individus haploïdes portant un seul chromosome de 10^7 paires de bases, avec des taux de mutation et de recombinaison de 10^{-8} par génération par site sous : (A) un modèle de Wright-Fisher avec une taille de population de 10000 individus haploïdes ; (B-C) un modèle en îles avec 20 sous-populations échangeant des migrants avec une probabilité de 0.05, chacune composée de 500 individus haploïdes et 50 individus échantillonnés avec un chromosome de (B) 10^7 paires de bases ou (C) 10^4 paires de bases.

3

La librairie de calcul de statistiques résumantes `GSumStat`

Afin d'effectuer des inférences par simulation dans le cadre des modèles spatialisés, en plus du simulateur décrit dans le chapitre précédent, il est nécessaire de pouvoir calculer des statistiques sur les données génomiques portant de l'information sur les paramètres d'intérêt des modèles spatiaux (densité de population, dispersion, ...) et adaptées aux différents jeux de données.

Afin de pouvoir calculer des statistiques diverses de façon efficace et de pouvoir complètement maîtriser ces calculs de statistiques résumantes, il est vite apparu nécessaire de développer une bibliothèque de calcul de statistiques résumantes nommée `GSumStat`. Un objectif plus spécifique de cette librairie est de traiter efficacement des données génomiques de grande taille et de calculer en un temps raisonnable des statistiques résumantes en lien avec la différentiation géographique et le déséquilibre de liaison présents dans les données génomiques.

Ce chapitre présente l'ensemble des statistiques utilisées dans les chapitres suivants. Si certaines sont très classiques, des différences ou nouveautés par rapport à la littérature seront ici présentées pour les statistiques descriptives des patrons de déséquilibre de liaison dans l'espace, et pour la prise en compte de certains types de données manquantes, en particulier pour le cas de génotypes diploïdes incomplets. L'objectif général du traitement des données manquantes est de limiter les biais d'estimation, notamment pour de faibles tailles d'échantillons avec des données manquantes, afin que les statistiques résumantes soient toujours facilement interprétables, voire restent de bons estimateurs de paramètres bien identifiés.

3.1 Définitions des statistiques résumantes

3.1.1 Statistiques de diversité

`GSumStat` peut calculer la moyenne et la variance des statistiques suivantes sur l'ensemble des locus de l'échantillon :

K_i : Nombre d'allèles dans l'échantillon au locus i.

K_d : Nombre d'allèles moyen au dème d pour tous les locus.

$H_{\mathrm{obs}i}$,$H_{\mathrm{ex}i}$: hétérozygotie observée et attendue selon la terminologie traditionnelle introduite par Nei (1973) au locus i. En particulier

$$H_{\mathrm{ex}i} = \left(1 - \frac{\sum_{\text{allèles } k=1}^{K} n_{k,i}^2}{n_i^2}\right) \frac{n_i}{n_i - 1} \tag{3.1}$$

avec n_i le nombre de copies de gène au locus i dans l'échantillon et $n_{k,i}$ le nombre de copies de l'allèle k au locus i. L'hétérozygotie attendue est aussi classiquement dénommée "diversité génique de Nei".

Var_i : Variance de la taille allélique, pour les marqueurs dont la variation est caractérisée par une longueur d'allèle, tels que les microsatellites.

MGW : Statistique M de Garza and Williamson (2001), définie aussi pour ce type de marqueurs comme le nombre d'allèles divisé par la différence de taille entre allèle le plus grand et le plus petit.

AFS : Spectre de fréquence allélique (voir Ewens 1972) ou, de manière équivalente, le spectre de fréquence par site pour les marqueurs bi-alléliques.

En cas d'échantillons de taille inégale entre loci, la meilleure façon de synthétiser les AFS des différents loci n'est pas évidente. Un AFS synthétique est ici construit en ne considérant que les loci dont la taille d'échantillon est au moins $n_{0.7} := 70\%$ de celle de l'échantillon maximal, et en remplaçant l'AFS de chacun de ces loci par l'AFS moyen de tous les sous-échantillons possible de taille $n_{0.7}$ construits par échantillonnage aléatoire dans l'échantillon total (cette moyenne peut être construite par un calcul analytique simple, sans simulation).

3.1.2 Statistiques de structure géographique

Wright (1943, 1951, 1965) a montré que les variations de fréquences des différents allèles entre les sous-populations peuvent être analysées au travers des indices de fixation ou statistiques F : F_{IT}, F_{IS} et F_{ST} que l'on peut interpréter comme des corrélations entre les états alléliques des gènes tirés à différents niveaux d'une population subdivisée (individus, intra-dème et entre dème). Ces corrélations peuvent être reliées aux probabilités d'identité de paires de (copies de) gènes (Cockerham and Weir 1987) :

$$F_{\mathrm{IS}} \equiv \frac{Q_{\mathrm{wi}} - Q_{\mathrm{biwd}}}{1 - Q_{\mathrm{biwd}}};\ F_{\mathrm{ST}} \equiv \frac{Q_{\mathrm{biwd}} - Q_{\mathrm{bd}}}{1 - Q_{\mathrm{bd}}};\quad F_{\mathrm{IT}} \equiv \frac{Q_{\mathrm{wi}} - Q_{\mathrm{bd}}}{1 - Q_{\mathrm{bd}}} \tag{3.2}$$

en terme des probabilités d'identité entre paires de gènes, au sein d'un même individu diploïde (Q_{wi}), entre deux individus différents dans le même dème (Q_{biwd}), et entre deux individus différents dans deux dèmes différents (Q_{bd}).

Les estimateurs multilocus des F statistiques, $(\widehat{F_{\mathrm{IS}}}, \widehat{F_{\mathrm{ST}}})$, sont ici calculés selon les même formules que dans `Genepop` (Rousset, 2008).

$$\widehat{F_{\mathrm{IS}}} = \frac{\sum_i [MSI - MSG]_i}{\sum_i [MSI + MSG]_i};\ \widehat{F_{\mathrm{ST}}} = \frac{\sum_i [MSP - MSI]_i}{\sum_i [MSP + (n_c - 1)MSI + n_c.MSG]_i} \tag{3.3}$$

avec

$$MSI = \frac{SS_{\mathrm{i}}}{S_1 - n_s};\ MSG = \frac{SS_{\mathrm{g}}}{S_1}; \tag{3.4}$$

$$n_c = \frac{S_1 - S_2/S_1}{n_s - 1}; MSP = \frac{SSp_{\mathrm{i}}}{n_s - 1}; \tag{3.5}$$

où toutes les sommes sont sur les loci et, pour chaque locus i, $S_1 = \sum_{j=1}^{j=n_s} n_j$ est le nombre d'individus génotypés (n_j par dème j échantillonné), et $S_2 = \sum_{j=1}^{j=n_s} n_j^2$, n_s au nombre de dèmes échantillonnés.

Comme détaillé dans la documentation de `Genepop`, les estimateurs unilocus ainsi définis sont équivalents à ceux de Weir and Cockerham (1984) mais, en cas d'échantillons de taille différentes par locus, les estimateurs sont un peu différents de ces derniers.

Les termes intermédiaires de ces calculs (SS_{i},SS_{g},SSp_{i}) sont les sommes, sur les différents allèles observés à un locus, des sommes de carrés qui apparaissent dans une analyse de variance de la fréquence de chaque allèle. Plus de détails peuvent être trouvés dans l'annexe de Rousset (2007).

3.1.3 Statistiques spatiales

Pour pouvoir décrire les patrons spatiaux de différentiation génétique en termes d'un petit nombre de statistiques résumantes, on s'appuie en particulier sur les résultats des modèles d'isolement par la distance qui montrent une relation "linéaire" (ou, plus précisément, affine) entre certaines mesures de différentiation et la distance ou le logarithme de la distance géographique (Rousset, 1997). On peut alors résumer l'information par les deux paramètres d'une régression linéaire entre différentiation génétique et (logarithme de la) distance géographique.

L'inverse de la pente de la régression est, de plus, à une constante près, un estimateur du produit $D\sigma^2$ de la densité de population D et du carré moyen σ^2 de la distance de dispersion parent-descendant. Cette statistique résumante est d'autant plus intéressante que, dans certaines conditions au moins, une comparaison à un estimateur de $D\sigma^2$ a priori plus puissant (car conçu pour approcher l'estimateur par maximum de vraisemblance) montre que la pente de la régression peut préserver l'essentiel de l'information statistique concernant $D\sigma^2$ (Rousset and Leblois, 2012). Néanmoins, en partant du principe que les coefficients de la régression ne contiennent pas toute l'information disponible dans les données concernant les paramètres du modèle, le calcul d'autres statistiques spatiales est inclus dans `GSumStat`.

La variable réponse de la régression est définie en fonction des fréquences de paires de gènes identiques par état, à différentes distances r dans l'espace. On peut voir ces fréquences comme des estimateurs des probabilités d'identité Q_{r} entre paires de gènes à distance r dans l'espace. Outre les statistiques plus spécifiques décrites ci dessous, `GSumStat` calcule des estimateurs binnés des Q_{r} sur des classes de distance, comme leur moyenne entre loci pour une classe de distance géographique donnée $[r, r + \mathrm{d}r]$ ($\mathrm{d}r$ dépendant du nombre arbitraire de classes choisi par l'utilisateur).

`GSumStat` peut calculer la pente et l'ordonnée à l'origine de la régression linéaire des statistiques résumantes multilocus suivantes par rapport à la distance géographique, ou à son logarithme :

$linF_{\mathrm{st}}$, a_{r} et e_{r} sont toutes des statistiques décrivant la différenciation entre dèmes (pour $linF_{\mathrm{st}} = \widehat{F_{\mathrm{ST}}}/(1 - \widehat{F_{\mathrm{ST}}})$) ou entre individus (pour a_{r} et e_{r}), classiquement utilisées pour les analyses d'isolement par distance, voir Rousset (1997) pour $linF_{\mathrm{st}}$, Rousset (2000) pour a_{r}, et Watts et al. (2007) pour e_{r}. En particulier e_{r} est une forme de distance génétique entre deux individus diploïdes, calculée comme

$$e_{\mathrm{r}} = \frac{\widehat{Q}_i + \widehat{Q}_j - \widehat{Q}_{ij}}{1 - \widehat{Q}_w} - \frac{\widehat{L_s}}{1 - \widehat{Q}_w} \tag{3.6}$$

où $\widehat{Q}_{ij}$ est la fréquence de paires de gènes identiques sur toutes les paires d'individus (i, j) séparées par une distance géographique r ; $\widehat{Q}_i$ (resp. j) est la fréquence de paires de gènes identiques au sein de l'individu i (resp. j) (c'est donc la fréquence de loci homozygotes au sein d'un individu) ; $\widehat{Q}_w$ est la fréquence de paires de gènes identiques sur l'ensemble des individus échantillonnés et $\widehat{L_s}$ le terme constant (par rapport à i, j) identifié par Watts et al. (2007) dans la statistique de Loiselle et al. (1995), qui est calculé comme suit :

$$\widehat{L_s} = \frac{n_p(\mathrm{mean}(\widehat{Q}_{ij})) + n(1/2 + \widehat{Q}_w/2)}{n_p + n} \tag{3.7}$$

où n est le nombre total d'individus échantillonnés et n_p le nombre de paires d'individus.

3.1.4 Statistiques déséquilibre de liaison

Les associations non aléatoires entre allèles à différents loci peuvent être décrites par des paramètres et des statistiques analogues à celles précédemment considérées pour la structuration de la variation à un locus. De façon classique, une association non aléatoire

entre un allèle en fréquence p_i au locus i et un allèle en fréquence p_j au locus j se mesure par $D_{ij} = P_{ij} - p_i p_j$ où P_{ij} est la fréquence d'haplotypes qui portent cet allèle. Dans les modèles de variation neutre, l'espérance de D est nulle et on s'intéresse à sa variance, que l'on peut relier au "déséquilibre d'identité", c'est-à-dire la probabilité que deux haplotypes portent la même combinaison d'allèles aux deux loci, probabilité qui généralise ainsi à deux loci le concept classique de probabilité d'identité à un locus.

Pour réduire les patrons de déséquilibre de liaison à un petit nombre de statistiques résumantes, il serait utile d'avoir des attendus théoriques pour les déséquilibres d'identité en fonction de la distance entre loci sur un chromosome, et de la distance spatiale entre individus. Toutefois, il n'existe pas de tels résultats sous isolement par la distance. Pour définir les statistiques résumantes, il faut s'appuyer sur des résultats plus partiels.

Tout d'abord, il est connu que, dans une population panmictique, l'attendu théorique pour une mesure normalisée de déséquilibre de liaison, en fonction de la probabilité de recombinaison ρ entre deux loci, est une fraction rationnelle en fonction du produit $N\rho$ (Hill and Weir, 1988, p. 76-77). Une approximation par une décroissance exponentielle a paru plus facile à généraliser aux statistiques définies ci-dessous pour prendre en compte la dimension spatiale.

Par ailleurs, divers descripteurs des déséquilibres de liaison ont été proposés pour tenir compte d'une structuration selon le modèle en île. On peut s'inspirer, en particulier, de la statistique de "standardized identity disequilibrium" définie par Vitalis and Couvet (2001), qui peut être vue comme un estimateur de

$$\frac{\Phi_{ij} - Q_{1i}Q_{1j}}{(1 - Q_{2i})(1 - Q_{2j})} \tag{3.8}$$

où Q_{1i} et Q_{1j} représentent les probabilités d'identité aux loci i et j entre haplotypes au sein d'un dème ; Q_{2i} et Q_{2i} représentent les mêmes probabilités mais pour des haplotypes pris dans des dèmes distincts ; et Φ_{ij} est la probabilité d'identité jointe aux loci i et j entre deux haplotypes pris dans un même dème.

Les statistiques de déséquilibre de liaison décrites dans la littérature tiennent compte du fait que l'on observe pas directement des haplotypes, mais des génotypes diploides où, pour les double hétérozygotes, la phase n'est pas connue, et donc l'identité allélique entre des haplotypes de différents individus ne peut être observée. Pour cette raison, on s'intéresse souvent à des statistiques de déséquilibre "génotypique" définies à partir de

fréquences de paires de gènes identiques qui peuvent être observées sans connaître la phase, et à la probabilité d'identité correspondante $\Phi := (\phi + 2\gamma + \delta)/4$. (voir figure 3.1 ; Weir and Cockerham, 1974)

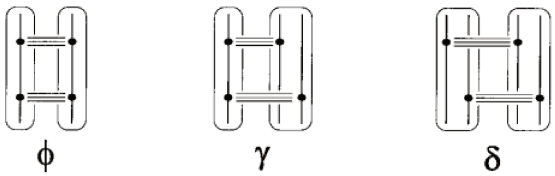


Figure 3.1 – Définitions des probabilités d'identités à deux locus ϕ, γ, δ impliquant respectivement deux, trois ou quatre haplotypes des deux génotypes diploides comparés (Figure 1, Vitalis and Couvet, 2001)

Une mesure de déséquilibre de liaison tenant compte de la distance géographique entre individus peut alors être définie comme suit :

$$\widehat{\eta}_{xy,ij} \equiv \frac{\widehat{\Phi}_{xy,ij} - \widehat{Q}_{1i}\widehat{Q}_{1j}}{(1 - \widehat{Q}_{2i})(1 - \widehat{Q}_{2j})} \tag{3.9}$$

où $\widehat{\Phi}_{xy,ij}$ est la fréquence des paires d'haplotypes, produites par tirage d'un haplotype dans le génotype diploide de chacun de deux individus pris dans les dèmes x et y, qui sont conjointement identiques aux loci i et j ; $\widehat{Q}_{1i}$ (resp. j) est la fréquence d'identité intra-dème observée au locus i (resp. j) et $\widehat{Q}_{2i}$ (resp. j) est la fréquence d'identité observée entre les dèmes au locus i (resp. j), pour les individus échantillonnés dans le dème x et le dème y ; et $\widehat{Q}_{2i}$ (resp. j) est la fréquences d'identité observée entre dèmes au locus i (resp. j) calculée sur tous les individus échantillonnés.

Pour résumer l'information contenue dans les $\widehat{\eta}_{xy,ij}$, on calcule d'abord une estimation unique $\widehat{\eta}_{d_g,d_c}$ pour toutes les paires de loci à distance chromosomique d_c dans des paires d'individus à distance géographique d_g. Cette estimation est la somme des numérateurs des $\widehat{\eta}_{xy,ij}$ concernés, divisée par la somme des dénominateurs des mêmes $\widehat{\eta}$. On ajuste ensuite aux $\widehat{\eta}_{d_g,d_c}$ par moindres carrés un modèle de régression non linéaire de forme

$$\eta_{d_g,d_c} = a + (b - a)\exp\left[-c_g d_g - c_c \rho(d_c)\right] \tag{3.10}$$

où $\rho(d_c)$ est la probabilité de recombinaison entre les loci (la probabilité qu'un gamète pris au hasard porte à ces loci des copies de gènes des deux haplotypes parentaux), déduite de la distance chromosomique (la distance selon une carte chromosomique, déduite en sommant les distances de recombinaison d_c entre loci adjacents), ici en inversant la fonction

de Haldane (1919) : $\rho(d_c) = [1 - \exp(-2d_c)]/2$. Il y a donc quatre paramètres a, b, c_g, c_c qui constituent les statistiques résumantes des patrons de déséquilibre de liaison dans l'échantillon.

Ce modèle de régression n'a pas de justification théorique précise, mais sa forme parait raisonnable au vu de simulations préliminaires visant à identifier un modèle de régression approprié. On avait d'abord essayé d'ajuster des modèles pour une variable dérivée de $\widehat{\eta}_{d_g,d_c}$ en y remplaçant les fréquences monolocus intradème par des fréquences monolocus à distance d_g. Mais il a paru difficile de trouver un bon modèle statistique pour cette variable, et également difficile d'ajuster les modèles, en raison de problèmes de convergence des algorithmes disponibles pour l'ajustement par l'approche des moindres carrés non linéaires.

3.2 Temps de calcul et optimisation

L'une des difficultés lorsque l'on cherche à calculer des statistiques liées au déséquilibre de liaison vient de leur complexité algorithmique. Par exemple, si la complexité algorithmique du calcul de la fréquence d'identité entre paires de gènes monolocus pour l'ensemble des locus des individus de la population est d'ordre $O(n^2.l)$ avec n le nombre d'individus de l'échantillon et l le nombre de locus (disons 12 s pour un jeu de données de 1000 locus sur n_x individus), alors la complexité algorithmique de la probabilité d'identité jointe entre paires de gènes pour deux locus est de $O(n^2.l^2)$ (soit plus de 3h pour les mêmes 1000 locus sur n_x individus). Cette explosion des temps de calcul peut devenir très pénalisante lorsqu'il s'agit de calculer de telles statistiques sur des jeux de données contenant des milliers voire des dizaines de milliers de marqueurs.

Exemple d'optimisation

Dans certains cas, le calcul des statistiques $\widehat{\eta}_{d_g,d_c}$ peut être effectué de façon particulièrement efficace en utilisant une représentation binaire de chaque haplotype. En particulier, un calcul spécifique a été implémenté dans `GSumStat` pour les jeux de données de SNPs uniformément répartis le long des haplotypes à étudier.

Pour illustrer cet algorithme, considérons 2 haplotypes phasés $A = \mathbf{0110}$; $B = \mathbf{0100}$ de 4 marqueurs répartis aux emplacements 1, 2, 3 et 4. Ici la fréquence d'identité jointe entre paires de gènes pour tout les locus à distance 1 : $\frac{1}{3}$, à distance 2 : $\frac{1}{2}$ et à distance 3 : $\frac{1}{1}$.

L'approche naïve consiste à comparer deux à deux l'ensemble des gènes (complexité de $O(l^2)$ comme décrit dans la section précédente). L'on compare ainsi 2 paires de gènes (une par locus) pour 6 paires distinctes non ordonnées de loci, soit un total de 12 opérations.

Une autre approche consiste à utiliser des opérations binaires sur ces haplotypes qui peuvent être vus comme des vecteurs de 0 et de 1. Ainsi, si on applique les opérations OUX[1] et NON[2] sur les vecteurs binaires représentant A et B (A OUX NON(B) = C) on obtient un vecteur C ayant pour valeur 0 aux loci pour lesquels les deux allèles sont différents dans A et B et 1 pour les loci pour lesquels les deux allèles sont identiques ($\mathbf{0110}$ OUX NON($\mathbf{0100}$) = $\mathbf{1101}$).

Une fois ce vecteur C obtenu, pour obtenir les fréquences d'identité jointes entre loci à une certaine distance d (en unité d'intervalle entre marqueurs adjacents) il est possible de le comparer (avec ET[3]) à une copie de lui-même décalée de d emplacements ($\ll_d$[4]) de d. Ainsi, on aligne toutes les paires de gènes à une distance d et on les compare (C ET C$\ll_d$ = D). Ne reste plus qu'à compter le nombre de 1 dans le vecteur D pour obtenir la fréquence d'identité jointe à distance d. Par exemple pour $d = 1$: $\mathbf{1101}$ ET $\mathbf{1101}$ $\ll_1$ = $\mathbf{1101}$ ET $\mathbf{1010}$ = $\mathbf{1000}$), et donc la fréquence d'identité jointe entre paires de gènes pour tous les locus à distance 1 est $\frac{1}{3}$.

Le calcul pour l'ensemble des distances possibles nécessite $1+3*(l-1) = 10$ opérations binaires, avec une complexité réduite à $O(l)$. Une élaboration de cette approche permet de prendre en compte des données non phasées ou des données réparties de manière hétérogène sur le chromosome pour un coût supplémentaire faible.

Ceci n'est qu'un parmi de nombreux exemples d'optimisation présents dans `GSumStat`.

1. 1 OUX 1 = 0 ; 1 OUX 0 = 1 ; 0 OUX 1 = 1 ; 0 OUX 0 = 0
2. $NON(1) = 0$; $NON(0) = 1$
3. 1 ET 1 = 1 ; 1 ET 0 = 0 ; 0 ET 1 = 0 ; 0 ET 0 = 0
4. $\ll_d$ permet de déplacer les valeurs du vecteur de d emplacements à droite (les d dernières emplacements du vecteur sont remplies avec des 0). $\mathbf{0001} \ll_2 = \mathbf{0100}$

4

Le pipeline d'inférence `gspace2infr`

`gspace2infr` est une librairie R implémentant un pipeline pour l'inférence par simulation, partant de la spécification des paramètres de simulation aux résumés de l'inférence statistique (estimations ponctuelles, intervalles de confiance, graphiques...). Pour cela, elle interface et relie les outils dédiés aux différentes étapes de l'inférence par simulation, notamment les programmes de simulation d'échantillons génétiques `IBDsim` et `GSpace`, le calcul de statistiques résumantes par la librairie `GSumStat`, et l'inférence statistique proprement dite par la méthode *summary likelihood* présentée dans l'introduction de cette thèse, et implémentée dans la librairie R `Infusion`. Elle inclut aussi des procédures de post-traitement de sorties de `GSumStat`. Notamment, elle effectue le calcul de la régression-non linéaire des statistiques $\widehat{\eta}_{xy,ij}$ de déséquilibre de liaison, décrites dans le chapitre précédent. Enfin, la libraire `gspace2infr` permet de faire des "post-analyses diagnostiques" d'une grand nombre d'inférences faites sur des données simulées à paramètres constants pour évaluer ces procédures d'inférence. Ces procédures d'évaluation seront décrites dans le chapitre suivant.

Bien que l'ensemble des fonctionnalités prévues de `gspace2infr` ne soient pas encore implémentées à ce jour, cette librairie permet déjà de tester `GSpace` dans un cadre inférentiel, de vérifier l'information que l'on peut extraire des données au moyen des statistiques résumantes et d'étudier la performance des inférences par `Infusion` pour l'ensemble des paramètres de scénarios d'isolement par la distance, et l'ensemble des statistiques résumantes décrites dans le chapitre précédent.

Ce chapitre présente un exemple d'inférence par *summary likelihood* utilisant le pipeline implémenté dans `gspace2infr`, afin de donner une idée plus concrète des méthodes utilisées et de la nature des résultats produits.

4.1 Le scénario biologique simulé

Nous allons ici considérer une population de taille fixe dans le temps d'un organisme à cycle de vie majoritairement diploïde, hermaphrodite sans auto-fécondation. Les individus sont répartis par couples sur les noeuds d'un réseau bidimensionnel. Chaque couple ne se reproduit qu'une seule fois (semelparité). Les descendants migrent selon un modèle d'isolement par la distance au sein de la population.

La simulation est effectuée dans `GSpace` en considérant une grille homogène carrée de dimension 70×70. Sur chaque noeud de cette grille on place un unique couple d'individus diploïdes. A chaque génération, ce couple se reproduit, et les descendants diploïdes peuvent migrer. La probabilité qu'un descendant donné émigre est 0.25 et la distribution axiale de distance de dispersion suit une loi géométrique de paramètre $g = 0.558615$ et de distance maximum de 20 dèmes (voir section 2.3.3). Si plus de deux individus sont présents sur le même dème après la phase de migration, un couple est tiré au hasard et occupe le noeud du réseau à la génération suivante.

L'échantillon est constitué de l'ensemble des couples d'une zone centrale carrée de l'habitat. Pour chaque individu, les génotypes de 500 loci bialléliques sont simulés. Ces loci sont répartis sur 10 chromosomes, à raison de 50 loci par chromosome, à intervalles réguliers en termes de probabilité de recombinaison. Cette probabilité est de 10^{-5} entre chaque locus pour chaque gamète formé. La probabilité de mutation par locus et par gamète est 5.10^{-5}.

Ce scénario permet de vérifier l'apport de trois fonctionnalités introduites dans `GSpace` et n'existant pas dans `IBDsim` :

— la présence de marqueurs liés sur plusieurs chromosomes ;
— la migration des individus. Les généalogies de chaque locus d'un échantillon simulé ne sont ainsi pas simulées indépendamment, même pour ceux portés par différents chromosomes, car elles sont toutes conditionnelles à la généalogie des individus dans la population d'où l'échantillon est tiré ;
— la reproduction sexuée, sans auto-fécondation.

4.2 Inférence

Sur cet échantillon simulé, on va utiliser `Infusion` pour inférer les quatre paramètres suivants : la taille de la population (plus précisément le nombre n_x de noeuds sur un axe du réseau, modélisant le nombre de dèmes sur ce même axe), la probabilité μ de mutation par locus par génération, la probabilité m d'émigration de chaque individu et le paramètre g de forme de la distribution géométrique de dispersion (voir tableau 4.1). Les autres paramètres de la simulation sont supposés connus (voir tableau 4.2).

		Bornes d'estimation	
Paramètre	Valeur	minimum	maximum
n_x	70	10	300
μ	5.10^{-5}	5.10^{-7}	5.10^{-2}
m	0.25	0.01	0.99
g	0.558615	0.01	0.99

Table 4.1 – Paramètres variables dans les simulations

On voit donc que l'inférence est en principe très différente de celle d'une inférence sous les approximations usuelles de la génétique des populations, de produits tels que $\theta = 2N_{\mathrm{T}}\mu$ en fonction de la taille totale de population N_{T} ; ou le nombre $2Nm$ de migrants dans le modèles en île ; ou encore le produit $D\sigma^2$ sous isolement par la distance (où D est la

Paramètre	Valeur
Dist disp max	20
Indiv par dème (N)	2
Ploïdie	2
Nb dèmes échant	10x10
Indiv échant par dème	2
Nb chr par indiv	10
Nb marq par chr	50
Modèle mutation	KAM, $K = 2$
ρ	5.10^{-5}

Table 4.2 – Paramètres fixés dans les simulations

densité de la population, exprimée dans la même unité de distance que les paramètres de dispersion g et σ^2, ce qui fait que $D = N$ individus diploïdes par élément de surface correspondant à un noeud du réseau dans notre exemple, voir section 3.1.3).

De fait, rien ne permet de dire a priori que les données contiennent de l'information permettant d'estimer les quatre paramètres n_x, μ, m et g (appelés *paramètres canoniques* par la suite). Cela dépend en particulier de ce que l'on suppose fixé et connu (par exemple, ici, la présence d'un couple par noeud du réseau). S'il n'y a pas d'information pour un paramètre, l'inférence devrait nous l'indiquer sous la forme d'un profil de vraisemblance (résumée) plat pour ce paramètre.

Pour cette raison, nous nous intéresserons aussi à des *paramètres composites*, c'est-à-dire à des fonctions d'au moins deux paramètres canoniques, ici $\theta = 2N_{\mathrm{T}}\mu = 2n_x^2 N\mu$ et σ^2 ($\sigma^2 = m\frac{(1+g)}{(1-g)^2}$ pour une distribution de dispersion géométrique), pour lesquels les données peuvent contenir plus d'informations que pour les paramètres canoniques du modèle.

Inférence par *summary likelihood*

La *summary likelihood* est une méthode d'inférence consistant à approximer la surface de vraisemblance des statistiques résumantes calculées sur des données génomiques (génotypes multilocus) simulées (voir figure 4.1) par rapport aux paramètres du modèle,

puis à utiliser cette surface de "vraisemblance résumée" afin d'effectuer l'inférence des paramètres d'une population à partir d'un ensemble de génotypes échantillonnés au sein de cette dernière.

Dans le cas de notre scénario biologique simulé, l'inférence des *paramètres canoniques* se déroule de cette manière :

Construction de la table de référence

La construction de la table de référence s'effectue d'abord en réalisant des tirages des valeurs des 4 paramètres d'intérêt à estimer. Ce tirage s'effectue de manière indépendante pour chaque paramètre dans des distributions uniformes (ou ici log(uniforme) si l'on veut explorer l'espace d'un paramètre sur une échelle logarithmique) entre deux extrêmes (min et max). Dans notre exemple, nous allons commencer par construire une table de référence de 200 simulations en explorant l'espace des paramètres selon les intervalles résumés dans le tableau **??**. En effet, pour chaque vecteur (n_x, μ, m et g) de paramètres du processus biologique, il est possible de simuler des génotypes multilocus et d'en extraire un certain nombre de statistiques résumantes (voir section 3.1 et figure 4.1b). Dans le modèle ici présenté, les données simulées pour les faibles valeurs de n_x et μ ne montrent souvent pas assez de polymorphisme pour permettre de calculer des statistiques telles que a_r, e_r ou η, dont les dénominateurs tendent vers zéro quand la diversité génétique diminue. Ces lignes de la table de référence sont ignorées dans la suite de l'inférence.

Réduction du nombre de statistiques résumantes

A cette étape va être créée une statistique synthétique pour chacun des paramètres à inférer. Cette étape est appelée *projection*.

Cette statistique synthétique peut être construite par différentes méthodes. L'on cherche ici à construire un prédicteur de chaque paramètre à partir des statistiques résumantes. La méthode utilisée pour cela est l'apprentissage de la relation entre valeurs de paramètres et statistiques résumantes, à partir de la table de référence. La méthode d'apprentissage utilisée par défaut dans `Infusion` est la régression non paramétrique par forêts aléatoires, implémentée dans la librairie R `ranger` (Wright and Ziegler, 2017). On exploite donc ici comme dans la méthode ABC-RF la facilité d'utilisation et l'efficacité de cette méthode

```
30 30 , 002002 001001 002002 001001 002002 001001 001001 001001 001001 002002 002002 001001 001001 002002 002002
001001 002002 002002 002002 002001 002002 002002 001001 001001 002001 002002 002002 002002 002002 001002 002002
001001 002002 002002 001001 001001 002001 001001 002002 001001 002002 002002 002002 002002 002002 001001 001001
001001 002002 001001 002001 001001 002002 001002 001001 001002 001001 002001 001001 001002 002002 001001 001001
001001 001002 002002 002002 002001 001002 002001 001002 002002 001002 001001 001002 002002 001001 002002 002002
001001 002002 001002 002002 002002 001001 001001 002002 001001 001001 002002 002002 002002 002002 001001 002002
002002 001001 002002 001001 001001 002002 001001 001001 002002 001001 001001 001001 001001 001002 002002 002002
002001 001001 002001 002001 001002 002001 002002 001001 001001 002002 002002 001001 001001 002002 001001 002002
002002 002001 002002 002002 002002 001001 001001 002001 002002 002002 002002 002001 001001 001001 002002 001001
002002 002002 001002 002001 001001 002001 001001 001001 002002 001002 002002 001001 002002 002001 002002 001002
001002 001001 002002 002002 001001 001001 002002 001001 002002 002001 002002 002002 002002 001001 002002 001001
002002 002002 001002 001002 002002 001001 001001 001001 001001 001001 002002 002002 002001 001001 002002 002002
002002 001001 001001 001001 001002 002001 001002 002002 001001 001001 002002 002002 001001 001001 002002 001002
002002 001001 002001 001002 002001 001002 002002 001001 001001 002002 001001 002002 002002 002001 001001 001001
002002 001002 001001 001001 002002 001001 001001 001002 001002 001001 002002 001001 001001 002002 001002 002002
001001 001002 002002 002002 001001 002001 002002 001001 002001 001002 001001 002002 001002 002002 002001 002002
001001 002002 001001 002002 002002 001001 001001 002002
pop
30 31 , 002002 001001 001002 001001 002002 001001 002002 001001 001001 001002 002002 001001 002001 002002 002002
001001 001001 002002 002002 002002 002002 002002 001001 001002 002002 002002 002002 002002 002002 002002 002002
001001 002002 001002 001001 002001 002002 001001 002002 002002 002002 002002 002002 001002 002002 001001 001001
001001 002002 001001 002002 001001 001002 001001 001001 001001 001001 001001 001001 001001 002002 001001 001001
002002 001001 002002 002002 001001 002002 001002 002002 002002 002002 001002 002001 002002 001001 002002 002002
002001 002002 002002 002002 001002 001001 001001 001001 001001 001001 002002 002002 002002 001001 002001 002002
002002 001001 002002 001001 001001 002002 001001 001001 002002 001001 001001 001001 001001 001001 002002 002002
002002 001001 001002 001002 002001 001002 002002 001001 001001 002002 002002 001001 001001 002002 002001 001002
002002 001002 002002 002002 002002 001001 001001 002002 002002 002001 002002 002002 002001 002001 002002 002001
002002 002001 001002 002001 001002 002002 001002 001001 002002 001001 002001 001002 002002 002002 002002 001002
001002 001001 001001 002002 001001 001001 002002 001001 002001 001001 002002 002002 002002 001001 002002 001001
002002 002001 001001 001001 001002 001001 001001 001001 002002 001001 002002 002002 001001 001001 001001 002002
001001 002001 001001 001001 001001 002002 001001 002002 001001 002002 001001 001002 001001 002002 002001 002002
002002 001001 002001 001001 002001 002001 001001 002001 002002 002002 002001 001002 002002 001001 002002 002002
002002 001002 001001 001001 001001 002002 001001 001001 001001 002002 002001 002002 002002 001001 001001 001001
002002 001002 001001 001001 002002 001001 001001 002002 002002 001001 002002 002001 001001 002002 001002 002002
001002 002001 002001 002002 001001 001002 001001 001002 002002 001002 001001 002001 001002 001002 002002 001002
002001 002002 002001 001002 002002 001001 001001 002001
```

(a) Ici est représenté un sous-ensemble de 239 loci présents sur 2 individus diploïdes. Ces loci sont associés à une carte génétique permettant de déterminer leurs emplacements sur les différents chromosomes. Données au format `Genepop` (voir la documentation de `Genepop` pour plus de détails.)

Hobs_mean	Hobs_var	Hexp_mean	Hexp_var	Nb_allele_mean	Nb_allele_var
Nb_allele_deme_mean	Nb_allele_deme_var	Var_mean	Var_var	MGW_mean	
MGW_var	Fis	Fis_var	Fst	Fst_var	Qwi_mean	Qwi_var
Qbiwd_mean	Qbiwd_var	Qbd_mean	Qbd_var	Qbi_mean	Qbi_var
Q0_mean	Q0_var	Q1_mean	Q1_var	Q2_mean	Q2_var	Q3_mean	Q3_var
Q4_mean	Q4_var	Q5_mean	Q5_var	Q6_mean	Q6_var	Q7_mean	Q7_var
Q8_mean	Q8_var	Q9_mean	Q9_var	lin_Fst_slope	lin_Fst_intercept	Ar_slope
Ar_intercept	Er_slope	Er_intercept				
0.21814	0.0344645	0.221588	0.0345337	1.868	0.114806	1.3874
0.00170792	0.110794	0.00863343	1	0	-0.0157289	0.0152212
0.0310105	0.00444794	0.78186	0.0344645	0.78626	0.0328614
0.778364	0.0345454	0.778403	0.0345354	0.78626	0.0328614
0.781226	0.0336443	0.778376	0.0345558	0.778273	0.034669
0.778315	0.0345736	0.778164	0.0346257	0.777618	0.0346928
0.778686	0.0348823	0.778691	0.0355179	0.785808	0.0353395
0.00554291	0.0374106	0.00612601	0.00671319	0.00417407	-0.00613945

(b) Ensemble des statistiques résumantes calculées à partir des génotypes simulés et qui seront utilisées lors de l'inférence.

Figure 4.1 – Génotypes générés par `GSpace` et leur représentation par des statistiques résumantes (voir section 4.1)

à faire le tri entre les statistiques porteuses d'informations sur le paramètre à estimer et celles qui n'apportent que peu ou pas d'information. La statistique synthétique qui résulte de cette projection est ainsi censée résumer avec peu de perte l'information disponible dans l'ensemble des statistiques résumantes.

Chaque projection se fait en pratique en deux étapes. La première étape est l'apprentissage, sur la table de référence, du prédicteur du paramètre considéré à partir des statistiques résumantes qu'elle contient. Dans un deuxième temps, l'on utilise le résultat de cet apprentissage pour produire la prédiction pour chaque ligne du tableau de référence (il s'agit ici des prédictions "out-of-bag" [1] pour éviter un surajustement des prédictions aux données d'entraînement), et pour le jeu de données que l'on cherche à analyser. Pour ce dernier, même si on l'a simulé ici sous des valeurs de paramètres connues, l'on se place dans la position d'un utilisateur de la méthode : les valeurs de paramètres ne sont pas connues, le jeu de données n'est donc pas utilisé pour l'apprentissage et la valeur prédite de chaque paramètre n'est donc pas "out-of-bag".

L'information conservée dans chaque statistique synthétique peut être visualisée par la dispersion de la relation entre valeurs de paramètres et valeurs prédites, qui donne une première idée de la capacité à inférer chaque paramètre avec précision. La figure 4.2 présente ces relations pour les données de la table de référence (toujours en utilisant les prédictions *out-of-bag*). L'on y voit par exemple que m devrait être estimé assez précisément, et inversement que n_x ne devrait pas l'être.

A l'issue de cette étape, l'on a donc réduit les nombreuses statistiques résumantes *brutes* à quatre statistiques résumantes *projetées*, une par paramètre estimé. Ce sont uniquement ces dernières qui sont utilisées à l'étape suivante de l'analyse.

1. Le bagging utilise le sous-échantillonnage avec remplacement pour créer des échantillons d'apprentissage. L'erreur "out-of-bag" est l'erreur de prédiction moyenne sur chaque échantillon d'apprentissage x_i, en utilisant uniquement les arbres non construits à partir de x_i.

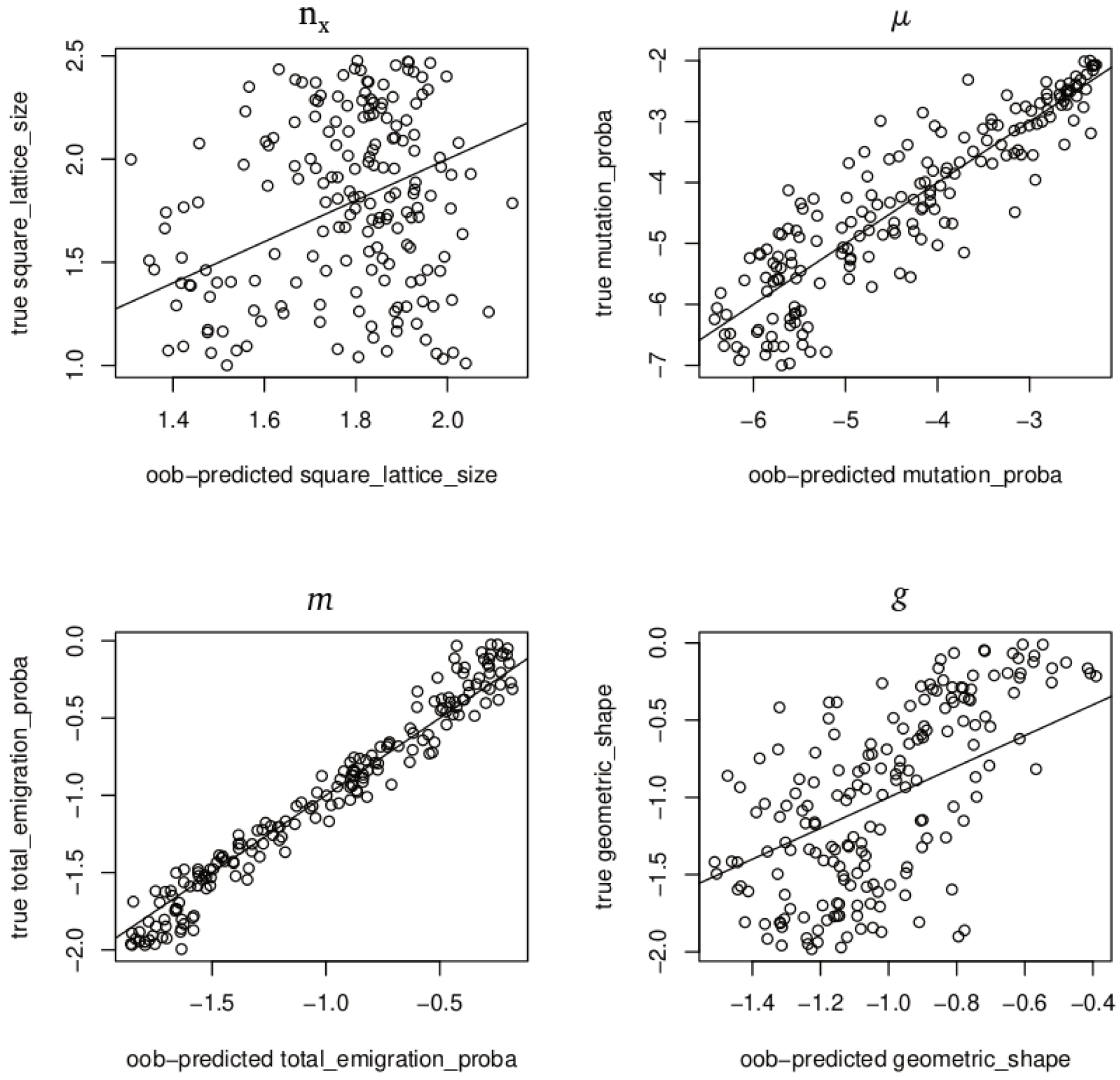


Figure 4.2 – Valeurs prédites des paramètres (de haut gauche à bas droite : n_x, μ, m et g) par régression non paramétrique par forêts aléatoires. La droite est la diagonale $y = x$.

Estimation de la surface de vraisemblance résumée

Cette étape consiste à estimer, à partir du tableau de référence, la distribution jointe des paramètres et des statistiques projetées produite à l'étape précédente, et d'en déduire une estimation de la surface de vraisemblance pour le jeu de données analysé réduit à ses statistiques résumantes projetées.

La distribution jointe (en dimension 8 dans cet exemple) est ici inférée comme un mélange de gaussiennes multivariées (en dimension 8, donc). La librairie R `Rmixmod` (Lebret et al., 2015) est utilisée pour ajuster ces modèles de mélange pour une gamme de valeurs candidates du nombre d'éléments gaussiens, et l'on sélectionne le nombre d'éléments par comparaison des AICs de ces différents ajustements.

A partir de cette distribution jointe estimée, l'on peut déduire la densité estimée des "données" (les statistiques projetées) pour chaque valeur du vecteur de 4 paramètres par application de la formule de Bayes, en divisant la densité jointe par la densité marginale des paramètres déduite de cette densité jointe (toute ces opérations sont effectuées par la librairie `Infusion` et n'ont pas fait l'objet d'une implémentation spécifique pour ce travail de thèse).

Cette densité estimée des données pour chaque valeur du vecteur de paramètres, vue comme une fonction des paramètres, constitue une estimation de la surface de vraisemblance des paramètres sachant les statistiques résumantes projetées calculées sur l'échantillon analysé.

Inférences à partir de la surface de vraisemblance résumée

Une fois la surface de vraisemblance estimée, son maximum est déterminé, et des intervalles de confiance à un seuil prédéfini par l'utilisateur sont construits par rapport de vraisemblance ou rapport de profil de vraisemblance. `Infusion` peut aussi effectuer une forme de bootstrap pour estimer l'incertitude sur ces différents résultats, et produire différentes représentations de la surface de vraisemblance, notamment des représentations des profils de vraisemblance de chaque paramètre (voir figure 4.3) et des rapports de profils de vraisemblance de paires de paramètres (voir figure 4.4).

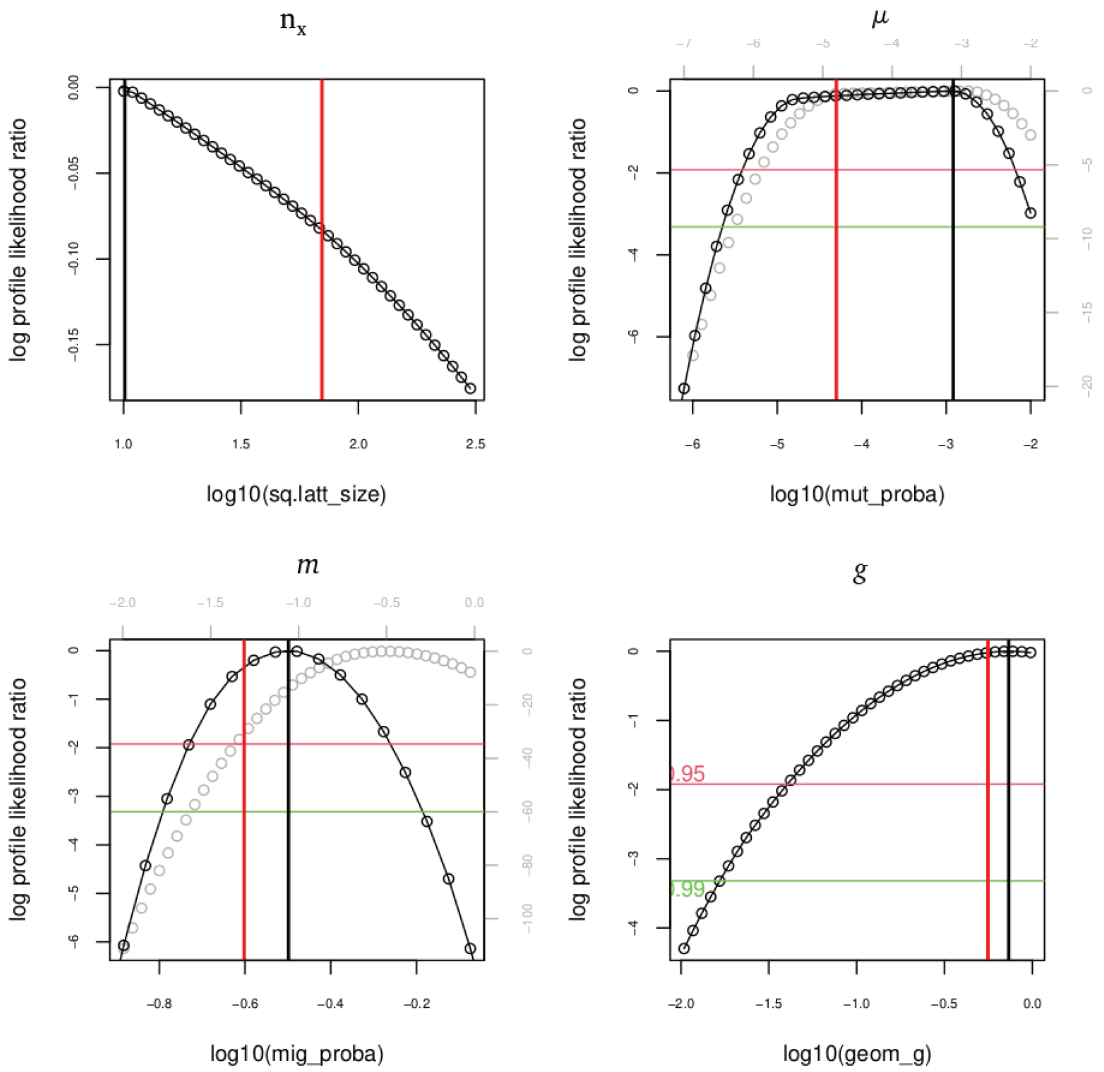
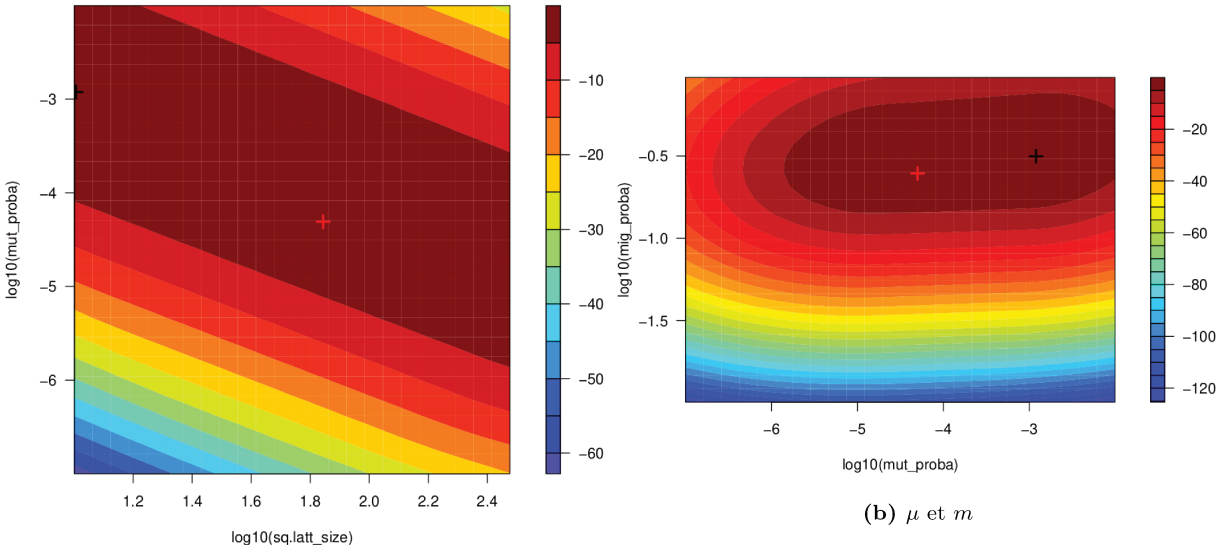


Figure 4.3 – Profils de vraisemblance résumée estimés pour chacun des paramètres à inférer (de haut gauche à bas droite : n_x, μ, m et g). Les points gris et l'échelle grise correspondante représentent l'ensemble de l'intervalle exploré. Les points noirs et l'échelle noire correspondante représentent un zoom sur la zone autour du maximum de vraisemblance estimé. Les traits rouges et verts représentent (quand ils sont calculables) respectivement les intervalles de confiance à 95% et à 90%. Ils ne s'appliquent qu'à la zone de maximum de vraisemblance. Le trait rouge vertical représente la valeur réelle de chaque paramètre utilisée pour simuler le jeu de données, et le trait noir la valeur estimée par maximisation de cette vraisemblance.



(a) n_x et μ

(b) μ et m

(c) m et g

(d) n_x et g

Figure 4.4 – Logarithme du rapport de profil de vraisemblance pour des paires de paramètres. La croix rouge représente les valeurs réelles de la paire de paramètres utilisée pour simuler le jeu de données, et la croix noire les valeurs estimées par maximisation de la vraisemblance.

On peut remarquer ici que le profil de vraisemblance est très plat pour le paramètre n_x contrairement à celui de m qui est beaucoup plus piqué. Ceci est cohérent avec les figures diagnostiques des projections.

Affinage de l'estimation de la surface de vraisemblance résumée

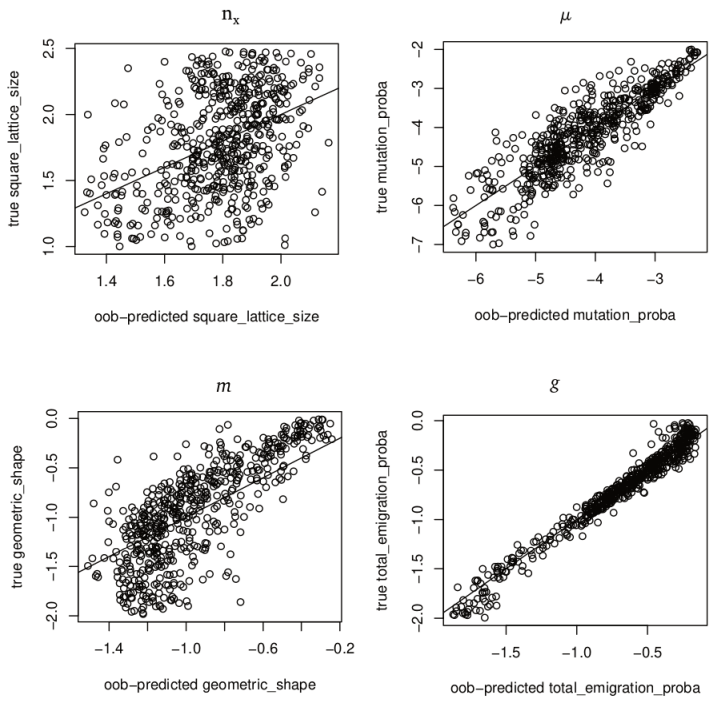
Par la suite, il est possible de relancer les étapes précédentes en tirant de nouvelles valeurs de paramètres des valeurs de paramètres dans des zones d'intérêt, et en simulant de nouveau le processus biologique pour ces nouvelles valeurs de paramètres (voir figure 4.5). Ces zones d'intérêt se trouvent notamment à proximité du maximum de vraisemblance, aux abords des bornes des intervalles de confiance, ou dans des zones où la vraisemblance est mal estimée. En pratique, `Infusion` échantillonne uniformément dans l'espace de paramètres défini par une valeur seuil minimum de la vraisemblance, définie par rapport au maximum inféré de la vraisemblance, ce seuil étant conçu pour inclure les bornes d'intervalles de confiance. La densité d'échantillonnage diminue ensuite progressivement pour les valeurs de vraisemblance inférées plus basses. Il s'ensuit que la densité simulée des paramètres évolue au cours de l'analyse et est uniquement définie dans un but d'estimation précise de la région haute de la surface de vraisemblance et non par référence à d'autres critères, tels que ceux qui peuvent être discutés pour l'élicitation de distributions *prior*.

Ces affinages successifs permettent :

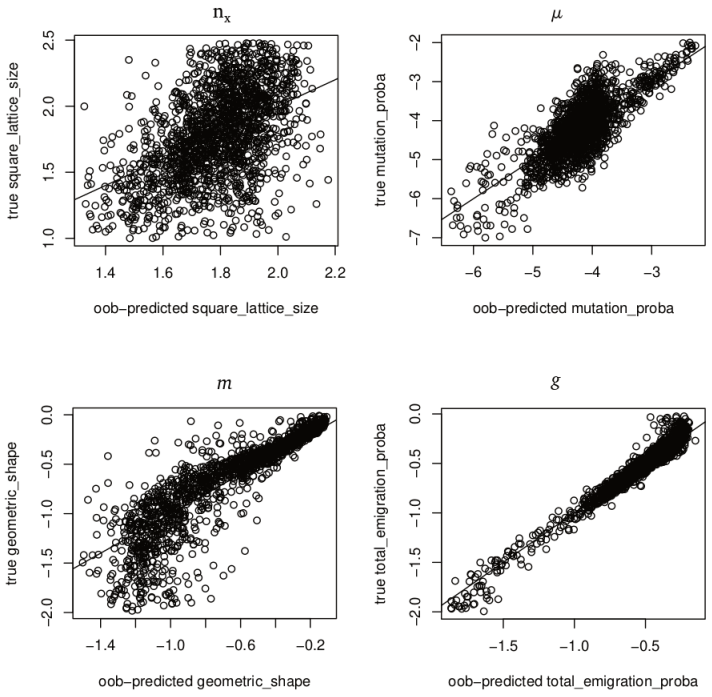
— d'entraîner la méthode de projection avec un plus grand nombre de données la rendant potentiellement plus précise, particulièrement dans la région de paramètres la plus intéressante (voir figure 4.5) ;

— de mieux estimer la surface de vraisemblance et donc de potentiellement améliorer les estimations des paramètres. La figure 4.6 montre par exemple une tendance à l'amélioration des estimations ponctuelles entre la 3e et la 8e itération.

Au fur et à mesure des itérations d'affinage, l'estimation de la surface de vraisemblance s'améliorant, l'on observe que les profils de vraisemblance deviennent plus piqués même si le profil de n_x reste beaucoup plus plat que celui de m.

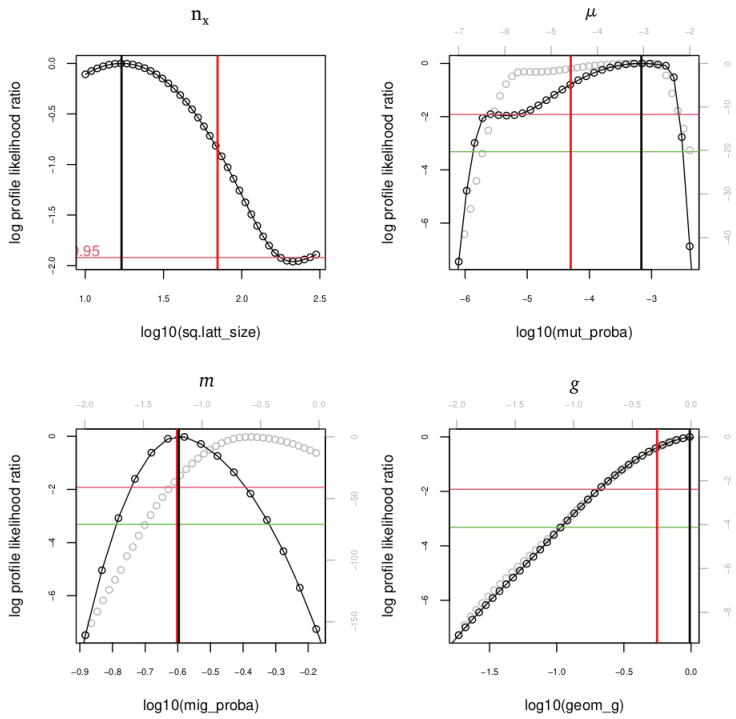


(a) 3ème affinage

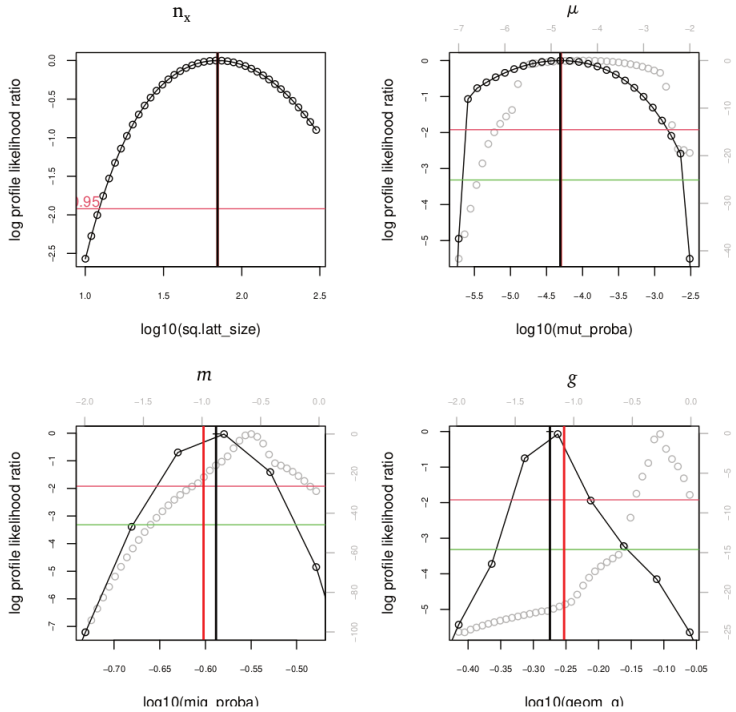


(b) 8ème affinage

Figure 4.5 – Valeurs prédites des paramètres par régression non paramétrique par forêts aléatoires en fonction des itérations d'affinage. La droite est la diagonale $y = x$.



(a) 3ème affinage



(b) 8ème affinage

Figure 4.6 – Profils de vraisemblance résumée à la 3e et la 8e itération d'affinage.

Inférence de paramètres composites

L'inférence précédente peut aussi s'effectuer sur la même table de référence en combinant des paramètres canoniques et des paramètres composites (à nombre total de paramètres constant). On peut par exemple considérer le paramètre composite classique θ, et le paramètre de dispersion σ^2 (ici équivalent à considérer le produit $D\sigma^2$ car la densité de population est connue et fixée à $D = 2$ individus diploïdes par élément de surface correspondant à un noeud du réseau).

Comme le montre le profil de vraisemblance bidimensionnel pour la taille de l'habitat et la probabilité de mutation (figure 4.7), le produit $\theta = 2n_x^2 N\mu$ (constant le long de diagonales descendantes sur cette figure en échelle des logarithmes des paramètres) est plus facilement estimable que chaque paramètre indépendamment. La crête de la surface de vraisemblance, visible le long d'une telle diagonale, se traduit en effet par des profils plats pour chaque valeur des paramètres du produit, alors qu'en se déplaçant orthogonalement à cette crête, on obtiendrait un profil bien plus variable pour le produit.

Les intervalles de confiance pour les paramètres composites sont calculés par rapport de vraisemblance comme pour les paramètres non composites.

Dans le cadre de `gspace2infr`, toutes les analyses présentées dans ce chapitre sont effectuées de façon automatisée sous le contrôle d'arguments d'une fonction `infusion_inference()`. En les appliquant à un grand nombre d'échantillons simulés, on va pouvoir évaluer la performance des méthodes d'inférence en fonction de la nature des données simulées et des statistiques résumantes utilisées.

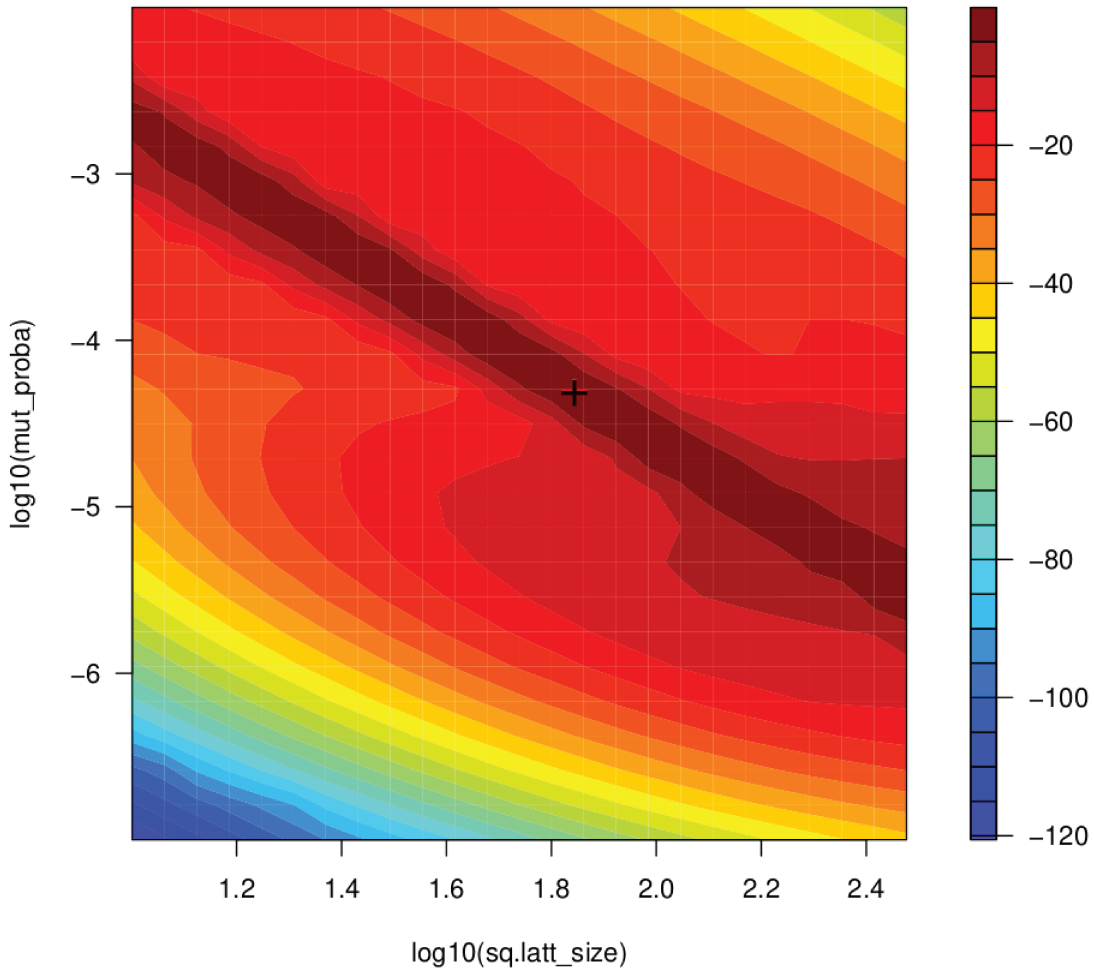


Figure 4.7 – Logarithme du rapport de profil de vraisemblance pour les paramètres n_x et μ à la 8ème itération d'affinage.

5

Performance des inférences

5.1 Test de performance des outils d'inférences

Comme nous l'avons vu au chapitre précédent, l'estimateur $\hat{\phi}$ d'un paramètre ϕ se base sur le maximum de vraisemblance (MLE pour "Maximum Likelihood Estimate") de la surface de vraisemblance estimée des paramètres, sachant les statistiques résumantes projetées calculées sur l'échantillon de génotypes analysé.

Pour évaluer la performance de la méthode on l'évalue sur un certain nombre n de jeux de données simulés pour des valeurs de paramètres connues. On peut alors évaluer des mesures classiques (biais, variance, erreur carrée moyenne...) de la performance de l'estimateur $\hat{\phi}$, sachant la valeur ϕ_0 du paramètre ϕ. Ici ces mesures sont calculées sur les estimations relatives $\hat{\phi}/\phi_0$:

— le biais des estimations relatives
— (l'estimation sans biais de) la variance des estimations relatives

— la moyenne de l'erreur absolue des estimations relatives (NMAE pour "Normalized Mean Absolute Error")

De plus, il est possible d'effectuer un test sur le rapport des vraisemblances (LRT pour "Likelihood-Ratio Test") de $\hat{\phi}$ et ϕ_0. La théorie stipule que sous l'hypothèse H_0 : $\phi = \phi_0$ la statistique utilisée par le LRT (qui est fonction du logarithme du rapport des vraisemblances) suit approximativement une loi de χ^2 à 1 degré de liberté lorsqu'il y a suffisamment d'informations sur le paramètre ϕ dans les données (Severini 2000).

Pour vérifier cela, nous représentons la fonction de distribution cumulative empirique (ECDF pour *Empirical Cumulative Distribution Function*) des p-values déduits sous l'hypothèse de la distribution du $\chi 2$ (voir figure 5.1) pour l'ensemble des n échantillons simulés. Cette distribution cumulative doit être proche de la diagonale 1 :1. Une déviation par rapport à cette distribution uniforme peut se produire si la vraisemblance est mal estimée ou si la statistique du test ne suit pas la distribution χ^2 attendue (car il n'y a pas assez d'informations sur le paramètre dans les données). Un test de Kolmogorov-Smirnov sur les n p-values permet de quantifier l'écart de leur distribution avec la distribution uniforme attendue. Inversement si la distribution du des p-values est uniforme, les intervalles de confiance seront corrects (contiendront la valeur de paramètre ϕ_0 avec la probabilité attendue).

Le calcul des "summary LRT" est implémenté dans `Infusion`. La librairie `gspace2infr` automatise la synthèse des résultats obtenus sur chaque jeu de données analysé et utilise des fonctions standard de R pour les analyses telles que le test de Kolmogorov-Smirnov.

Nous présentons ici deux résultats liés à l'analyse préliminaire de la fiabilité et de la performance de nos outils d'inférence. Le premier est une tentative de quantifier l'information portée par le déséquilibre de liaison. Le second est l'analyse de l'impact de la non-indépendance des arbres de coalescence des marqueurs non liés.

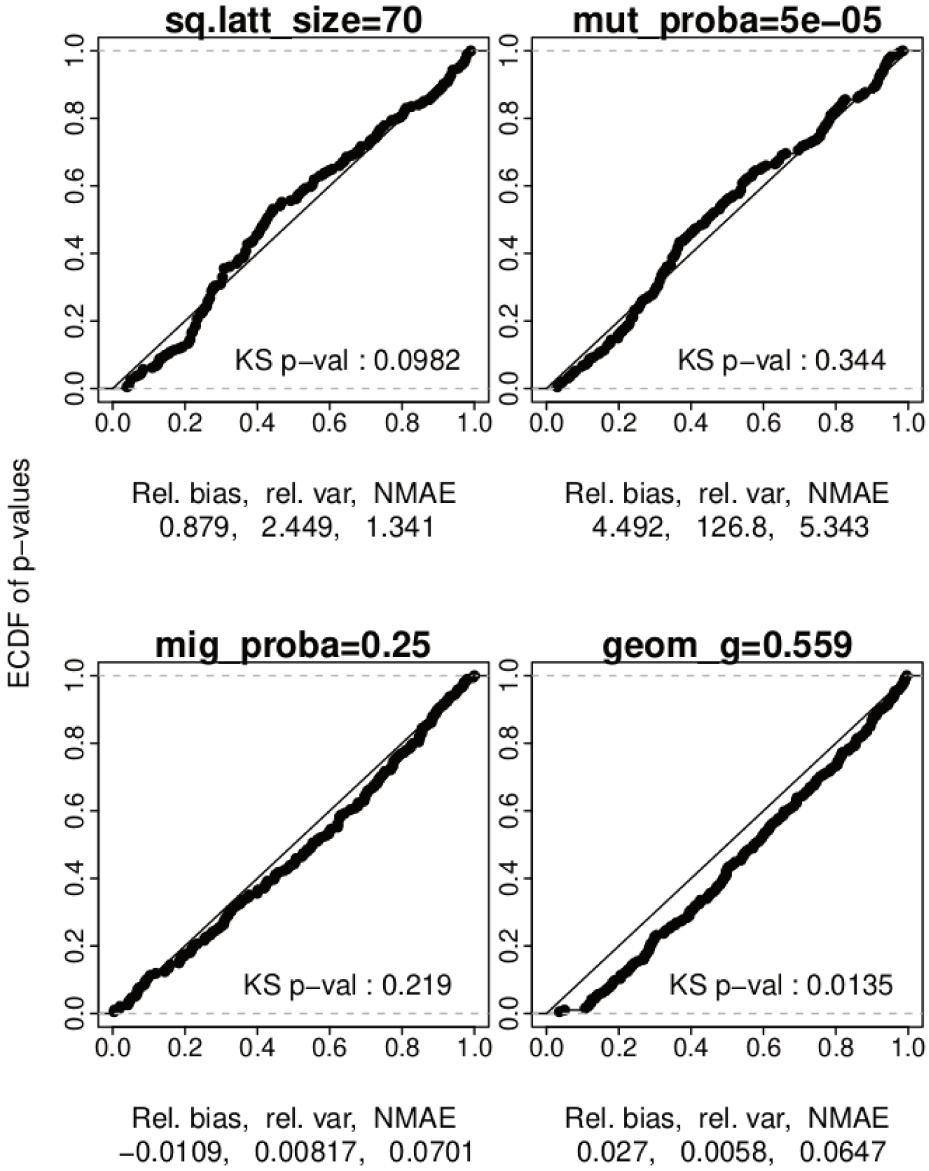


Figure 5.1 – Distribution des p-values pour chacun des paramètres estimés (μ, m, θ et σ^2). Sont par ailleurs représentés pour chaque paramètres la p-value du test de Kolmogorov-Smirnov (KS p-value), le biais relatif (Rel. biais), la variance relative (rel. var) et le NMAE.

5.2 Information apportée par le déséquilibre de liaison entre marqueurs liés

Pour trouver les conditions où les déséquilibres de liaison entre marqueurs peuvent apporter de l'information concernant les paramètres du modèle, on peut considérer les probabilités relatives de coalescence et de recombinaison dans un graphe ancestral de recombinaison.

La probabilité que les arbres de gènes de deux marqueurs deviennent indépendants en remontant dans le temps augmente avec la distance chromosomique séparant ces marqueurs (en augmentant la probabilité de survenue d'un évènement de recombinaison entre eux). Dans cette perspective, les motifs dus au déséquilibre de liaison de marqueurs éloignés renseigneraient presque exclusivement sur des évènements de coalescence récents (dans le graphe ancestral de recombinaison) là où les motifs dus au déséquilibre de liaison de marqueurs proches porteraient de l'information sur des événements plus anciens (Hayes et al., 2003 ; Al-Asadi et al., 2019).

La distance chromosomique entre les marqueurs est fonction de la probabilité de recombinaison entre marqueurs adjacents par génération (ρ, voir section 3.1.4). Ainsi, une manière d'obtenir l'information la plus complète possible concernant les différentes périodes du graphe ancestral de recombinaison serait de choisir la position des marqueurs de sorte qu'une partie des paires de marqueurs aient une forte probabilité de rester liés avant de coalescer et inversement, que cette probabilité soit faible pour une autre partie des marqueurs. Pour cela, on choisit conjointement le nombre de marqueurs par chromosome et un ρ "optimal".

En connaissant approximativement le temps de coalescence moyen entre deux lignées ($\approx N_{\mathrm{T}}$ avec N_{T} le nombre de copie de gène dans la population, voir équation 2.3), la distance entre les marqueurs (en nombre de marqueurs dans l'intervalle) et ρ il est possible de trouver le nombre moyen d'évènements de recombinaison pour toutes les paires de marqueurs.

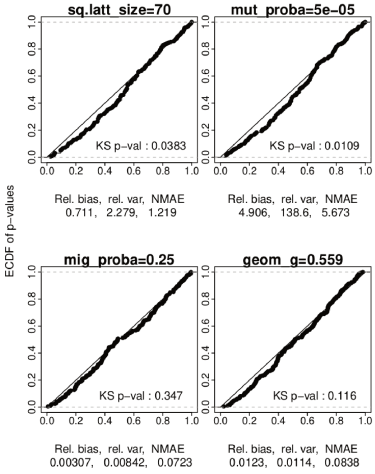
Pour vérifier cette idée, on va comparer l'information apportée par les déséquilibres de liaison dans ce cas, par rapport aux cas où les marqueurs sont systématiquement plus liés, ou systématiquement moins liés que selon la valeur "optimale" de ρ précédemment

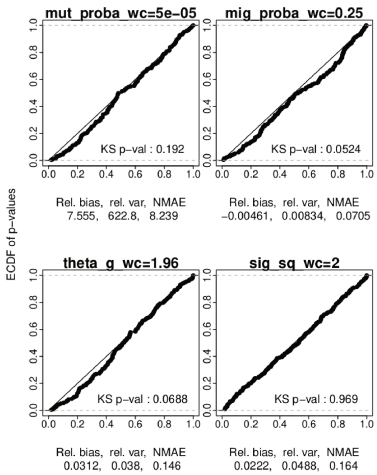
identifiée. Cette information peut être mesurée soit en inférant les paramètres (canoniques ou composites) uniquement à partir des statistiques de déséquilibre de liaison, soit en comparant les précisions des inférences, avec ou sans ces statistiques, mais incluant les autres statistiques présentées au chapitre 3.

Dans les tests de performance suivants, on suppose le scénario biologique présenté à la section 4.1. Pour chaque évaluation de performance, $n = 200$ jeux de données furent simulés et leurs paramètres inférés. L'inférence de ces paramètres fut effectuée avec `Infusion` sur une table de référence d'environ 1800 lignes générées lors de 7 affinages successifs. Les valeurs de paramètres choisies ainsi que les bornes de l'espace des paramètres exploré pour l'estimation sont résumées dans le tableau **??**.

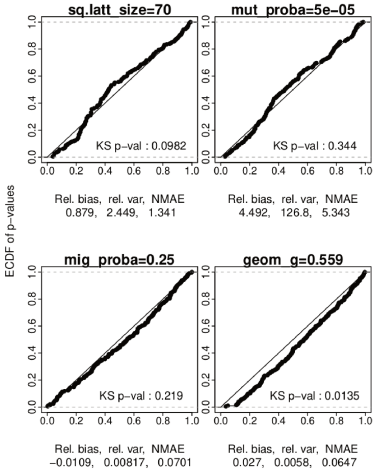
Dans ce scénario, $N_{\mathrm{T}} = 19600$ individus diploïdes et 50 marqueurs sont considérés par chromosome. $\rho = 10^{-5}$ semble alors être un choix intéressant, le nombre d'évènements de recombinaison pour chaque évènement de coalescence variant entre $[0.196; 9.6]$ selon la distance entre marqueurs. Les temps de calcul de l'inférence par échantillon analysé variant de 8 heures (avec $\rho = 10^{-7}$) à 48 heures (avec $\rho = 10^{-2}$), soit 1600 à 9600 heures de calcul pour 200 échantillons, seules quelques conditions de simulation furent comparées pour limiter l'impact écologique de l'étude. Ici seront présentés trois de ces tests de performance (avec respectivement $\rho = 10^{-7}$, $\rho = 10^{-5}$ et $\rho = 10^{-3}$). Dans chaque cas, la performance est présentée pour l'estimation conjointe des quatre paramètres canoniques (colonne de gauche de la figure 5.2), et pour l'estimation conjointe de deux paramètres composites (θ et σ^2, colonne de droite de la même figure) et deux paramètres canoniques.

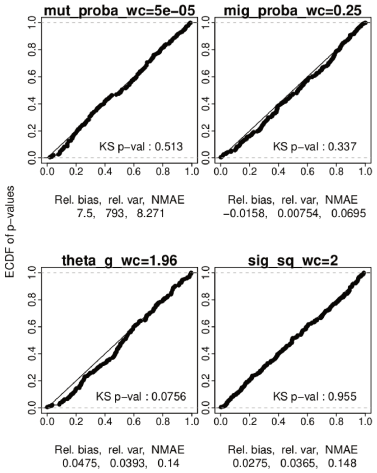
$\rho = 10^{-7}$



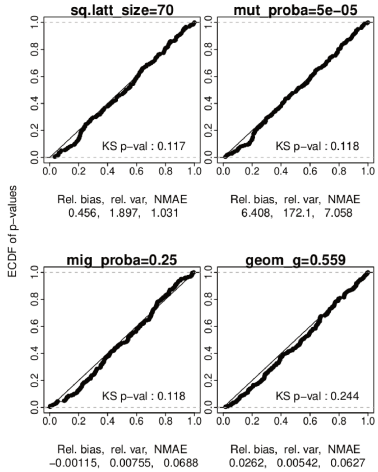


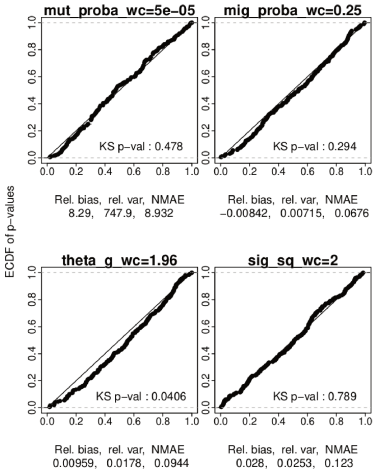
$\rho = 10^{-5}$





$\rho = 10^{-3}$



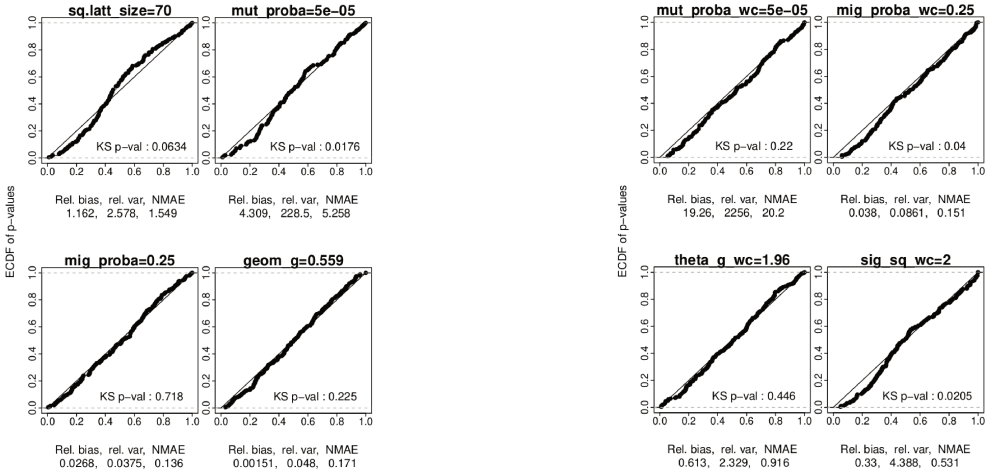


(a) Paramètres canoniques uniquement

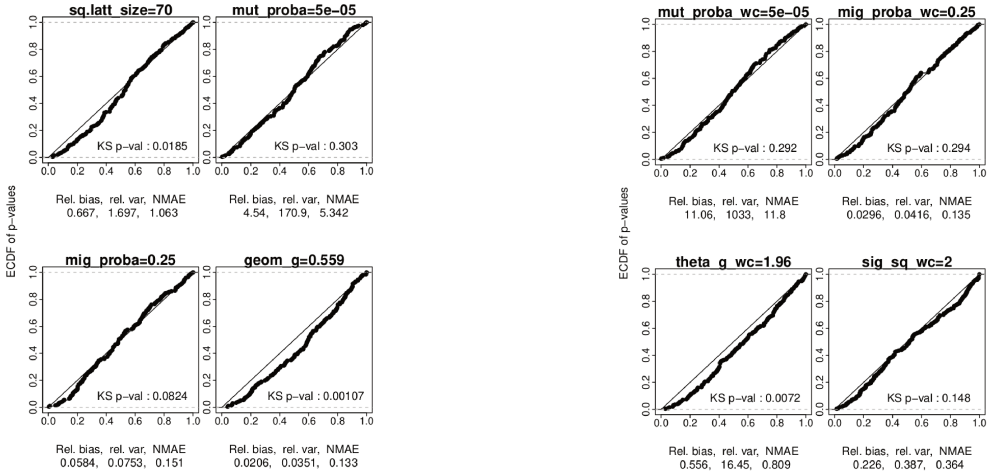
(b) Paramètres canoniques et composites

Figure 5.2 – Résultats des tests de performances sur des inférences effectuées avec l'ensemble des statistiques résumantes calculables par `GSumStat`(voir chapitre 3).

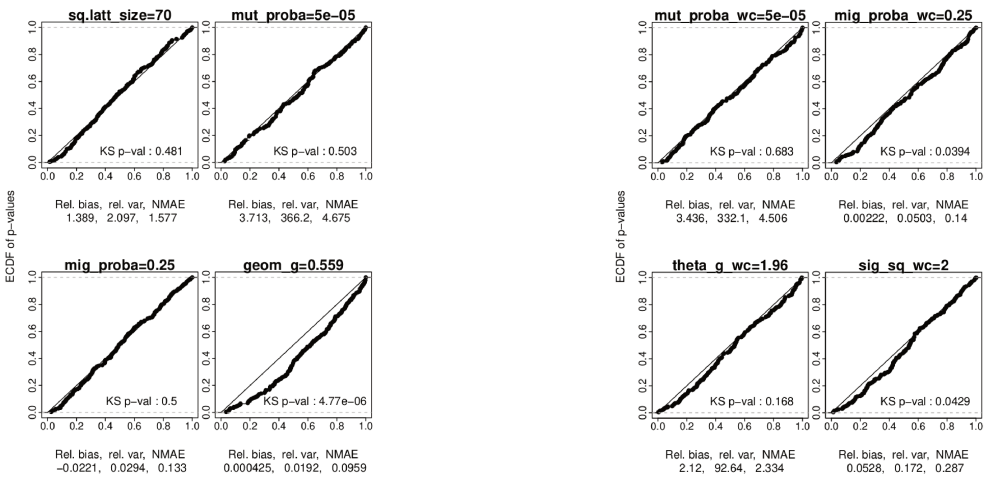
$\rho = 10^{-7}$



$\rho = 10^{-5}$



$\rho = 10^{-3}$

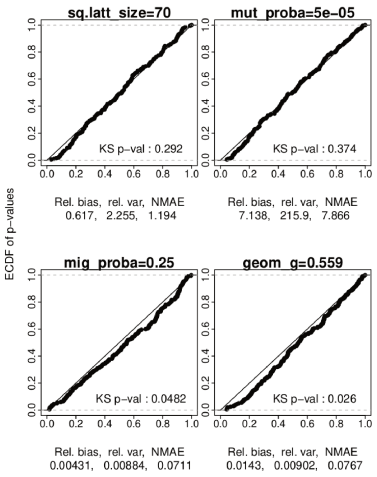


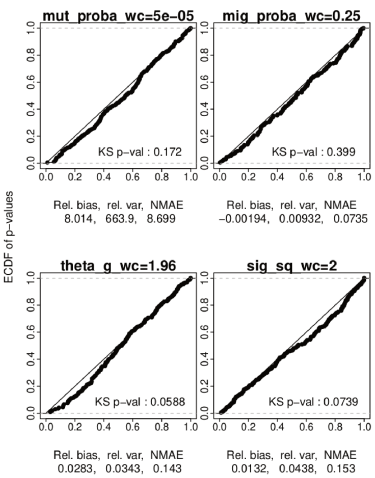
(a) Paramètres canoniques uniquement

(b) Paramètres canoniques et composites

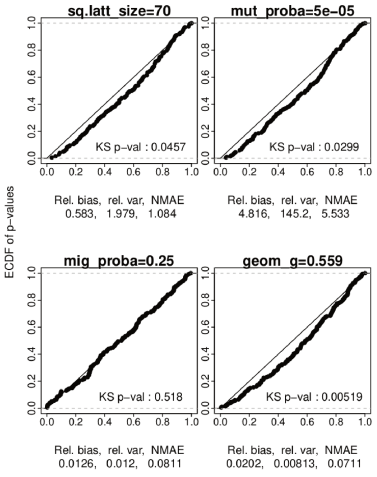
Figure 5.3 – Résultats des tests de performances sur des inférences effectuées avec uniquement les quatre statistiques résumantes qui décrivent les variations de η.

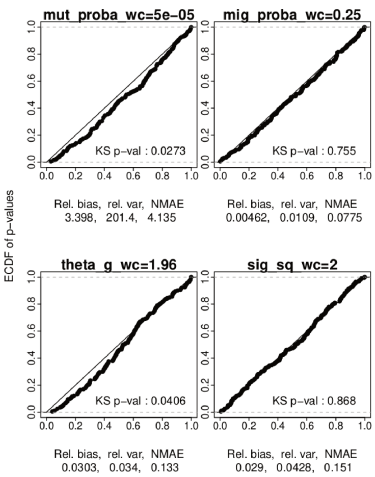
$\rho = 10^{-7}$



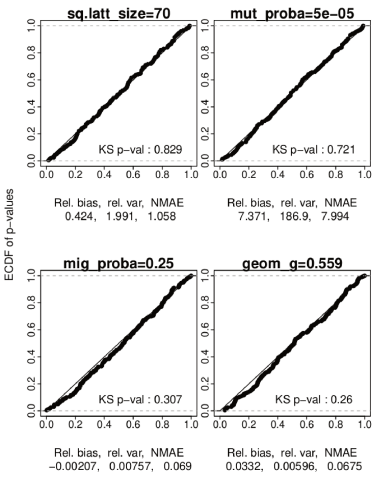


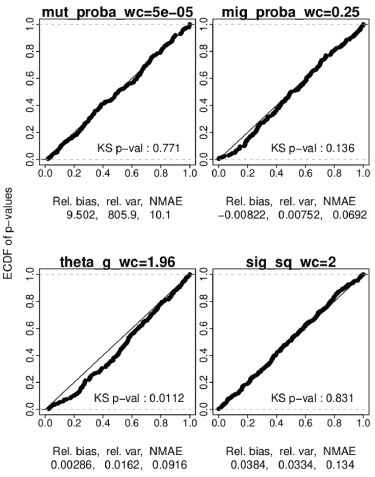
$\rho = 10^{-5}$





$\rho = 10^{-3}$





(a) Paramètres canoniques uniquement **(b)** Paramètres canoniques et composites

Figure 5.4 – Résultats des tests de performances sur des inférences effectuées avec l'ensemble des statistiques résumantes calculables par `GSumStat` mais sans les quatre statistiques résumantes qui décrivent les variations de η.

Comme nous pouvons le voir sur les différents tests de performance (figure 5.2), la variation de ρ ne semble pas affecter sensiblement la performance des outils d'inférence. L'analyse des résultats des tests de performance sur des inférences effectuées avec uniquement la statistique η (figure 5.3), ou plus précisément les quatre statistiques résumantes décrivant les variations de η dans un jeu de données, et sans elles (figure 5.4) nous permet d'observer que :

— si en effet la statistique η semble apporter de l'information sur les paramètres de dispersion (m, g et par extension σ^2) dans ce modèle, cette information semble être déjà portée par d'autres statistiques spatiales de manière bien plus efficace (le biais, la variance et le NMAE de l'estimation de ces paramètres est systématiquement plus faible dans les inférences utilisant toutes les statistiques).

— la variation de ρ n'affecte pas du tout l'information apporté par les η de la manière attendue. Les estimations dans le cas $\rho = 10^{-5}$ "optimal" n'apparaissent pas être plus précises que dans les autres cas.

Au delà de cette comparaison, il convient de souligner que les distributions des p-values des LRT sont approximativement uniformes, ce qui montre qu'il est possible de calculer des intervalles de confiance corrects par "rapport de vraisemblance résumée" pour les différents paramètres, et ceci, sur la base de tables de référence d'environ 1800 lignes pour l'inférence de quatre paramètres. Bien que la précision des tests de performance serait sûrement améliorée par l'utilisation de plus grandes tables de référence, ce résultat est à souligner car les inférences par simulation sont généralement très coûteuses en terme de temps de simulation. Par exemple, Pudlo et al. (2016) recommandent de simuler entre 5000 et 10000 échantillons par modèle pour la comparaison de modèles.

5.3 Information apportée par le déséquilibre de liaison entre marqueurs non-liés

Le déséquilibre de liaison peut s'observer même lorsque que l'on considère des paires de marqueurs non liés (c'est à dire sur des chromosomes différents). En effet il existe une corrélation entre les arbres de coalescence de tous les marqueurs du fait qu'ils sont tous containts par l'arbre généalogique réalisé des individus de la population (pour cette raison, on parle souvent de "déséquilibre gamétique" plutôt que de "déséquilibre de liaison").

Ce déséquilibre gamétique se retrouve dans les données simulées par `GSpace` car il simule des génomes individuels (voir section 2.3.2) mais pas dans celles simulées par `IBDsim` car ce dernier indépendamment chaque locus. On peut donc tester l'impact de l'absence de déséquilibre gamétique dans les données simulées part `IBDsim` sur l'inférence des paramètres en comparant les performances des inférences dans un même modèle mais dans un cas en utilisant `GSpace` pour simuler les données et dans un autre cas en utilisant `IBDsim`.

Le scénario démographique commun utilisé dans cette comparaison est assez proche du précédent, mais `IBDsim` étant relativement lent et ne pouvant simuler des marqueurs liés, le hypothèses concernant les marqueurs ont été modifiées afin de simuler des locus microsatellites, moins nombreux mais plus informatifs car plus polymorphes. Le modèle de mutation utilisé à cette fin est le modèle de mutation par pas (pour *strict stepwise mutation model* ou SMM, Ohta and Kimura 1973), où chaque évènement de mutation ajoute ou supprime une unité répétée à l'allèle muté dans une gamme de valeurs comprises entre des valeurs minimale et maximal. Un seul microsatellite est simulé par chromosome (voir tableau 5.2).

		Bornes d'estimation	
Paramètre	Valeur	minimum	maximum
n_x	70	10	300
μ	5.10^{-4}	1.10^{-6}	2.10^{-1}
m	0.25	0.01	0.99
g	0.558615	0.01	0.99

Table 5.1 – Paramètres variables dans les simulations du modèle utilisé pour la comparaison de `GSpace` et `IBDsim`.

A ce stade du développement de `gspace2infr`, et comme évoqué précédemment, les jeux de statistiques résumantes calculées sur les données issues des deux simulateurs diffèrent car elles sont calculées par des algorithmes différents : `GSumStat` calcule toutes les statistiques décrites dans la section 3 pour les données issues de `GSpace` alors qu'`IBDsim` calcule un plus petit nombre de statistiques sur ses propres données simulées. Afin de ne pas introduire un facteur confondant dans la quantification de l'information apportées

Paramètre	Valeur
Dist disp max	10
Indiv par dème (N)	2
Ploïdie	2
Nb dèmes échant	10x10
Indiv échant par dème	2
Nb chr par indiv	20
Nb marq par chr	1
Modèle mutation	SMM, 200 allèles
ρ	NA

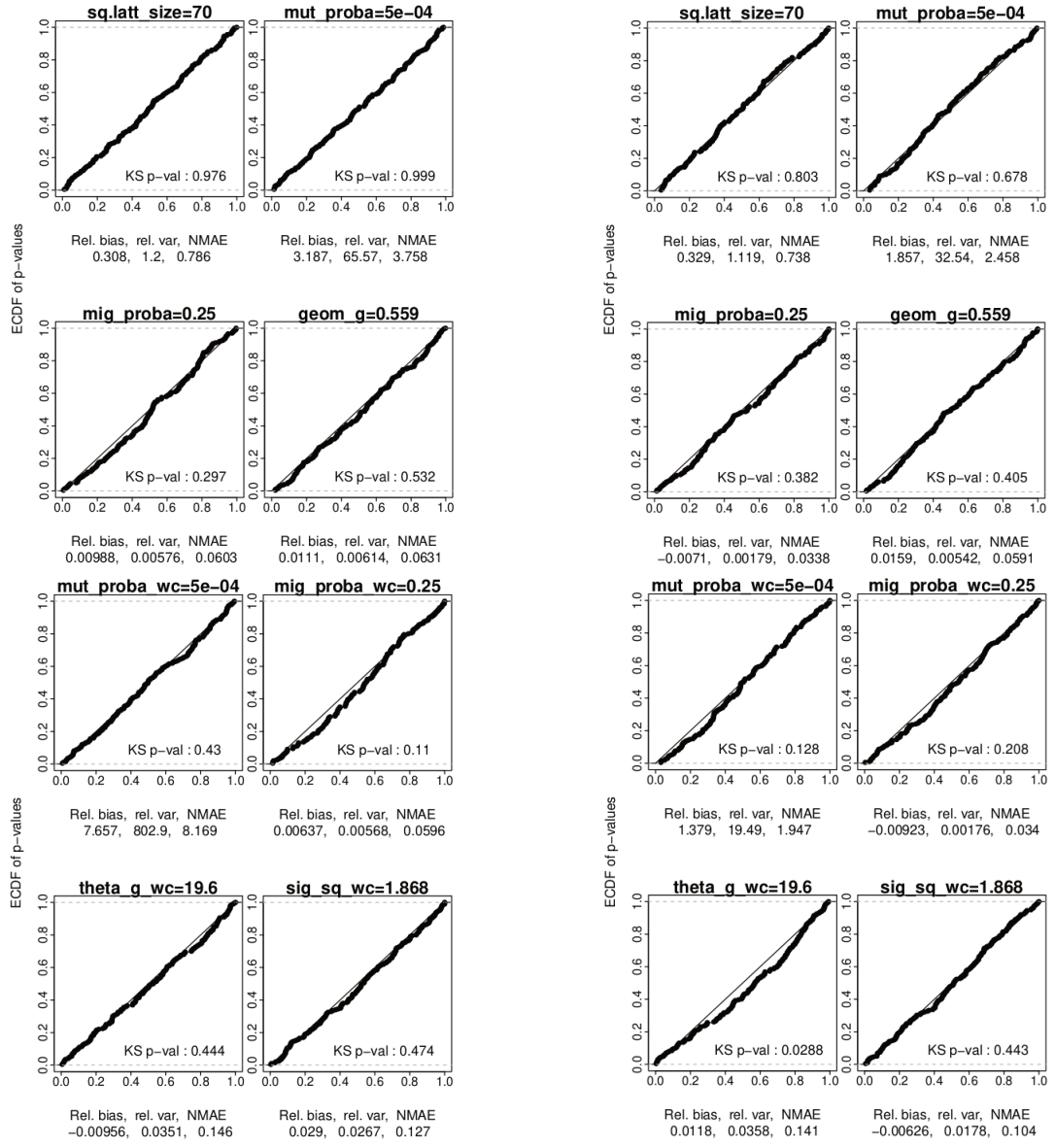
Table 5.2 – Paramètres fixés dans les simulations du modèle utilisé pour la comparaison de `GSpace` et `IBDsim`.

par les déséquilibres entre marqueurs non liés, on a réduit les statistiques calculées au plus grand ensemble de statistiques communes calculable à la fois par `GSumStat` et par `IBDsim`.

De plus, pour éviter un autre facteur confondant, on a aussi simulé sous `GSpace` la migration en phase haploïde, soit une migration indépendante pour les lignées situées sur des chromosomes homologues.

Pour vérifier la pertinence des différences de performances observables selon les simulateurs (voir figure5.5), l'on a comparé les distributions des erreurs quadratiques $\hat{\phi} - \phi_0)^2)$ des estimateurs dans les deux séries d'inférences (respectivement pour `GSpace` et `IBDsim` de taille 200 et 199), issues des deux tests de performances sur le même modèle (voir tableau 5.3).

Les comparaisons des tests de performances entre `GSpace` et `IBDsim` semble montrer qu'il existe bien des différences. Au jour du rendu du manuscrit définitif et après de nombreuses analyses la raison de cette différence est inconnue.



(a) `GSpace` **(b)** `IBDsim`

Figure 5.5 – Résultats des tests de performances des inférences des paramètres dans le cadre de la comparaison entre `GSpace` et `IBDsim`"0307

Paramètre	p-valeur à 5%
n_x	0.3533
μ	0.3023
m	$> 7.10^{-10}$
g	0.3686

(a) 4 paramètres canoniques

Paramètre	p-valeur à 5%
μ	0.3749
m	$> 3.10^{-9}$
θ	0.4459
σ^2	0.0798

(b) 2 paramètres canoniques et 2 paramètres composites

Table 5.3 – Résultats des tests non paramétriques de Wilcoxon-Mann-Whitney (Wilcoxon 1945 ; Mann and Whitney 1947) comparant les distributions des erreurs carrées pour les paramètres inférés dans le cadre de la comparaison entre `GSpace` et `IBDsim`"0307

6

Discussion et Conclusion

Apport de nos outils dans un contexte d'inférence de paramètre démographiques sous des modèles IBD

Malgré le peu de résultats obtenus pendant cette thèse concernant la performance de nos outils, il semble important de souligner les résultats suivants :

— Dans nos tests de performance, il ressort que l'estimation de m et g se fait de manière assez (voir même très) précise pour tous les scénarios explorés, malgré le petit nombre de marqueurs considérés. L'inférence de ces deux paramètres de dispersion semble ainsi pouvoir être effectuée de manière indépendante de la taille de la population ou de sa structure, et ce sans a priori sur ces dernières.

— Les estimations des paramètres composites tel que θ sont elles aussi assez précises avec de faibles biais et des variances réduites quel que soit le modèle.

— Ces inférences par simulation se déroulent sur des tables de référence de taille très réduite. Il est donc possible d'inférer des paramètres dans des modèles complexes tout en conservant une bonne estimation d'un certain nombre de paramètres.

Tout cela laisse présager des résultats intéressants pour les analyses ultérieures, bien que deux points concernant le modèle avec lesquels les résultats précédents ont été obtenus nécessitent d'être discutés.

Premièrement, ce modèle considère la densité de la population par unité de distance en maille de la grille ($D = N_{ind}$ avec N_{ind} le nombre d'individus par dème) comme fixé. Dans un modèle spatial où l'on sait classiquement estimer $D\sigma^2$, fixer l'un des deux paramètres aura tendance à faciliter l'estimation de l'autre (et donc des statistiques le composant). Relâcher cette hypothèse simplificatrice nécessiterait de considérer un modèle où la densité est un paramètre à estimer (ce modèle ne fut pas construit au cours de cette thèse par manque de temps).

Deuxièmement, l'utilisation d'une distribution de dispersion aussi contrainte que la loi géométrique n'est pas entièrement satisfaisante. En effet, la question se pose de la signification de la valeur de g si la distribution de dispersion des individus de la population suit un autre type de loi (typiquement une loi dont la kurtosis est plus élevée). Une idée pour pallier ce biais potentiel serait d'utiliser une distribution de Sichel pour modéliser la dispersion (plus versatile, voir section 2.3.3) et de chercher à inférer les paramètres de cette dernière (a, b, p). Ce point a cependant son contrecoup puisque cela signifierait inférer deux paramètres supplémentaires, alourdissant d'autant plus nos tests de performances.

Un canevas d'inférence prometteur pour de futurs développements et tests de performance

L'apport principal de cette thèse consiste en la mise en place et le test d'un ensemble d'outils (`GSpace`, `GSumStat` et `gspace2infr`) permettant d'effectuer des inférences par simulation. Ces outils ont pour avantage d'être flexibles dans le choix et la paramétrisation des modèles, des statistiques résumantes et le couplage avec d'autres méthodes d'inférence comme par exemple ABC-RF.

En ce qui concerne les modèles et leur paramétrisation, les limitations des outils utilisés sont celles des simulateurs utilisables (pour l'instant `GSpace` et `IBDsim`), à savoir :

1. être capable de simuler le modèle. Il est actuellement possible de travailler sur n'importe quel modèle d'isolement par la distance en réseau, homogène dans l'espace et constant au cours du temps (`GSpace` dans sa version actuelle) ou potentiellement variable dans le temps et hétérogène dans l'espace (`IBDsim`). Les prochaines étapes du développement de `GSpace` devraient permettre de simuler sur des réseaux avec de l'hétérogénéité spatiale et temporelle. De plus, des apports substantiels ont été apportés à `GSpace` en parallèle de cette thèse dans le cadre de l'utilisation de modèles en espace continu (combinant simulation *forward* et coalescence, collaboration avec I. Bonnici).

2. être capables de simuler les données en un temps raisonnable. Les méthodes d'inférence par simulation nécessitant au minimum un nombre de simulation de l'ordre de quelques milliers (pour des modèles comportant jusqu'à 6 paramètres à inférer) et sans doute des dizaines de milliers pour des modèles avec plus de 10 paramètres à inférer, il est pour le moment difficile d'effectuer des inférences sur des modèles dont la simulation prend en moyenne plus que quelques dizaines de minutes. Cela n'est pas tant un problème pour l'analyse d'un jeu de données (facilement parallélisable par `Infusion`) mais reste un problème pour évaluer la performance d'une méthode sur un grand nombre de jeux de données tout en restant relativement sobre sur les ressources utilisées. Une optimisation des procédures de gestion des individus et de dispersion de `GSpace` sera notamment nécessaire pour simuler de plus gros volumes de données génomiques et/ou des réseaux hétérogènes.

L'implémentation et le calcul des statistiques résumantes (via `GSumStat`) étant indépendants des simulateurs utilisés, ces derniers sont théoriquement interchangeables[1] (dans la limite de leurs capacités à simuler sous les mêmes modèles), sans pour autant que cela ait un impact sur l'analyse. Ce point a pour effet émergent une facilitation du test, de la comparaison et de la validation de différents simulateurs, même dans le cas où il n'existe aucun attendu théorique calculable sur les modèles simulés.

1. À l'heure où ces lignes sont écrites, les procédures permettant une réelle indépendance entre les simulateurs et le calcul des statistiques résumantes ne sont pas complètement implémentées (mais le seront très bientôt).

De plus `GSumStat` étant optimisé pour travailler avec de grandes quantités de marqueurs, ce canevas est une base solide pour permettre le passage à l'échelle de l'inférence par simulation dans le cadre de données génomiques.

Dans cette thèse, le temps a manqué pour implémenter dans `gspace2infr` d'autres méthodes d'inférence par simulation qu'`Infusion`. Cependant, une grande part du travail conceptuel permettant l'intégration de méthodes *Approximate Bayesian Computation* (utilisation d'ABC-RF à la place d'`Infusion`) a déjà été effectuée et l'intégration de cette méthode ne devrait pas poser de problème majeur. De plus, en fournissant un cadre standardisé et robuste, `gspace2infr` permettra de facilement comparer la performance des méthodes d'inférence par simulation, actuelles ou ou à venir, et de tester la pertinence de nouvelles statistiques résumantes.

Information portée par les patrons de déséquilibre de liaison

Ce canevas d'inférence a permis de mesurer l'information portée par les quatre statistiques résumant les patrons de déséquilibre de liaison en fonction de la distance chromosomique et géographique, ainsi que leur capacité à estimer un ou plusieurs paramètres. Au moment ou ces lignes sont écrites nous n'arrivons pas à obtenir de meilleures inférences en ajoutant ces statistiques à toutes les autres, qu'en les ignorant. Ce résultat peut paraître décevant à la lumière d'autres travaux (par exemple, Al-Asadi et al. 2019). Dans la mesure où ces autres travaux cherchent de l'information sur les changements des paramètres démographiques au cours du temps, alors qu'il n'y a pas de tels changements dans notre scénario, ces résultats ne sont pas nécessairement contradictoires. Cependant, l'on peut se demander si ce résultat est dû à un problème dans nos outils, ou à une statistique mal calibrée ne permettant pas de résumer les patrons de déséquilibres de liaison spécifiquement informatifs sur les paramètres d'intérêt dans nos scénarios biologiques. À l'inverse, il est possible que les autres statistiques résumantes utilisées dans ce travail aient été choisies et exploitées de façon particulièrement efficace, de sorte que l'information que l'on pourrait chercher dans les déséquilibres de liaison aura été aussi extraite de ces autres statistiques.

Ce canevas d'inférence a aussi permis l'analyse de l'impact de la prise en compte de la non-indépendance des arbres de coalescence des marqueurs non liés. Il semblerait, à l'heure de l'écriture de cette thèse, que les processus créant du déséquilibre gamétique affectent

nos inférences, et au final limitent la précision des inférences possibles, particulièrement concernant les distance de migration élevées. De la même manière que pour les résultats précédents l'on peut se demander si ce résultat est dû à un problème dans nos outils, ou s'il traduit un réel problème à traiter des marqueurs non liés comme pratiquement indépendants.

Références

Al-Asadi H, Petkova D, Stephens M, Novembre J. 2019. Estimating recent migration and population-size surfaces. *PLoS genetics*. 15 :e1007908.

Beaumont MA. 2010. Approximate Bayesian Computation in Evolution and Ecology. *Annual Review of Ecology, Evolution, and Systematics*. 41 :379–406.

Beaumont MA, Zhang W, Balding DJ. 2002. Approximation Bayesian computation in population genetics. *Genetics*. 162 :2025–2035.

Bell G, Koufopanou V. 1991. The architecture of the life cycle in small organisms. *Philosophical Transactions of the Royal Society of London Series B : Biological Sciences*. 332 :81–89.

Breiman L. 2001. Random Forests. *Machine Learning*. 45 :5–32.

Caswell H, Lensink R, Neubert M. 2003. Demography and dispersal : Life table response experiments for invasion speed. *Ecology*. 84 :1968–1978.

Caughley G. 1994. Directions in Conservation Biology. *Journal of Animal Ecology*. 63 :215–244.

Chesson P, Lee CT. 2005. Families of discrete kernels for modeling dispersal. *Theoretical Population Biology*. 67 :241–256.

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

Chikhi L, Sousa VC, Luisi P, Goossens B, Beaumont MA. 2010. The confounding effects of population structure, genetic diversity and the sampling scheme on the detection and quantification of population size changes. *Genetics*. 186 :983–995.

Cockerham CC, Weir B. 1987. Correlations, descent measures : drift with migration and mutation. *Proceedings of the National Academy of Sciences*. 84 :8512–8514.

Endler JA. 1977. Geographical variation, speciation, and clines. Princeton : Princeton University Press.

Ewens WJ. 1972. The sampling theory of selectively neutral alleles. *Theoretical Population Biology*. 3 :87–112.

Excoffier L, Dupanloup I, Huerta-Sanchez E, Sousa VC, Foll M. 2013. Robust Demographic Inference from Genomic and SNP Data. *PLoS Genetics*. 9 :e1003905.

Excoffier L, Ray N. 2008. Surfing during population expansions promotes genetic revolutions and structuration. *Trends in Ecology & Evolution*. 23 :347–351.

Fisher RA. 1930. The genetical theory of natural selection. Clarendon Press.

Fu YX. 2006. Exact coalescent for the Wright-Fisher model. *Theoretical Population Biology*. 69 :385–394.

Garza J, Williamson E. 2001. Detection of reduction in population size using data from microsatellite loci. *Molecular ecology*. 10 :305–318.

Goldstein DB, Harvey PH. 1999. Evolutionary inference from genomic data. *BioEssays*. 21 :148–156.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS genetics*. 5 :e1000695.

Haldane JBS. 1919. The combination of linkage values and the calculation of distance between the loci fo linked factors. *Hournal of Genetics*. 8 :299–309.

Hanski I. 1998. Metapopulation dynamics. *Nature*. 396 :41–49.

Hastings A, Cuddington K, Davies KF, Dugaw CJ, Elmendorf S, Freestone A, Harrison S, Holland M, Lambrinos J, Malvadkar U, et al. 2005. The spatial spread of invasions : new developments in theory and evidence. *Ecology Letters*. 8 :91–101.

Hayes B, Visscher P, McPartlan H, Goddard M. 2003. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Research*. 13 :635–643.

Heller NE, Zavaleta ES. 2009. Biodiversity management in the face of climate change : A review of 22 years of recommendations. *Biological Conservation*. 142 :14–32.

Hey J, Nielsen R. 2007. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proceedings of the National Academy of Science*. 104 :2785–2790.

Hill WG, Weir BS. 1988. Variances and covariances of squared linkage disequilibria in finite populations. *Theoretical Population Biology*. 33 :54–78.

Hudson RR. 1983. Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*. 23 :183–201.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*. 18 :337–338.

Kelleher J, Etheridge AM, McVean G. 2016. Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLOS Computational Biology*. 12 :e1004842.

Kimura M. 1968. Evolutionary rate at the molecular level. *Nature*. 217 :624–626.

Kimura M, Weiss GH. 1964. The Stepping Stone Model of Population Structure and the Decrease of Genetic Correlation with Distance. *Genetics*. 49 :561–576.

Kingman JFC. 1982. The coalescent. *Stochastic Processes and their Applications*. 13 :235–248.

Kot M, Lewis MA, van den Driessche P. 1996. Dispersal data and the spread of invading organisms. *Ecology*. 77 :2027–2042.

Leblois R, Estoup A, Rousset F. 2009. IBDSim : a computer program to simulate genotypic data under isolation by distance. *Molecular Ecology Resources*. 9 :107–109.

Leblois R, Estoup A, Streiff R. 2006. Genetics of recent habitat contraction and reduction in population size : does isolation by distance matter ? *Molecular ecology*. 15 :3601–3615.

Leblois R, Pudlo P, Néron J, Bertaux F, Beeravolu CR, Vitalis R, Rousset F. 2014. Maximum-Likelihood Inference of Population Size Contractions from Microsatellite Data. *Molecular Biology and Evolution*. 31 :2805–2823.

Lebret R, Iovleff S, Langrognet F, Biernacki C, Celeux G, Govaert G. 2015. Rmixmod : The R Package of the Model-Based Unsupervised, Supervised, and Semi-Supervised Classification Mixmod Library. *Journal of Statistical Software*. 67 :1–29.

Lehmann L, Perrin N, Rousset F. 2006. Population demography and the evolution of helping behaviors. *Evolution*. 60 :1137–1151.

Lewis WJ, van Lenteren JC, Phatak SC, Tumlinson I J H. 1997. A Total System Approach to Sustainable Pest Management. *Proceedings of the National Academy of Science*. 94 :12243–12248.

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

Loiselle BA, Sork VL, Nason J, Graham C. 1995. Spatial genetic structure of a tropical understory shrub, Psychotria officinalis (Rubiaceae). *American journal of botany*. 82 :1420–1425.

Lopez S, Rousset F, Shaw FH, Shaw RG, Ronce O. 2009. Joint Effects of Inbreeding and Local Adaptation on the Evolution of Genetic Load after Fragmentation. *Conservation Biology*. 23 :1618 – 1627.

Luikart G, England P, Tallmon D, Jordan S, Taberlet P. 2003. The power and promise of population genomics : From genotyping to genome typing. *Nature Reviews Genetics*. 4 :981–994.

Malécot G. 1966. Probabilité et Hérédité, volume 47 of *Travaux et Documents*. Institut national d'études démographiques, presses universitaires de france edition.

Mann HB, Whitney DR. 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*. 18 :50–60.

Marchi N, Schlichta F, Excoffier L. 2021. Demographic inference. *Current biology : CB*. 31 :R276–R279.

Mazet O, Rodríguez W, Grusea S, Boitard S, Chikhi L. 2016. On the importance of being structured : instantaneous coalescence rates and human evolution–lessons for ancestral population size inference ? *Heredity*. 116 :362–371.

Möhle M. 1998. Robustness results for the coalescent. *Journal of Applied Probability*. 35 :438–447.

Nei M. 1973. Analysis of Gene Diversity in Subdivided Populations. *Proceedings of the National Academy of Sciences*. 70 :3321–3323.

Nelson D, Kelleher J, Ragsdale AP, Moreau C, McVean G, Gravel S. 2020. Accounting for long-range correlations in genome-wide simulations of large cohorts. *PLOS Genetics*. 16 :e1008619.

Neyman J. 1935. On the problem of confidence intervals. *Annals of Mathematical Statistics*. 6 :111–116.

Norris K. 2004. Managing threatened species : the ecological toolbox, evolutionary theory and declining-population paradigm. *Journal of Applied Ecology*. 41 :413–426.

Ohta T, Kimura M. 1973. A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genetics Research*. 22 :201–204.

Osborne J, Loxdale H, Woiwood I. 2002. Monitoring insect dispersal : methods and approaches. In : Bullock J, Kenward R, Hails R, editors, Dispersal Ecology. British Ecol Soc, pp. 24–49.

Peron G, Crochet PA, Doherty PF Jr, Lebreton JD. 2010. Studying dispersal at the landscape scale : efficient combination of population surveys and capture-recapture data. *Ecology*. 91 :3365–3375.

Petkova D, Novembre J, Stephens M. 2016. Visualizing spatial population structure with estimated effective migration surfaces. *Nature genetics*. 48 :94–100.

Pudlo P, Marin JM, Estoup A, Cornuet JM, Gautier M, Robert CP. 2016. Reliable ABC model choice via random forests. *Bioinformatics*. 32 :859–866.

Raynal L, Marin JM, Pudlo P, Ribatet M, Robert CP, Estoup A. 2018. ABC random forests for Bayesian parameter inference. *Bioinformatics*. 35 :1720–1728.

Rousset F. 1997. Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics*. 145 :1219–1228.

Rousset F. 2000. Genetic differentiation between individuals. *Journal of Evolutionary Biology*. 13 :58–62.

Rousset F. 2007. Inferences from spatial population genetics. In : Balding DJ, Bishop M, Cannings C, editors, Handbook of statistical genetics, Chichester, U.K. : Wiley. third edition, pp. 945–979.

Rousset F. 2008. genepop'007 : a complete re-implementation of the genepop software for Windows and Linux. *Molecular ecology resources*. 8 :103–106.

Rousset F. 2022. Infusion : Inference using simulation. R package version 1.5.66. https ://gitlab.mbb.univ-montp2.fr/francois/Infusion.

Rousset F, Beeravolu CR, Leblois R. 2018. Likelihood computation and inference of demographic and mutational parameters from population genetic data under coalescent approximations. *Journal de la société Française de Statistique*. 159 :142–166.

Rousset F, Gouy A, Martinez-Almoyna C, Courtiol A. 2017. The summary-likelihood method and its implementation in the Infusion package. *Molecular Ecology Resources*. 17 :110–119.

Rousset F, Leblois R. 2012. Likelihood-based inferences under isolation by distance : two-dimensional habitats and confidence intervals. *Molecular Biology & Evolution*. 29 :957–973.

Saccheri IJ, Rousset F, Watts PC, Brakefield PM, Cook LM. 2008. Selection and gene flow on a diminishing cline of melanic peppered moths. *Proceedings of the National Academy of Sciences*. 105 :16212–16217.

Severini TA. 2000. Likelihood methods in statistics. Oxford statistical science series. Oxford University Press.

Slatkin M. 1985. Gene flow in natural populations. *Annual Review of Ecology and Systematics*. 16 :393–430.

Vitalis R, Couvet D. 2001. Estimation of effective population size and migration rate from one- and two-locus identity measures. *Genetics*. 157 :911–925.

Wakeley J. 2009. Coalescent theory : an introduction. Roberts & Co. Publishers.

Watts PC, Rousset F, Saccheri IJ, Leblois R, Kemp SJ, Thompson DJ. 2007. Compatible genetic and ecological estimates of dispersal rates in insect (*Coenagrion mercuriale* : Odonata : Zygoptera) populations : analysis of 'neighbourhood size' using a more precise estimator. *Molecular Ecology*. 16 :737–751.

Weir B, Cockerham C. 1974. Behavior of pairs of loci in finite monoecious populations. *Theoretical Population Biology*. 6 :323–354.

Weir BS, Cockerham CC. 1984. Estimating F-Statistics for the Analysis of Population Structure. *Evolution*. 38 :1358.

Whitlock M. 2002. Selection, load and inbreeding depression in a large metapopulation. *Genetics*. 160 :1191–1202.

Wilcoxon F. 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*. 1 :80.

Wright MN, Ziegler A. 2017. ranger : A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*. 77 :1–17.

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

Wright S. 1943. Isolation by distance. *Genetics*. 28 :114–138.

Wright S. 1951. The genetical structure of populations. *Annals of Eugenics*. 15 :323–354.

Wright S. 1965. The Interpretation of Population Structure by F-Statistics with Special Regard to Systems of Mating. *Evolution.* 19 :395.

Annexe

GSpace: an exact coalescence simulator of recombining genomes under isolation by distance

Thimothée Virgoulay, François Rousset, Camille Noûs, Raphaël Leblois

Bioinformatics, 2021, 1–3
doi: 10.1093/bioinformatics/btab261
Advance Access Publication Date: 8 May 2021
Applications Note

Genetics and population analysis

GSpace: an exact coalescence simulator of recombining genomes under isolation by distance

Thimothée Virgoulay [1,2,*], François Rousset [1], Camille Noûs[3] and Raphaël Leblois [2]

[1]Institut des Sciences de l'Evolution, Univ Montpellier, CNRS, IRD, EPHE, Montpellier, France,[2]CBGP, INRA, CIRAD, IRD, Montpellier SupAgro, Univ Montpellier, Montferrier sur Lez, France and [3]Laboratoire Cogitamus, Univ Montpellier, Montpellier, France

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz
Received on December 17, 2020; revised on April 16, 2021; accepted on April 27, 2021editorial decision on April 20, 2021;

Abstract

Motivation: Simulation-based inference can bypass the limitations of statistical methods based on analytical approximations, but software allowing simulation of structured population genetic data without the classical *n*-coalescent approximations (such as those following from assuming large population size) are scarce or slow.
Results: We present GSpace, a simulator for genomic data, based on a generation-by-generation coalescence algorithm taking into account small population size, recombination and isolation by distance.
Availability and implementation: Freely available at site web INRAe (http://www1.montpellier.inra.fr/CBGP/software/gspace/download.html).
Contact: thimothee.virgoulay@umontpellier.fr

1 Introduction

GSpace is a program that can simulate neutral genomic data with recombination under a wide range of demographic models. It is based on a backward in time generation-by-generation (gen-by-gen) approach, coupled with an efficient recombination algorithm and flexible models for dispersal and subpopulation sizes. It simulates the ancestry of a sample of haploid or diploid individuals (in both cases following a standard haplo-diploid sexual life cycle), carrying one or more chromosomes.

Individual dispersal is generally restricted in space in natural populations (isolation by distance: Endler, 1979; Rousset, 1997; Wright, 1943). To represent this fact, GSpace considers a lattice of subpopulations of any size (down to single individuals) connected by limited dispersal, according to different possible dispersal distributions, including in particular fat-tailed distributions, such as the Zeta (discretized Pareto, Patil and Joshi, 1968) and the Sichel (Chesson and Lee, 2005).

The case where each subpopulation on the lattice hosts a single individual or a mating pair and dispersal is mostly restricted to a few steps apart is suitable to represent a range of territorial species inhabiting a continuous habitat (e.g. Rousset, 2000), but cannot be simulated when considering extensions of Kingman's (1982) *n*-coalescent which assume large sub-population sizes and small migration rates. To simulate small population sizes and high dispersal rates without biases (Nelson *et al.*, 2020), coalescence probabilities exact for small population size (Fu, 2006) must be used in a gen-by-gen simulation until the common ancestors of the whole simulated sample have been found. Such simulations are required to assess any inference framework which might for example allow separate estimation of sub-population size, mutation and migration probabilities, something that is not possible under *n*-coalescent approximations. In all these respects, GSpace retains and extends for genomic data some of the previous features of the IBDsim software (Leblois *et al.*, 2009). The current version considers only time-homogeneous models but time-heterogeneous models will be implemented in future versions similarly to IBDsim.

As gen-by-gen algorithms are expected to be slower than those involving *n*-coalescent approximations, we performed simulations to check the feasibility of simulating genomic data by such algorithms, and compared computation times with those of alternative software based on *n*-coalescent approximations, such as msprime (Kelleher *et al.*, 2016), FastSimcoal2 (Excoffier *et al.*, 2013), exact coalescence algorithms implemented as DTWF in msprime python package [back-in-time Wright-Fisher simulator, Nelson *et al.* (2020)], IBDSim (Leblois *et al.*, 2009) and forward algorithms, such as SimBit (Matthey-Doret, 2020).

The gen-by-gen algorithm in GSpace is slower than those involving *n*-coalescent approximations but much faster than IBDSim or forward simulators like SimBit in most cases (see Section 4).

2 Implementation

GSpace combines some parts of the modified Hudson's algorithm (Hudson, 1983) for recombination and coalescence implemented in msprime with previous features from IBDSim, in a new implementation in modern C++, as follows. At each generation going backward in time,

the program considers all possible migration, recombination and coalescence events, until all common ancestors have been found. Because neutral genetic data are simulated, genetic states do not affect genealogical trees and mutations can then be added downwards to the gene tree of each chromosome segment that did not recombine. Implementation details of such algorithm can be found in Kelleher *et al.* (2016) and Nelson *et al.* (2020) for the approximated and gen-by-gen algorithms for coalescence with recombination, respectively; and in Leblois *et al.* (2009) for gen-by-gen algorithms under isolation by distance. We only highlight below what can make GSpace different from other software.

At each generation t, the coordinates of the parent of each individual carrying ancestral lineages are randomly drawn in a 2D backward dispersal distribution of the position of a parent given the position of the lineage. The backward distributions are deduced by assuming that dispersal occurs independently in each dimension forward in time, and can automatically handle spatial heterogeneity (i.e. different forward migration rates and size of sub-populations on the lattice) as well as various edge effects. The program can consider (i) uniform, geometric and discretized Gaussian, Zeta and Sichel forward dispersal distributions, including the stepping stone and island models as special cases, as well as (ii) a custom forward migration rate matrix. Each chromosome harbors multiple discrete potentially recombining sites and the program handles multiple recombination events per chromosome, even in a single generation. When a recombination event occurs in a diploid genome in the backward simulation, the segments on each side of the recombination point originate from each of the two parental chromosomes and have a distinct coalescence history further backward in time. When a coalescence event occurs, ancestral segments of all descendant chromosomes have a unique parental segment and share a common coalescence and migration history until a recombination event occurs. The combination of such gen-by-gen diploid coalescence, migration and recombination algorithms simulates the exact patterns of linkage disequilibrium expected under a haplo-diploid life cycle.

Mutations are then added independently on each gene tree, going forward in time on each branch, from the common ancestor to the leaves. As the underlying algorithm assumes a finite number of mutable sites GSpace can handle numerous nucleotidic and allelic mutation models (e.g. IAM, KAM, JC69, see user manual for more models) but not the infinite site model.

3 Compilation, automated checks, inputs and outputs

The program is written in modern C++ (17) and can be compiled on any operating system with a modern compiler (g++ $\geq$ 7.5, clang $\geq$ 6.0) with simple command line arguments, or using the CMake build system (both the command line arguments and the CMake commands are provided in the manual). The CMake build includes unit tests for each part of the program and functional tests comparing simulation outcomes in terms of probability of identity of pairs of genes at one and two loci to theoretical results (see Rousset, 2004 and Vitalis and Couvet, 2001).

GSpace's runs can be controlled both by a settings file and by command-line arguments, which together allow the easy specification of many parameters, and quick changes of selected parameters between simulations. The settings file is exacxtread first, and allows the user to control all options of GSpace (detailed in the user manual). These options can then be altered by the command line arguments. Results can be saved in three different file formats for individual genetic data: Genepop for allelic data, Fasta and VCF (v4.3) for sequence data; as well as in the new binary treeSequence format (see tskit documentation) for efficient storage of trees with mutations.

4 Comparison

Gspace is, at the time of writing, the only gen-by-gen coalescence simulator specifically designed to handle recombination and allowing easy specification of various forward dispersal distributions. Thus, it cannot be compared in terms of computation time to other simulators not sharing such features, but it has been compared to algorithms from five other simulators in simpler cases: the n-coalescent approximations implemented in msprime v1.0.0a5 ('msprime') and FastSimcoal2 v2.6.0.3 ('fsc2'); the gen-by-gen algorithms implemented in msprime v1.0.0a5 ('DTWF') and IBDSim v2.0; and the forward simulator SimBit v3.9.13. Results for IBDSim are not detailed here because it cannot consider recombination and is not designed to handle long DNA sequences (e.g. $> 10^5$ bp). However, without recombination and for many allelic loci, GSpace is two to fifty time faster. Other simulations are detailed in Table 1 and show that although GSpace is not the fastest simulator, its speed approaches that of the approximate simulators rather than that of other generation-by generation ones.

Table 1. Comparison of computation times between GSpace and other simulators under three different demographic and mutational schemes.

	Method (see text for details)				
Case	msprime	fsc2	GSpace	DTWF	SimBit
A	0.320	3.959	6.231	7.096	56.781
B	14.507	5.664	7.471	36.335	52.121
C	0.048	0.016	0.028	2.459	39.055

Note: Mean run time in seconds over 100 (10 for SimBit) replicates for the simulation of a sample of 1000 haploid individuals carrying a single chromosome of 10^7 base pairs, with mutation and recombination rates of 10^{-8} per generation per site under: (A) a Wright–Fisher model with a population size of 10000 haploid individuals; (B and C) an island model with 20 subpopulations of 500 haploid individuals each and 50 sampled chromosomes of (B) 10^7 base pairs or (C) with 10^4 base pairs

Acknowledgements

We thank A. Dehne-Garcia, F.-D. Collin and M. Navascues for initial discussions on algorithms and code, as well as J. Kelleher and P. Ralph for constructive comments and help with tskit during the review process.

Funding

This work used the following HPC platforms: INRA MIGALE (http://migale.jouy.inra.fr) and GENOTOUL (Toulouse Midi-Pyrénées), Montpellier Bioinformatics Biodiversity supported by the LabEx CeMEB (ANR-10-LABX-04-01), and CBGP host platform. All authors were supported by the Agence Nationale de la Recherche (RL & TV: projects GENOSPACE ANR-16-CE02-0008 and Labex Cemeb ProLag; FR & RL: project INTROSPEC ANR-19-CE02-0011).

Conflict of Interest: none declared.

References

Chesson,P. and Lee,C.T. (2005) Families of discrete kernels for modeling dispersal. *Theor. Popul. Biol.*, **67**, 241–256.
Endler,J.A. (1979) Gene flow and life history patterns. *Genetics*, **93**, 263–284.
Excoffier,L. *et al.* (2013) Robust demographic inference from genomic and snp data. *PLoS Genet.*, **9**, e1003905.
Fu,Y.-X. (2006) Exact coalescent for the wright–fisher model. *Theor. Popul. Biol.*, **69**, 385–394.
Hudson,R.R. (1983) Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.*, **23**, 183–201.
Kelleher,J. *et al.* (2016) Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Comput. Biol.*, **12**, e1004842.
Kingman,J. (1982) The coalescent. *Stoch. Process. Their Appl.*, **13**, 235–248.

Leblois,R. *et al.* (2009) Ibdsim: a computer program to simulate genotypic data under isolation by distance. *Mol. Ecol. Res.*, **9**, 107–109.

Matthey-Doret, R. (2021) SimBit: A high performance, flexible and easy-to-use population genetic simulator. *Mol Ecol Resour.* 10.1111/1755-0998.13372

Nelson,D. *et al.* (2020) Accounting for long-range correlations in genome-wide simulations of large cohorts. *PLoS Genet.*, **16**, e1008619.

Patil,G.P. and Joshi,S.W. (1968) *A dictionary and bibliography of discrete distributions*. Published for the International Statistical Institute by Oliver and Boyd Edinburgh.

Rousset,F. (1997) Genetic differentiation and estimation of gene flow from f-statistics under isolation by distance. *Genetics*, **145**, 1219–1228.

Rousset,F. (2000) Genetic differentiation between individuals. *J. Evol. Biol.*, **13**, 58–62.

Rousset,F. (2004) *Genetic Structure and Selection in Subdivided Populations*. Monographs in population biology. Princeton University Press, Princeton University, New Jersey.

Vitalis,R. and Couvet,D. (2001) Estimation of effective population size and migration rate from one- and two-locus identity measures. *Genetics*, **157**, 911–925.

Wright,S. (1943) Isolation by distance. *Genetics*, **28**, 114–138.