Approche multilocus du génome dans les modèles de génétique des populations

Mathieu Tiret

AgroParisTech
INSTITUT DES SCIENCES ET INDUSTRIES DU VIVANT ET DE L'ENVIRONNEMENT
PARIS INSTITUTE OF TECHNOLOGY FOR LIFE, FOOD AND ENVIRONMENTAL SCIENCES

Approche multilocus du génome dans les modèles de génétique des populations

Thèse de doctorat de l'Université Paris-Saclay
préparée à AgroParisTech

École doctorale n° 581 : Agriculture, alimentation, biologie, environnement et santé (ABIES)
Spécialité de doctorat : Génétique animale

Thèse présentée et soutenue à Paris, le 08/03/18, par

Mathieu Tiret

Composition du Jury :

Étienne Verrier Professeur, AgroParisTech	Président
Lounès Chikhi DR2, CNRS (UMR 5174)	Rapporteur
Maria Martinez DR2, INSERM (IRSD)	Rapporteur
Emmanuelle Baudry Maitre de conférence, ESE (UMR 8079)	Examinateur
Guillaume Achaz DR2, MNHN (UMR 7025)	Examinateur
Frédéric Hospital DR2, INRA (UMR 1313)	Directeur de thèse

Résumé

La génétique des populations est l'étude de l'évolution des fréquences alléliques au sein d'une population et de l'influence des pressions évolutives sur ces fréquences. Au sein de cette discipline, des modèles de population et des mesures génétiques sont développés pour pouvoir expliquer et prédire les données génétiques. Toutefois, au fur et à mesure des avancées technologiques, de nouveaux types de données sont disponibles, et il devient primordial de développer de nouveaux modèles et de nouvelles mesures pour pouvoir expliquer ces nouvelles données génétiques, plus denses et plus riches en marqueurs génétiques grâce à l'avènement de techniques comme la Next Generation Sequencing. Pour ce faire, nous proposons dans cette thèse de développer de nouvelles mesures avec une approche dite multilocus, qui considère le génome comme un tout plutôt que comme un agglomérat de locus indépendants. Dans un premier temps, nous avons tenté de construire une base théorique de l'approche multilocus en génétique des populations. Ensuite, nous avons illustré une telle approche à travers l'étude de l'identité par descendance, des graphes de recombinaison ancestraux et des autocorrélogrammes dans les modèles de génétique des populations. À travers ces différentes études de cas, nous avons tenté d'identifier les principaux enjeux et questions que soulève la génétique des populations multilocus.

Abstract

Population genetics is the study of the evolution of allelic frequencies within a population and the influence of evolutionary pressures on these frequencies. Within this field, one could develop population models and measures to explain and predict genetic data. However, as technologie evolves new types of data are available, and it becomes essential to develop new models and new measures to reflect these new genetic marker data, increasingly richer and denser thanks to the advent of new techniques such as the Next Generation Sequencing. To this end, we propose in this thesis to develop new measures with the so-called multilocus approach, which considers the genome as a whole rather than an agglomerate of independent loci. We have first tried to build a theoretical basis for the multilocus approach in population genetics. Then, we have illustrated this multilocus approach with the case studies of identity by descent, ancestral recombination graphs and autocorrelograms in population genetics models. Through these different studies, we tried to identify the main issues and questions that the multilocus population genetics raises.

Table des matières

Introduction

La génétique des populations est l'étude théorique de l'évolution au sein d'une population des fréquences alléliques et de l'influence de la démographie et des pressions évolutives sur ces fréquences. Les travaux de S. Wright, R. A. Fisher et de J. B. S. Haldane, considérés comme les pères fondateurs de cette discipline, ont amplement contribué à son développement, au point que leurs travaux sont encore largement utilisés de nos jours. Aujourd'hui, la génétique des populations est avant tout une discipline qui fournit des outils mathématiques simples pour l'étude et l'interprétation de données génétiques.

La génétique des populations est une discipline vaste qui peut être abordée sous plusieurs angles : la modélisation (construction de modèles de population), la méthodologie (développement d'outils pour l'étude de modèles de population), la sélection de modèle (recherche d'un modèle de population approprié à une population réelle donnée) ou encore l'analyse de données (inférence de paramètres clés et interprétation des données). Nous nous concentrerons dans cette thèse sur la modélisation et la méthodologie, et ne traiterons pas à proprement parler de la sélection de modèle ni de l'analyse de données. En revanche, nous tenterons de définir le plus précisément possible la population, objet mathématique au cœur de cette discipline, et les objets

associés pour pouvoir clarifier les hypothèses que posent certains modèles de population ; cette tentative devrait grandement faciliter le travail de sélection de modèle et améliorer la qualité des analyses de données.

Cette thèse porte sur la modélisation et la méthodologie des caractères dits multilocus : conception et développement de modèles de population multilocus d'une part, et définition, développement et étude de mesures multilocus d'autre part. Dans ce vaste domaine qu'est la génétique des populations multilocus, nous nous sommes concentrés sur la transmission (multilocus) du génome, et sur la façon de rendre compte de cette transmission ; en d'autres termes, nous avons essayé de suivre et de décrire l'évolution au cours du temps des associations alléliques (ou haplotypes). Nous montrerons dans un premier temps en quoi cette transmission est, dans les modèles de population multilocus, "en bloc". Ensuite, nous essaierons d'en rendre compte à travers l'identité par descendance (certains l'appelleront identité par ascendance), les graphes de recombinaison ancestraux et les autocorrélogrammes dans ces modèles de population.

Cette thèse de génétique des populations est un travail de recherche fondamentale, nous avons donc essayé de fonder tout ce que nous avons entrepris, en définissant les termes les plus élémentaires et en clarifiant au mieux les axiomes des modèles. Ce travail peut sembler futile, mais la polysémie est fortement présente en génétique des populations, notamment pour le concept de population (Debouzie, 1999; Hartl et al., 1997). Qu'entend-on par "la" génétique des populations lorsque le concept même de population est pluriel ? Par conséquent, nous n'avons pas supposé de prime abord "une" génétique des populations, mais essayé de préciser *quelle* génétique des populations nous avons étudiée, en clarifiant le plus possible notre travail avec un vocabulaire

préexistant ou ad hoc. L'élément essentiel de notre thèse étant le locus, nous avons essayé autant que possible de constuire les objets à travers le lien qu'ils partagent avec le locus.

Le travail de modélisation (la première partie) mené dans cette thèse a été un travail de définition des objets théoriques. Il est important de constater qu'un objet théorique n'est pas un objet réel, en ce sens que l'essence d'un objet théorique est complètement épuisée lorsqu'il est défini, mais pas celle d'un objet réel. Pour bien comprendre, prenons l'exemple d'un objet réel simple, une aubergine : la définir n'épuise pas son essence, puisqu'une aubergine est bien plus que sa définition. Par contre, définir une variable x dans $\mathbb{R}$ la définit totalement, toutes propriétés annexes sont des tautologies de la définition de x ; autrement dit, une définition est une propriété parmi d'autres, dans un ensemble de propriétés équivalentes. Nous ne travaillerons ici qu'avec des objets théoriques, ce qui implique que le travail de modélisation sera un travail de définition des termes : bien définir les objets *est* la construction de ces derniers. Par conséquent, notre première partie sur la modélisation est un ensemble de définitions et d'intrications entre ces définitions. Ces définitions pourront paraitre triviales, mais nous les avons spécifiées précisément parce qu'elles ne le sont pas et qu'elles auraient pu être tout autre.

Enfin, fort de ce travail de modélisation, nous développerons les statistiques multilocus en lien avec notre problématique : la transmission génétique "en bloc", ou multilocus. Dans un premier temps, notre étude portera sur l'identité par descendance (ou IBD pour identity by descent) dans les modèles de population multilocus. Cette étude, principalement de modélisation, nous permettra de prédire, pour tous les modèles de population (à quelques conditions près), l'évolution au cours du temps de la taille des blocs IBD.

Dans un deuxième temps, nous étudierons le “chromosome painting” en combinaison avec les graphes de recombinaison ancestraux. Nous introduirons le concept de “cluster”, et étudierons numériquement son comportement autant dans les modèles “forwards” que “backwards”. Cette étude nous permettra d’illustrer la grande variance des mesures multilocus, ainsi que leur forte sensibilité aux différentes pressions évolutives. Enfin, dans un troisième projet nous étudierons les autocorrélogrammes, exclusivement de façon numérique, et leur capacité à concilier l’aspect multilocus et l’applicabilité.

Chapitre 1

Une génétique des populations

1.1 La population, un objet inscrit dans le temps

Définissons une population comme un ensemble d'individus qui n'intéragissent avec aucun élément extérieur. Les individus d'une population ont chacun une date d'apparition et de disparition, et peuvent ainsi être répartis le long d'un axe continu, celui du temps. Une population est donc un objet inscrit, à travers ses individus, dans le temps, et à un instant donné il existe un ensemble unique d'individus qui représentent cette population. Par la suite, nous appellerons t-population l'ensemble des individus observés à l'instant t. Par abus de langage, on appelle plus communément "population" ce que nous avons ici défini comme "t-population", mais par souci de précision et de clarté, nous continuerons à distinguer ces deux termes.

Une population peut être représentée comme une suite de t-populations, l'indice de cette suite étant le temps. Cette représentation est équivalente à une autre, plus centrée sur l'invididu et qui consisterait à enregistrer les dates d'apparition et de disparition de chaque individu. Dans le cadre de

la génétique des populations qui, comme son nom l'indique, est davantage centrée sur les populations, nous préfèrerons la première. En effet, ce qui nous intéresse ici ne sont pas les individus en soi, ni leur pedigrée, ni leur durée de vie à chacun, mais seulement ce qu'ils représentent ensemble dans une population.

Définissons une statistique comme une valeur réelle (ou un vecteur réel) issue d'une mesure sur une t-population. Il est ainsi possible de représenter une population avec une suite de statistiques (de valeurs réelles donc). Cette opération engendre une nécessaire perte d'information liée à la mesure, nous pourrions davantage parler de projection d'une population en une suite de statistiques, plutôt que d'une représentation. Une suite de statistiques étant beaucoup plus simple à étudier qu'une population en soi, la possibilité d'une telle projection est primordiale.

Enfin, précisons que cette façon de procéder – projeter une population en une suite de statistiques malgré la nécessaire perte d'information – est motivée par une hypothèse, celle de l'existence d'une bijection entre une population et l'ensemble des statistiques mesurables sur cette dernière ; autrement dit, si on pouvait tout mesurer sur une population, on la connaitrait parfaitement. N'ayant évidemment pas accès à toutes les mesures possibles, l'objectif sera de trouver les statistiques les plus informatives pour avoir une bonne image d'une population. C'est ainsi que par la suite, nous étudierons une population à travers des statistiques, autant que nous voulons bien en considérer de différentes, qui, ensemble, formeront une approximation de cette population.

1.2 Les individus, le tissu de la population

Nous avons vu précédemment que les individus, à travers leurs dates d'apparition et de disparition, donnent la dimension temporelle à la population. Définissons maintenant les individus en procédant par étapes et en commençant par décrire l'objet réel (le référent) que nous voulons modéliser ; autrement dit, identifions les propriétés qui nous semblent pertinentes. L'aspect de l'individu réel qui nous intéresse dans le cadre de la génétique des populations est son génome. Un gène chez un individu est caractérisé par sa position (ou locus) et sa séquence (ou allèle). De plus, un individu porte un nombre donné d'allèles par locus, nombre que l'on appelle ploïdie et qui peut avoir une valeur différente pour chaque locus : par exemple, chez les humains la ploïdie des locus des autosomes vaut 2, alors que celle des locus des gonosomes vaut 1. Notez également que la ploïdie n'a rien à voir avec le nombre de parents d'un individu : certains individus diploïdes peuvent provenir d'une reproduction asexuée (avec un seul parent donc) ; ou encore, certaines variétés de blé peuvent être polyploïdes (diploïdes ou triploïdes) tout en n'ayant que deux parents (Levy et Feldman, 2004). Les supports biologiques de ces locus (chromosomes, plasmides...) ne nous intéressent que dans leurs différentes façons de se transmettre, et dans le cadre de cette thèse nous ne considèrerons que le cas des chromosomes ; en d'autres termes, tous nos modèles s'inspireront de chromosomes réels et de leur mode de transmission.

Fort de ces constats, nous définirons un individu comme un ensemble de chromosomes, ensemble que nous appellerons génome. Nous supposerons de plus par souci de simplicité que tous les locus ont la même ploïdie (nous

parlerons ainsi de *la* ploïdie d'un individu), et que l'individu ne porte que des chromosomes homologues, c'est-à-dire des chromosomes qui portent les mêmes locus. En résumé, ici, un individu est un ensemble de chromosomes homologues (dont le nombre dépend de *la* ploïdie), dont les locus sont les positions sur lesdits chromosomes et les allèles d'un locus les versions possibles du génome portées par ledit locus – le contenant (locus) et le contenu (allèle) en quelque sorte. C'est une définition classique en génétique des populations, bien que détaillée ici, mais il convient toutefois de noter que nous négligeons avec cette définition une multiltude d'éléments considérables de l'individu (épigénétique, plasmides, mitochondries...).

Nous supposerons par la suite que tous les individus d'une population ont les mêmes locus en leur sein – une hypothèse souvent implicite en génétique des populations qui exclut par là même les reproductions hybrides, ou encore l'haplodiploïdie (chez les abeilles par exemple). Nous parlerons ainsi de *la* ploïdie d'une population. Fort de cette hypothèse – les individus d'une population ont tous les mêmes locus – nous parlerons par la suite de locus indifféremment dans un individu ou dans une population.

Enfin, voyons en quoi les individus sont le tissu de la population. Appelons évènement de reproduction l'apparition d'un individu dans la population au cours duquel l'individu hérite, selon un processus déterminé, de tout ou partie des génomes de son ou ses parents. Considérons de plus que tout les individus proviennent d'un évènement de reproduction ; le ou les parents de cet individu sont dans la même population, mais dans une t-population en amont dudit évènement de reproduction. Les seuls individus n'ayant pas de parent dans cette population sont les "fondateurs" de cette dernière, et sont les premiers individus de cette population. Nous avons ainsi construit

au sein de la population une cohésion reproductive qui implique qu'entre deux t-populations, certains individus de l'une sont descendants de certains individus de l'autre. Il est à noter que la construction d'une telle cohésion au sein d'une population est pertinente en génétique des populations, mais qu'il aurait été possible de construire une cohésion tout à fait différente dans un autre domaine (écologique, sociale...).

1.3 Les modèles de population et stochasticité

À l'aide des objets préalablement définis, la population et l'individu, définissons maintenant les modèles de population, objets au cœur de la génétique des populations. Procédons une fois de plus par étapes, en commençant par constater que dans un environnement donné, deux populations peuvent évoluer différemment à cause de phénomènes aléatoires. Ainsi, une population telle que définie précédemment n'est finalement qu'une réalisation particulière de lois sous-jacentes dépendantes d'un environnement donné ; en d'autres termes, une réalisation d'un modèle sous-jacent. Définissons ainsi un modèle de population comme un ensemble des lois sous-jacentes qui régissent une population. Nous pouvons également définir un modèle de manière strictement équivalente comme un ensemble de populations possibles.

Ce concept de modèle de population nous est essentiel dans le cadre de la génétique des populations. En effet, forts de cette définition nous étudierons désormais une population en tant que réalisation du modèle de population sous-jacent. Par ailleurs, en analyse de données, les données réelles ne peuvent être interprétées que sous le prisme d'un modèle choisi arbitrairement ou rationnellement : les paramètres inférés à partir d'une population sont ceux du

modèle de population correspondant. Les modèles de population sont donc au cœur de cette discipline, et ne sont pas considérés comme des simplifications de populations réelles, mais comme générateurs desdites populations.

Nous nous intéresserons par la suite aux modèles de population stochastiques qui prennent en compte certains aléas pouvant subvenir dans une population, telle que la disparition accidentelle de certains individus, même si ces derniers étaient adaptés à leur environnement. En d'autres termes, les modèles de population intégrant la stochasticité permettent de prendre en compte l'inattendu et l'inconnu, et donc d'intégrer la probabilité dans les modèles. Les modèles de population déterministes (qui ne prennent pas en compte l'aléa), bien qu'il y en ait de pertinents en quantité importante, ne feront pas l'objet de cette thèse. Comme nous l'avons dit précédemment, un modèle de population est un ensemble de populations, par conséquent un modèle de population stochastique est un ensemble de populations formant un espace probabilisé ; autrement dit, un modèle de population attribue une probabilité à chaque population.

En outre, nous présenterons les modèles de population sous un angle génératif : à l'instar des ensembles mathématiques que l'on peut définir en extension (en donnant la liste complète des éléments qui composent cet ensemble) ou en compréhension (en donnant les propriétés qui permettent de générer ses éléments), un modèle de population, qui est un ensemble de populations, peut être défini en extension ou en compréhension. Dans cette thèse, nous les définirons en compréhension et nous nous concentrerons par la suite à fournir les propriétés qui permettent de générer des populations.

Enfin, sachant qu'une population peut être projetée en une suite de statistiques, nous pouvons dire qu'un modèle de population peut l'être en un

ensemble de suites de statistiques. De plus, un modèle de population stochastique attribue une probabilité à chaque suite de statistiques. Si chaque suite de statistiques, dont l'indice est pour rappel le temps, est associée à une probabilité, en "empilant" toutes les suites nous pouvons obtenir pour chaque instant t une distribution de statistiques (sur l'ensemble des t-populations). Il est ainsi possible de construire à partir d'un modèle de population stochastique une suite de distributions, ou par souci de simplicité des premiers moments, d'une statistique : ce sera par la suite le sujet de notre étude.

1.4 Les modèles en génération

Nous avons vu précédemment qu'une population peut être représentée sous la forme d'une suite (de t-populations) avec comme indice le temps, a priori continu. Une approche possible et plus simple consiste à étudier la population uniquement à certains temps clés – les générations. Ainsi, une population peut être représentée en une suite dont les indices sont désormais les générations. Après cette discrétisation nous n'étudions plus toute la suite de t-populations mais seulement une sous-suite de celle-ci, ce qui facilite grandement son étude aux dépens d'une nécessaire perte d'information.

Il existe plusieurs types de générations, mais nous nous concentrerons uniquement sur les générations dites non chevauchantes. Le temps entre générations non chevauchantes (ou temps intergénérationnel) est défini ici comme le temps d'attente minimum et nécessaire pour renouveler toute la t-population, de sorte que deux t-populations à deux générations différentes n'ont aucun individu en commun. Cette définition implique, d'une part, que les temps intergénérationnels n'ont pas nécessairement la même durée, et d'autre part

que deux individus d'une même génération ne proviennent pas nécessairement du même nombre d'évènements de reproduction depuis la génération précédente. Pour surmonter ce problème et faciliter l'analyse, nous supposerons que deux générations successives sont éloignées d'un seul évènement de reproduction pour chaque individu : tous les individus d'une génération ont donc nécessairement leur(s) parent(s) dans la t-population de la génération précédente ; autrement dit, les générations ne sont pas chevauchantes. Notons que définir le temps intergénérationnel (pour les générations non chevauchantes) comme le temps d'attente nécessaire pour renouveler toute la t-population ne contraint absolument pas l'ensemble de populations que nous étudions, il existe effectivement de telles générations dans toutes les populations possibles ; en revanche, supposer que les parents des individus d'une génération sont dans la génération précédente contraint ledit ensemble, et dans ce sens cette hypothèse est forte.

Discrétiser ainsi le temps sur lequel la population est définie permet de faciliter grandement l'introduction des schémas de reproduction dans les modèles de population, schémas que nous définissons comme fonctions permettant de passer d'une t-population d'une génération à celle de la suivante. Un schéma de reproduction est la loi sous-jacente des évènements de reproduction. Nous pouvons distinguer dans les schémas de reproduction le schéma d'accouplement (processus déterminant le(s) parent(s) d'un enfant) et le schéma de méiose (processus d'héritage à proprement parler du génome du ou des parents par les enfants).

Le deuxième intérêt de cette discrétisation est de pouvoir normaliser les moments où la démographie et les pressions évolutives influenceront la population : lors du passage d'une génération à la suivante, c'est-à-dire au cours

du schéma de reproduction. Plus précisément, la démographie interviendra uniquement dans le schéma d'accouplement, tandis que les pressions évolutives peuvent intervenir à la fois dans le schéma d'accouplement et dans le schéma de méiose.

Enfin, illustrons maintenant les modèles en générations non chevauchantes avec un modèle de population très classique : le modèle de Wright-Fisher, dans sa version haploïde. Les individus sont haploïdes, n'ont qu'un seul parent et ne portent qu'un seul locus. La population est définie d'un état initial composé d'individus généralement différents à un état final de fixation (un allèle a envahi toute la population). Les générations sont supposées, comme dit précédemment, éloignées d'un seul évènement de reproduction pour chaque individu. Le schéma d'accouplement est panmictique (chaque individu a la même probabilité de s'accoupler) et le schéma de méiose est un héritage simple (l'enfant hérite sans équivoque de l'allèle de son parent). La taille de population est constante le long des générations. Il existe d'autres variantes possibles de ce modèle, du fait que l'on puisse faire varier la ploïdie des individus, le nombre de parents par individu, le nombre de locus par individu, l'état initial et final de la population, le schéma de reproduction et l'évolution de la taille de population. Les différentes variantes du modèle ont toutes en commun l'organisation de la population en générations non chevauchantes. Par la suite, nous nous concentrerons uniquement sur les modèles de population (stochastiques) en génération.

1.5 Phénotype et schéma d'accouplement

Comme nous l'avons dit précédemment, dans un modèle de population en génération il est possible de définir un schéma d'accouplement. Les caractéristiques des individus, notamment leur phénotype, déterminent les accouplements. Nous définissons d'ailleurs tautologiquement le phénotype d'un individu comme les caractéristiques qui interviennent pendant le schéma d'accouplement – tous les autres aspects du phénotype réel ne nous intéressent pas ici. Le phénotype d'un individu le favorise ou non, relativement à d'autres phénotypes, dans la participation auxdits accouplements.

L'exemple canonique du schéma d'accouplement est la panmixie, au cours de laquelle tout individu a la même chance de s'accoupler ; autrement dit, tous les phénotypes sont identiques. Il est à noter, cependant, que la probabilité d'un tel schéma extrême diminue lorsque le nombre de locus augmente : en effet, supposer qu'*un* locus soit sous panmixie est concevable, beaucoup moins quand il s'agit d'un chromosome entier.

1.6 Génotype et schéma de méiose

Décrivons maintenant le schéma de méiose qui, contrairement au schéma d'accouplement, prend directement en compte le génome des individus – et ainsi les locus. Ce schéma définit comment à partir du ou des génomes parentaux obtenir le génome enfant, en considérant les locus (via le brassage inter- et intrachromosomique, nous y reviendrons) et éventuellement l'effet de certains allèles (en particulier les distorseurs de ségrégation). Nous négligerons par la suite l'effet des allèles dans les schémas de méiose.

Rappelons qu'un locus est une position (unique) sur le génome, et que

dans un schéma de méiose c'est la relation entre locus qui nous intéresse. La nature de ces relations est un peu particulière en raison de l'intérêt particulier que l'on y porte : tout ce qui nous intéresse ici est de savoir si des allèles portés par *un* des chromosomes homologues et sur, disons, deux locus différents sont transmis ensemble ou non d'une génération à l'autre. Autrement dit, c'est la conservation des associations formées par les allèles et la probabilité de cette conservation qui nous intéressent. Historiquement, cette relation a été quantifiée entre deux locus avec les taux de recombinaison, qui représentaient la probabilité de non-conservation des associations à ces deux locus. Ainsi, chaque paire de locus a un taux de recombinaison associé.

En utilisant les taux de recombinaison, on peut déterminer l'ordre entre trois locus A, B et C, qui sont dits alignés dans l'ordre A-B-C lorsque $r_{AC} < r_{AB} + r_{BC}$, où r_{AB}, r_{BC} et r_{AC} sont les taux de recombinaison entre les locus A et B, B et C, et A et C respectivement. Théoriquement, si les locus A, B et C sont alignés dans cet ordre, la formule suivante est respectée :

$$1 - 2\, r_{AC} = (1 - 2\, r_{AB})(1 - 2\, r_{BC}) \tag{1.1}$$

en utilisant les mêmes notations que précédemment. Si on connait pour chaque locus son taux de recombinaison avec n'importe quel autre locus, grâce à l'équation (1.1), il est possible de déduire pour toute paire de locus son taux de recombinaison.

Le fait que le taux de recombinaison soit une probabilité et non une distance complexifie significativement la modélisation : par exemple, le taux de recombinaison entre les locus A et C n'est pas la somme du taux de recombinaison entre A et B et de celui entre B et C. Par la suite nous considèrerons

comme relation entre locus la distance en Morgan (et le modèle de recombinaison de Haldane associé). Cette distance est une transformation directe du taux de recombinaison entre deux locus, formulée comme suit (Haldane, 1919) :

$$d_{AB} = -0,5 \cdot ln(1 - 2\, r_{AB}) \tag{1.2}$$

où d_{AB} est la distance en Morgan entre les locus A et B, et r_{AB} le taux de recombinaison entre les locus A et B. Cette relation a les propriétés d'une distance, et est notamment additive. Cette distance pose néanmoins l'hypothèse que les interférences génétiques (une recombinaison influence le taux d'apparition d'autres recombinaisons dans son voisinage) sont négligeables. De plus, on ne peut considérer de taux de recombinaison supérieur à 0,5 pour que la distance en Morgan soit définie. Nous avons d'ailleurs supposé auparavant que les individus étaient uniquement composés de chromosomes homologues, et nous en voyons enfin les conséquences : tous les taux de recombinaison entre locus sont par là même supposés strictement inférieurs à 0,5 (sachant que le taux de recombinaison entre deux locus sur deux chromosomes non homologues est *égal* à 0,5). Les équations (1.1) and (1.2) définissent chacun ce que nous appelerons une métrique.

Les deux relations caractérisent toujours une paire de locus. La différence fondamentale entre le taux de recombinaison et la distance en Morgan est simplement que la deuxième est effectivement une distance, modifiant ainsi toute la modélisation subséquente : on parle ainsi de *longueur*, de *position*... termes qui n'ont de sens qu'avec une métrique donnée, habituellement additive. Les taux de recombinaison permettraient de définir une telle "position"

pour chaque locus, mais il faudrait passer par l'équation (1.1) à chaque fois, là où avec la distance en Morgan une simple opération arithmétique suffit. Une fois le terme de "position" définie avec une métrique donnée, nous pouvons dire que, si le locus est la position sur le chromosome, alors le locus *est* la relation avec les autres locus : sa position le définit par rapport aux autres locus, sa relation avec les autres est ainsi définie.

Nous pouvons donc parler de distance entre locus, et donc de longueur de chromosome, que nous considèrerons finie dans cette thèse, c'est-à-dire qu'il existe entre les locus une distance maximale. Deux locus sont dits adjacents s'il n'y a pas de locus entre lesdits locus, et par conséquent, en supposant une longueur de chromosome finie, chaque locus d'un chromosome a deux voisins adjacents, à l'exception de deux locus qui n'en ont qu'un (que nous appellerons bordures de chromosome). Par convention, les positions des locus seront toujours données en fonction d'une même bordure de chromosome.

Précisons désormais comment les locus interviennent dans le schéma de méiose qui détermine les allèles parentaux transmis à l'individu enfant lors d'une méiose. Le schéma de méiose que nous adopterons met en œuvre un processus de recombinaison : dans le cas d'individus haploïdes, les parents produisent un gamète qui *est* l'individu enfant ; dans le cas d'une ploïdie de valeur supérieure (diploïdie, triploïdie...), chaque parent produit un ou plusieurs gamètes, qui, ensemble, forment l'individu enfant. Dans les deux cas, chaque gamète est haploïde, et est obtenu via un processus au cours duquel, à une bordure de chromosome, le gamète hérite d'un des chromosomes parentaux tiré au hasard (c'est le brassage interchromosomique), puis continue d'hériter de ce chromosome parental jusqu'à ce qu'une coupure survienne (c'est le brassage intrachromosomique), après quoi le gamète hérite

d'un autre chromosome parental. Ainsi, sur le gamète l'origine parentale diffère de part et d'autre d'une coupure (ou crossover), d'où les blocs d'origines parentales différentes : c'est pourquoi nous parlons de "transmission en bloc", que nous évoquerons également par transmission multilocus.

Par la suite, nous utiliserons plus précisément le processus de recombinaison de Haldane, dont les crossovers surviennent selon un processus de Poisson d'intensité 1 le long du chromosome du gamète. En d'autres termes, les distances (en Morgan) entre crossovers sont des variables aléatoires distribuées selon une loi exponentielle de paramètre 1. Ceci revient à dire que le nombre de crossovers est une variable aléatoire distribuée selon une loi de Poisson de paramètre l, où l est la longueur du chromosome, et dont les positions sont des variables aléatoires indépendantes et uniformément distribuées entre 0 et l.

1.7 Les données en génétique des populations

Une fois que nous avons défini la population, l'individu, le modèle de population et le schéma de reproduction, il nous reste à décrire le dernier objet théorique fondamental de la génétique des populations : les données. Nous ne décrirons pas ici comment les obtenir, mais plutôt ce qui est perçu comme donnée en génétique des populations. Cet objet porte très mal son nom, puisqu'il donne l'impression d'être une réalité qui est "donnée à voir" (Sellars et al., 1956; Thouzeau, 2017), alors qu'une donnée se construit et s'interprète à travers une théorie : il n'existe pas de donnée hors théorie. La donnée est donc un objet que nous devons construire et ce, en particulier dans le cadre de la génétique des populations. Pour ce faire, nous allons

construire la donnée à travers les relations qu'elle peut avoir avec les objets précédemment définis.

Définissons les données comme les sorties mesurables d'un modèle de population. Puisque les modèles de population sont les générateurs des populations, toutes les statistiques mesurées sur ces populations sont des données. Comprenons bien que nous parlons de la population telle que définie précédemment, les statistiques sur une telle population ressortent nécessairement de la génétique des populations. Par la suite, nous distinguerons deux types de données : les données réelles et les pseudo-données.

Les données réelles sont des statistiques mesurées sur une population réelle, sans connaitre a priori le modèle sous-jacent à cette population ; le travail de sélection de modèle prend ainsi tout son sens. Pour être plus précis, la sélection de modèle inclut la sélection du scénario démographique le plus approprié à une population (par exemple, la présence ou l'absence de goulot d'étranglement), ainsi que le calibrage des modèles (par exemple, la date du goulot d'étranglement). Il existe deux façons principales d'effectuer ce travail, le calcul de vraisemblance ou l'approche bayésienne (la computation bayésienne approchée par exemple), que nous ne faisons que mentionner. Une fois qu'un modèle de population est sélectionné, les données réelles sont utilisées en analyse de données pour inférer les paramètres du modèle de population correspondant.

Les pseudo-données, en revanche, sont littéralement générées par un modèle de population et ne sont utilisées qu'à des fins méthodologiques. Dans ce cas, le travail de sélection de modèle est inutile, ou plutôt le meilleur travail possible a été effectué puisque le modèle correspond exactement à la population. Ces pseudo-données sont extrêmement utiles pour pouvoir développer

des statistiques informatives sur un modèle donné, en d'autres termes mener à bien un travail de méthodologie en génétique des populations. Dans cette thèse qui est un travail de modélisation et de méthodologie, seules des pseudo-données seront considérées.

1.8 Choix des statistiques, choix de la problématique

Définir les pseudo-données nous a permis d'identifier leur rôle central dans la méthodologie, permettant de développer des statistiques informatives sur les populations. À noter que ce que nous appelons une "statistique informative" dépend de la problématique considérée : par exemple, lorsque l'on essaie de détecter une signature de sélection dans le génome, on attend d'une statistique informative qu'elle puisse effectivement la détecter avec une puissance et une robustesse satisfaisantes.

Il ne faut cependant pas oublier que le comportement d'une statistique dépend fortement du modèle de population considéré, il n'y a aucune garantie qu'une statistique se comporte de façon similaire dans tous les modèles. Le D de Tajima, par exemple, est une statistique connue pour pouvoir détecter des signatures de sélection ; toutefois, dans certains modèles, son comportement est juste un artefact donnant une illusion de signature de sélection qui peut être, entre autres, dû à des démographies complexes. L'étude d'une statistique est ainsi indissociable du modèle dans lequel cette statistique est étudiée. La méthodologie est en cela fastidieuse qu'il faut, pour être exhaustif, renseigner le comportement d'une statistique dans tous les modèles de population.

Pour résumer, nous avons dit qu'une statistique était issue d'une mesure

réelle (valeur ou vecteur) sur une t-population, et ainsi qu'un modèle de population stochastique pouvait être projeté en une suite de distributions de statistiques : à chaque temps (ou génération), nous aurions une distribution d'une statistique (sur l'ensemble des t-populations). L'attribution des probabilités sur l'ensemble des t-populations provient de la formulation dudit modèle de population stochastique. Par souci de simplicité, nous étudierons plutôt une suite des premiers moments d'une statistique.

Enfin, puisque les statistiques impliquent une perte d'information (raison pour laquelle on cherche des statistiques informatives par rapport à d'autres qui ne le seraient pas), il nous faut choisir une (ou plusieurs) statistique pour étudier une population, choix qui fait pleinement partie de la problématique. Dans cette thèse, et nous le verrons dans la partie suivante, nous nous intéresserons aux statistiques multilocus, c'est-à-dire des statistiques qui considèrent le génome comme un tout et non comme des parties indépendantes.

1.9 La génétique des populations multilocus

Définissons une mesure multilocus comme une mesure qui intègre plusieurs locus et leurs relations (la distance en Morgan dans notre cas) ; ou plutôt, puisque le locus est la relation avec les autres locus (1.6), une mesure multilocus est tout simplement une mesure qui prend en compte les locus. De telles mesures se focalisent principalement sur le concept de voisinage entre locus, et donc entre allèles portés par lesdits locus. Un haplotype est un ensemble d'allèles voisins (c'est-à-dire portés par un seul chromosome et sur des locus voisins), et ce concept est à la base de nombreuses mesures multilocus : l'Extended Haplotype Homozygosity – ou EHH – (Sabeti et al.,

2002; Mueller et Andreoli, 2004; Bersaglieri et al., 2004), ou encore les Runs of Homozygosity – ou ROH – (McQuillan et al., 2008; Bosse et al., 2012; Pemberton et al., 2012; Szpiech et al., 2013; Curik et al., 2014). Dans les modèles de population avec des individus diploïdes, ces mesures sont les longueurs des régions homozygotes – où les haplotypes sont identiques entre les deux chromosomes d'un individu. L'homozygotie est intéressante dans l'étude d'une mesure multilocus en tant qu'approximation de l'identité par descendance – ou IBD pour identity by descent. Deux haplotypes sont dits IBD s'ils sont des copies héritées d'un ancêtre commun, nous voyons ainsi en quoi l'IBD est au cœur des études sur la transmission génétique, multilocus ou non.

L'étude de la longueur de régions IBD, ou homozygotes par approximation, permet de considérer l'aspect multilocus de la transmission génétique, qui est dans notre cas "en bloc" à cause du processus de recombinaison. Comme dit précédemment, puisque la démographie et les pressions évolutives interviennent lors de la transmission (d'un passage d'une génération à une autre), les mesures multilocus visant à caractériser la transmission sont elles-mêmes affectées par la démographie et les pressions évolutives. Les mesures multilocus rendent ainsi compte de la dynamique des populations dans un contexte multilocus.

L'IBD est un objet théorique difficile à étudier (Stam, 1980; Chapman et Thompson, 2003; Ball et Stefanov, 2005), nous essaierons ainsi de la traiter le plus simplement possible dans le chapitre suivant (à travers les blocs IBD). Les données génétiques comportent rarement les informations concernant directement l'IBD, et est ainsi souvent étudiée à travers ses approximations, comme l'EHH ou le ROH (Albrechtsen et al., 2009; Browning et Browning,

2012; Park et al., 2015; Wang et al., 2017; Li et al., 2017). Dans cette thèse, nous étudierons uniquement l'IBD en soi, et non ses nombreuses approximations, mais il est important de garder à l'esprit qu'un travail complet devrait aborder la question de l'applicabilité aux données réelles de l'IBD, et à défaut de ses approximations. En effet, nous ne développons pas des mesures multilocus uniquement pour des études théoriques (comme ici), mais parce que de nouvelles données multilocus sont disponibles et à interpréter. Grâce au développement de nouvelles techniques, comme celles de Next Generation Sequencing (Mardis, 2008; Behjati et Tarpey, 2013), les données actuellement disponibles ne sont plus celles d'antant : elles comportent une densité importante de marqueurs génétiques, avec éventuellement une connaissance précise quant à la distance (en Morgan) séparant les marqueurs. Développer des mesures multilocus est nécessaire pour pouvoir interpréter de telles données sans perte d'information. Réfléchir à l'applicabilité de l'IBD est ainsi l'étape suivante de notre travail et est primordiale.

L'IBD reste néanmoins très difficile à analyser mathématiquement. Nous avons donc essayé de simplifier le concept, tout en gardant l'aspect multilocus, en développant le concept de cluster, ou l'ensemble des régions qui ont la même origine ancestrale. Nous verrons dans le chapitre sur les clusters qu'un tel concept permet une analyse méthodologique autant dans les approches dites "forwards" (Chapman et Thompson, 2002*a*) que dans les approches dites "backwards" en utilisant les graphes de recombinaison ancestraux – ou ARG pour Ancestral Recombination Graph. Dans ce chapitre, nous serons en mesure de pousser l'étude plus loin que dans le cas de l'IBD et d'analyser numériquement le comportement des clusters dans différents modèles de population.

Dans un troisième projet court, nous avons tenté d'illustrer une approche purement numérique et méthodologique, sans prédiction mathématique, de l'autocorrélogramme de l'hétérozygotie. Nous voulions dans cette thèse une illustration d'un projet méthodologique avancé, facile à utiliser et à mettre en place, tout en restant puissant dans la détection de sélection, contrairement aux deux précédents projets pour lesquels la construction des objets – la modélisation – était une part importante du travail.

Chapitre 2

Identité par descendance

2.1 Introduction à l'article

Dans ce chapitre, nous étudierons l'évolution de l'identité par descendance (ou IBD) au cours des générations dans les modèles de population – ceci est une introduction à l'article qui suit (Tiret et Hospital, 2017). L'IBD est à la base un concept défini entre deux allèles à un locus donné (deux allèles sont dits IBD s'ils sont deux copies héritées du même allèle ancestral), mais dans cette partie nous nous sommes concentrés sur sa définition étendue : deux haplotypes homologues (c'est-à-dire deux ensembles d'allèles voisins, chaque ensemble étant sur un chromosome) sont dits IBD s'ils sont des copies héritées du même haplotype ancestral – concept que nous avons appelé IBD multilocus. Nous nous sommes intéressés à l'IBD multilocus, car elle est une façon d'étudier la transmission génétique dans un contexte multilocus, comme mentionné précédemment.

L'IBD est toujours définie relativement à une population initiale – ou fondatrice, qui, dans notre cas, est supposée être composée de chromosomes

tous différents. Ainsi, au bout de plusieurs générations, à cause du processus de recombinaison, les chromosomes seront constitués de blocs d'origines ancestrales différentes (voir la Figure 1 de l'article), dont les bordures de blocs sont des crossovers ou les bordures de chromosome. La dynamique des bordures de ces blocs a été étudiée dans Chapman et Thompson (2002*a*). Dans notre cas, nous avons considéré des paires de chromosomes homologues, et étudié les blocs IBD – c'est-à-dire les blocs homologues d'origines ancestrales communes. Nous l'avons fait en étudiant la dynamique des bordures de ces blocs IBD, appelées jonctions, et utilisé la théorie sous-jacente Fisher (1949, 1954).

Dans ce projet, nous avons choisi d'étudier la taille des blocs IBD comme statistique multilocus. Comme décrit précédemment, cette taille permet de rendre compte de la transmission génétique dans un contexte multilocus. Nous avons effectué une recherche bibliographique sur l'étude des blocs IBD à travers les processus de marches aléatoires sur les hypercubes (Ball et Stefanov, 2005; Martin et Hospital, 2011), mais notre travail principal a été de développer une prédiction théorique de la taille moyenne des blocs IBD avec un processus de renouvellement – prédiction principalement basée sur celle de Stam (1980). Nous avons dit précédemment qu'un modèle de population stochastique décrivait un espace probabilisé, et ce sans vraiment l'illustrer ; dans l'article qui suit, nous avons explicitement formulé cet espace probabilisé pour pouvoir déduire mathématiquement l'espérance de la taille des blocs IBD, comme suit :

$$\mathbb{E}(L) = \frac{\mathbb{E}(D)}{\mathbb{E}(K)} \tag{2.1}$$

où L est la longueur d'un bloc IBD, D la longueur totale des blocs IBD sur un individu et K le nombre de blocs IBD sur un individu. Cette relation est vraie pour toute t-population et pour tout modèle de population de type Wright-Fisher (à quelques conditions près, spécifiées dans l'article). Comme illustré dans l'article, cette formule est généralisable à presque tous les schémas de reproduction, il suffit de connaitre $\mathbb{E}(D)$ et $\mathbb{E}(K)$ pour ces schémas.

Nous avons illustré cette formule générale avec un modèle de population de taille constante, avec des générations non-chevauchantes, sans autre pression évolutive que la dérive, composée d'individus diploïdes à deux parents, avec un schéma d'accouplement panmictique sans autofécondation et le schéma de méiose décrit dans (1.6), à savoir le modèle de recombinaison de Haldane. La population initiale est composé d'individus tous différents, c'est-à-dire qu'à chaque locus il y a initialement $2N$ allèles différents dans la population. Les chromosomes sont modélisés comme des objets continus de longueur finie. Ce modèle de population est celui qui a été utilisé dans les travaux de Stam (1980) et Chapman et Thompson (2003), et cet article ayant pour objectif de les étendre nous avons voulu utiliser le même modèle. C'est d'ailleurs ce modèle qui sera utilisé pour générer les pseudo-données utilisées dans l'analyse. Le principal intérêt de ce modèle de population est que l'on connait $\mathbb{E}(D)$ et $\mathbb{E}(K)$ au cours du temps d'après Stam (1980), ce qui nous permet de connaitre, d'après l'équation (2.1), $\mathbb{E}(L)$ au cours du temps ; en d'autres termes une prédiction de l'évolution de la taille moyenne des blocs IBD. Les formules exactes de $\mathbb{E}(D)$ et de $\mathbb{E}(K)$ sont dans Stam (1980).

Pour pouvoir tester cette prédiction, nous avons implémenté un programme en C++ (compilé avec GCC v5.1), permettant de simuler des populations selon le modèle de population décrit ci-dessus, et ainsi de produire les

pseudo-données nécessaires à la mesure de taille de blocs IBD. Le schéma de reproduction pour générer la (t+1)-population a été modélisé comme suit : un couple d'individus est tiré au hasard dans la t-population, et un gamète est généré par individu en suivant le modèle de recombinaison de Haldane. Ainsi, nous obtenons un individu enfant et diploïdes. Le programme génère les individus N fois, de manière que la taille de population reste constante. Il est à noter que l'équation (2.1) est valable pour tous les modèles démographiques, nous avons ici juste choisi de l'illustrer à travers un modèle simple et courant. Les hypothèses que pose ce modèle sont effectivement fortes, et il serait avisé de les évaluer avant d'appliquer ce modèle sur de vraies données.

Nous nous sommes rendus compte après l'implémentation d'un tel programme que considérer plusieurs locus ouvrait une dimension supplémentaire, et par conséquent qu'il n'y avait pas "une" mesure de cette taille moyenne, mais trois. Notre formule, basée sur les travaux de Stam (1980), nous a permis de prédire très précisément l'une des mesures (notée L_{AR} dans l'article).

2.2 Extension de l'article

Cette partie concerne l'extension de l'article qui n'a pas pu être publié. Les points précis sont dans l'article, mais sa lecture n'est pas nécessaire pour comprendre cette partie. Les principaux travaux menés en aval de la publication ont porté sur les deux mesures qui n'ont pas pu être prédites dans le cadre de l'article. Il est difficile de savoir à quoi correspond chaque mesure et comment les interpréter, mais nous avons néanmoins précisé dans l'article les utilisations possibles selon le mode d'échantillonage des blocs. Par ailleurs,

nous pouvons voir que l'une des mesures (notée L_{IW} dans l'article) converge, d'après la loi des grands nombres, vers l'espérance de la taille des blocs IBD si l'on suppose pour une génération donnée que cette taille ne dépend pas de l'individu dans lequel se trouve le bloc, ni de sa position sur le chromosome. Il est donc contre-intuitif de voir que ce n'est pas cette mesure qui converge vers notre prédiction ; malgré tout, il est possible que cette mesure converge vers ladite prédiction en considérant certaines valeurs de paramètres du modèle de population.

À défaut de pouvoir prédire ces mesures, nous avons donc voulu savoir si la différence de valeurs entre ces mesures existait pour tous les paramètres ; autrement dit, existe-t-il des paramètres (même asymptotiques) qui rendent ces trois mesures identiques, ce qui permettrait grâce, à notre formule, d'en avoir une prédiction ? Nous avons ici fait varier deux paramètres : la longueur de chromosome et la taille de population.

Nous avons d'abord considéré la longueur de chromosome. Ce paramètre nous semblait intéressant, dans le sens où si sa valeur augmente, l'hypothèse selon laquelle la taille des blocs IBD ne dépend pas de sa position sur le chromosome semblait de plus en plus valide (par stationnarité) ; et ainsi, nous verrions L_{IW} converger vers notre prédiction. Nous voyons dans les Figures 2.1, 2.2, 2.3 et 2.4 qu'effectivement, augmenter la longueur du chromosome (bien que complètement irréaliste) rapproche les trois mesures.

Ensuite, nous avons fait varier la taille de population. Nous voyons dans la Figure 2.5 que la convergence entre les mesures L_{IW} et L_{AR} (L_{PW} étant toujours entre les deux autres, nous l'avons pas tracée) n'est pas évidente. Par contre, ce que nous pouvons voir, en combinaison avec les précédentes figures, est que l'augmentation de la taille de population ou de la longueur de

chromosome s'accompagne d'une augmentation de la durée pendant laquelle la taille moyenne des blocs IBD est presque nulle. Durant cette période, les trois mesures sont très proches. Fort de ce constat, nous formulons le postulat que dans une population idéale de taille infinie, ou de taille finie mais avec des chromosomes de longueur infinie, les trois mesures sont identiques et égales à notre prédiction. La question dans les études futures sera de savoir à quel point ce postulat est valide dans des populations réelles avec des tailles finies et des longueurs de chromosome finies.

Une autre mise à jour de notre travail serait de considérer la distribution des tailles de blocs IBD en lieu et place de la moyenne, ce qui permettrait d'implémenter des tests statistiques de neutralité par exemple. La distribution est autrement plus complexe à considérer dans le sens où il y a un bord absorbant : dans une population de taille finie, au bout d'un certain temps tous les chromosomes deviennent identiques (la population est dite fixée), ce qui implique que la taille moyenne de blocs IBD vaut la longueur du chromosome, comme nous pouvons le voir sur la Figure 2.1. Il nous faudra donc considérer des distributions conditionnelles au bord absorbant.

De plus, nous pouvons ici discuter de l'applicabilité de telles mesures sur de vraies données en s'interrogeant premièrement sur le type de données disponibles et des difficultés techniques potentielles : dans nos modèles l'origine parentale des allèles est parfaitement connue, les haplotypes n'ont pas besoin d'être phasés... Pour utiliser de vraies données, il faudra approximer l'origine parentale soit par la séquence (Identity by State), soit par des chaînes de Markov cachées (Browning et Browning, 2012). Par ailleurs, le nombre de Single Nucleotide Polymorphisms (ou SNPs) disponibles par chromosome dans les données nuancerait grandement l'applicabilité de l'équation (2.1) de

notre article, puisque nous y supposons que les chromosomes sont des objets continus, portant une infinité de locus.

Enfin, nous ne pouvons directement modifier l'article mais nous tenons à fournir quelques précisions pour clarifier certains passages de l'article :

— La légende de la Figure 2 serait plus claire comme suit : *Comparing the different measures* L_{AR}, L_{IW} *and* L_{PW} *in lines and the* values *of equation (11) in dots. These values were obtained from simulations of a population of* $N = 20$ *diploid individuals, with a chromosome length of* $l = 1$ *Morgan, over 500 generations. 1,000,000 replicates were simulated.*

— Page 12, le deuxième paragraphe serait plus clair comme suit : *In equation (10),* $\mathbb{E}(K)$ *could be seen as half of the number of IBD block edges, i.e. half of the number of external junctions over* l *Morgans plus half of the number chromosome tips for which a fraction* $(1 - \mathbb{E}(H))$ *is IBD (knowing that each point of the chromosome has a probability* $1 - \mathbb{E}(H)$ *to be IBD).*

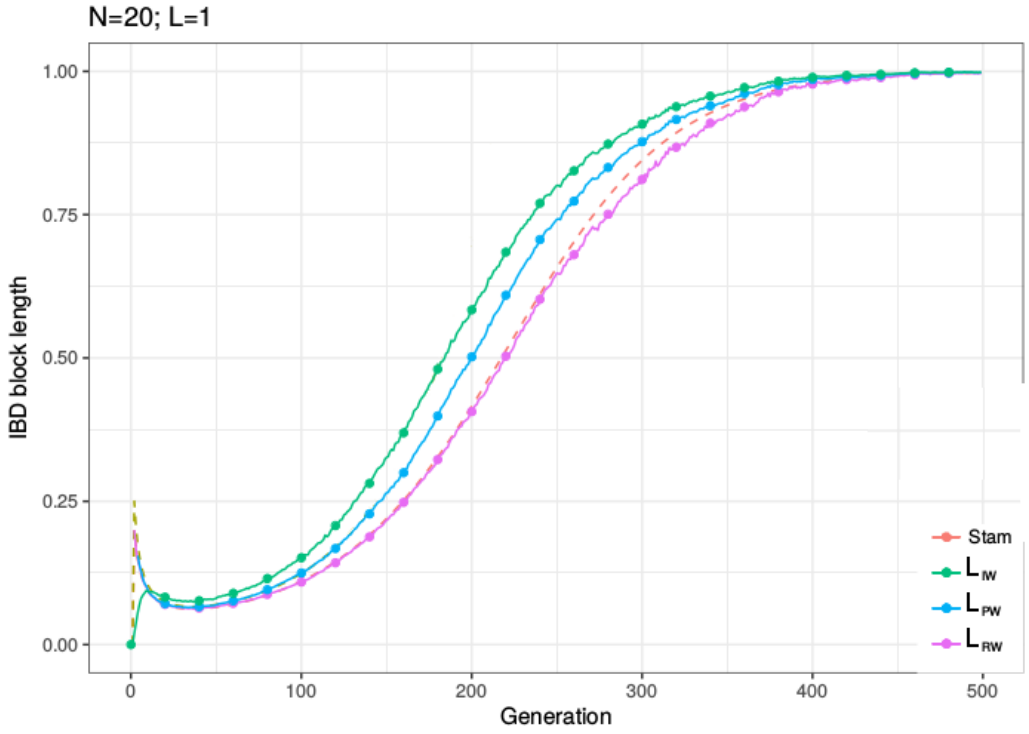

FIGURE 2.1 – Longueurs moyennes de blocs IBD au cours des générations selon les trois mesures L_{IW}, L_{PW} et L_{AR}, et notre prédiction (notée "Stam"). La taille de population est de 20 individus diploïdes, et la longueur de chromosome de 1 Morgan, sur 10,000 simulations.

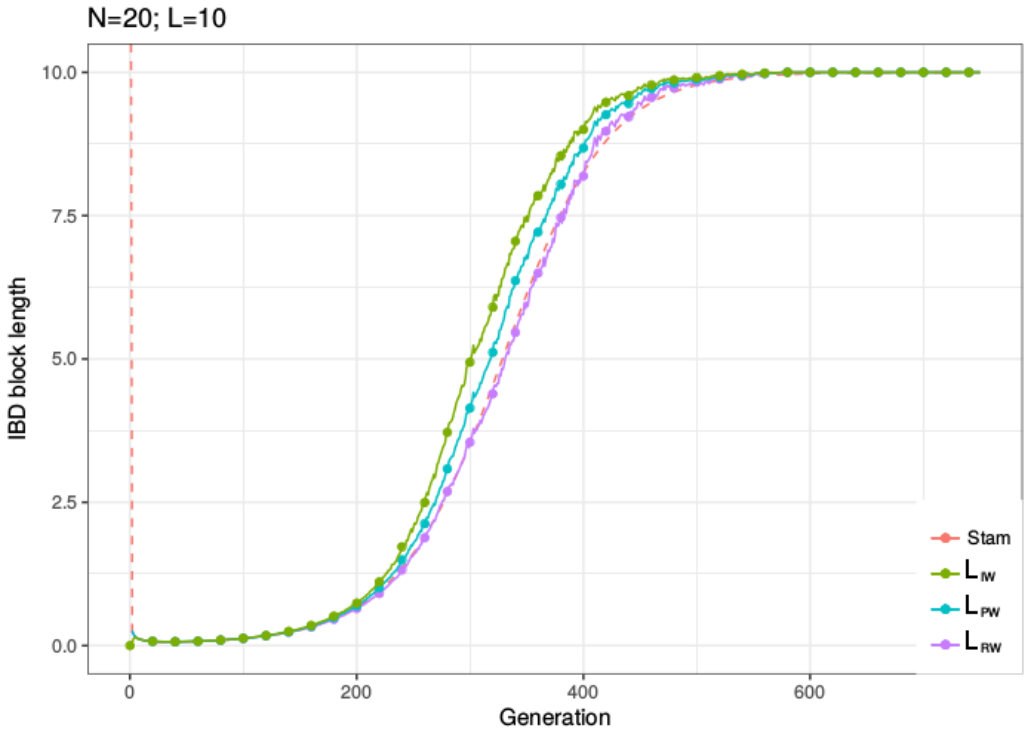

FIGURE 2.2 – Idem pour une longueur de chromosome de 10 Morgans.

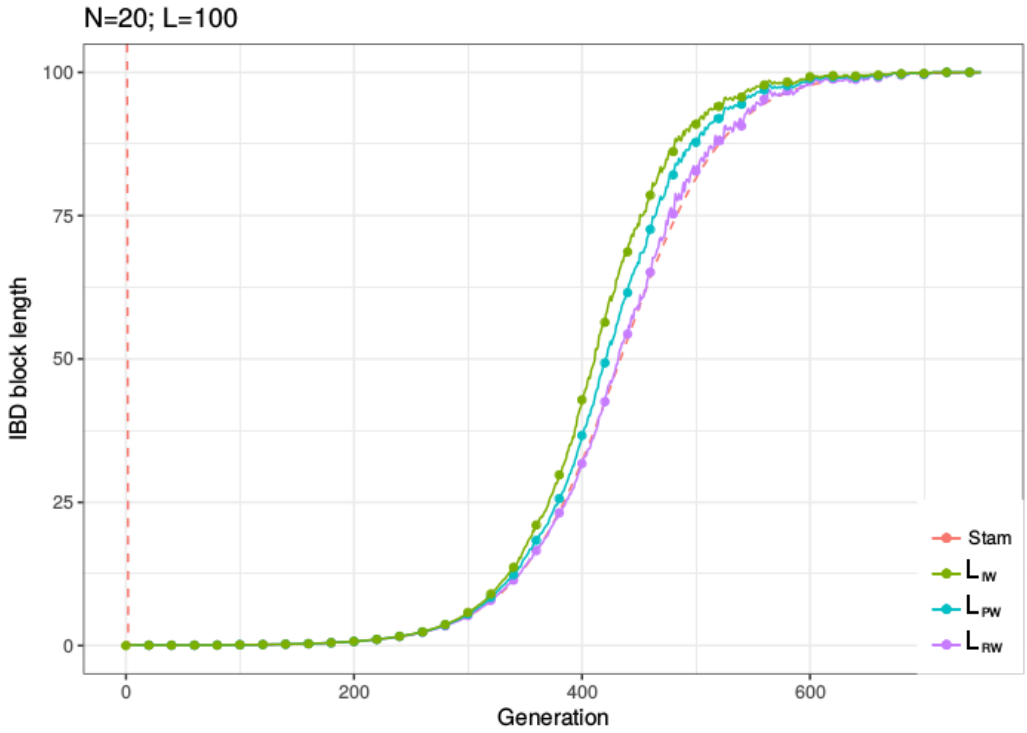

FIGURE 2.3 – Idem pour une longueur de chromosome de 100 Morgans.

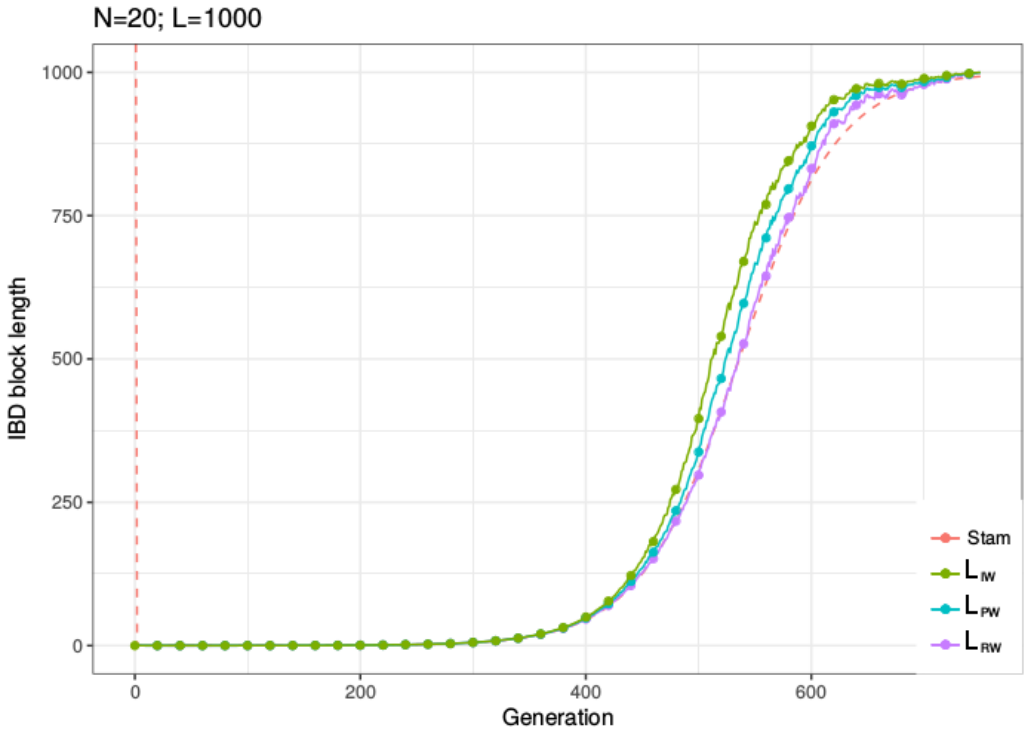

FIGURE 2.4 – Idem pour une longueur de chromosome de 1000 Morgans.

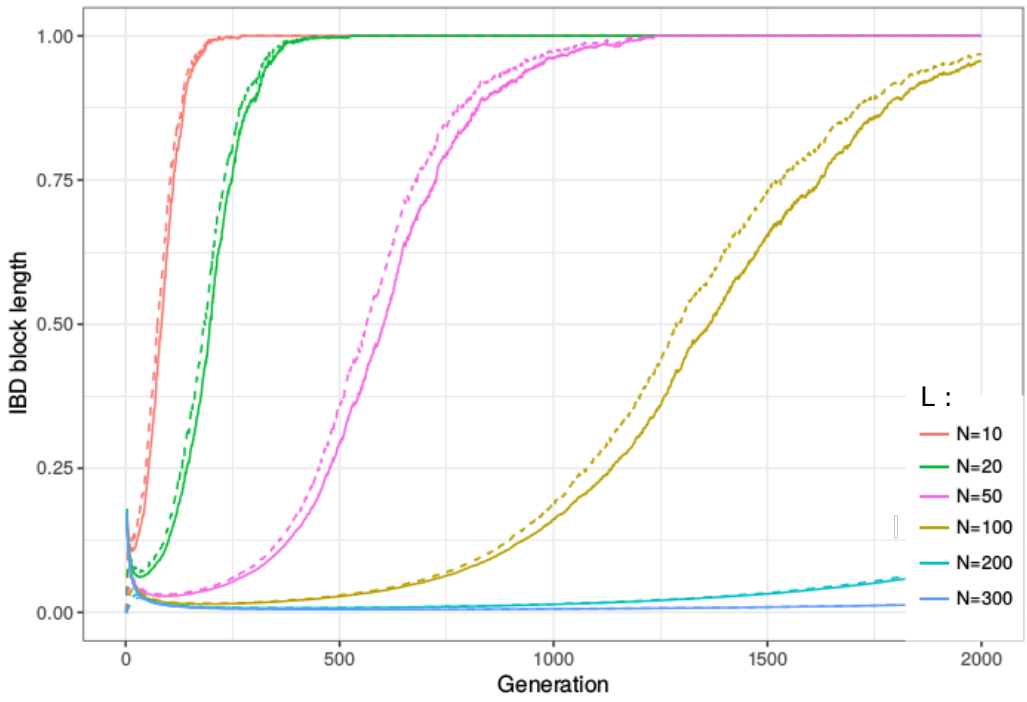

FIGURE 2.5 – Longueurs moyennes de blocs IBD au cours des générations selon les deux mesures L_{IW} (en pointillé) et L_{AR} (en ligne). Plusieurs tailles de population N ont été testées (comme indiqué dans la légende), avec une longueur de chromosome de 1 Morgan, sur 10,000 simulations.

Blocks of chromosomes identical by descent in a population: models and predictions

Mathieu Tiret[1*], Frédéric Hospital[1]

1 UMR 1313 Génétique Animale et Biologie Intégrative, INRA, Jouy-en-Josas, France

* Corresponding author: mathieu.y.tiret@gmail.com

Abstract

With the highly dense genomic data available nowadays, ignoring linkage between genes would result in a huge loss of information. One way to prevent such a loss is to focus on the blocks of chromosomes shared identical by descent (IBD) in populations. The development of the theoretical framework modelling IBD processes is essential to support the advent of new tools such as haplotype phasing, imputation, inferring population structure and demographic history, mapping loci or detecting signatures of selection. This article aims to present the relevant models used in this context, and specify the underlying definitions of identity by descent that are yet to be gathered at one place. In light of this, we derived a general expression for the expected IBD block length, for any population model at any generation after founding.

1 Introduction

Two alleles are said to be identical by descent (IBD) if they are inherited copies of the same ancestral allele. In the past, IBD was mostly studied at one locus or a few independent loci. Nowadays, with the advent of Next Generation Sequencing techniques, new models and concepts integrating several loci at once ('multilocus IBD') have become prominent in genome scan analyses. The idea is

to take full account not only of the high number of available marker loci, but also of their high density per genome length (in Morgan). In such analyses, linkage and linkage disequilibrium can no longer be ignored as was the case in the past with scarcer maps. Indeed, integrating haplotype information in genome scan analyses adds value to multilocus IBD studies (Browning and Browning, 2012). In this paper, we will focus on IBD blocks of chromosomes, or contiguous IBD loci, and thereby account for linkage between loci. Note that it is also possible to study probabilities of several disruptive loci to be IBD (Hill and Hernández-Sánchez, 2007; Hill and Weir, 2011), but this is a different approach of multilocus IBD that will not be considered here.

Developing the theoretical framework underlying IBD processes has become essential for the development of new tools suitable for high density genomic data, such as haplotype phasing and imputation (Carmi et al., 2013), inference of population structure and demographic history (Carmi et al., 2013; Palamara et al., 2012), mapping loci or detecting signatures of selection (Ødegård and Meuwissen, 2014; Kardos et al., 2017).

In the literature, several alternative definitions of an IBD block exist. We will first try to properly define the concepts and clarify implicit considerations for each definition. Then, we will present some of the relevant models used to study IBD blocks in a population. Practical applications of these models were thoroughly reviewed in Browning's article (Browning and Browning, 2012).

1.1 Diversity of definitions

From here onwards, let us call a 'locus' a common position over a set of n homologous chromosomes, and a 'segment' a set of adjacent loci. The concept of IBD is always relative to a founder population. It could be defined for k loci over n homologous chromosomes. It has already been thoroughly defined at one locus ($k = 1$) for any number of homologous chromosomes, and we are trying here to define it properly for a segment, for any k and any n. Paraphrasing some articles of the literature on IBD studies (Chapman and Thompson, 2003; Clark, 2004; Ball and Stefanov, 2005; Browning, 2008), we suggest in this article

that n homologous tracts of chromosomes are IBD if they are inherited copies of the same ancestral homologous tract of a chromosome. By the definition of segment, we are only considering homologous chromosomes, excluding transposable elements. Specifying that they are 'inherited' excludes horizontal gene transfer.

Identity by descent is a powerful concept with which it is possible to describe how genetic material is transmitted or lost over time. Assuming that genetic material could be split into a 'container' and a 'content', studying the containers independently of the content is a matter of IBD. On the other hand, studying the content is a matter of identity by state (IBS), not of IBD. Therefore, everything that concerns the content, namely the sequence, such as IBS or mutations, is not accounted for here: they are issues of allelic variation, not of descent. One should account for mutations only when approximating IBD through IBS. On the contrary, recombination events have to be taken fully into account. In this paper, we will not be considering crossovers among non-homologous chromosomes. There are two types of crossovers: those that occur between two tracts that are IBD and thus invisible; and the others that are called 'junctions' (Fisher, 1949, 1954). Describing and predicting the dynamics of junctions is a core part of IBD studies.

Furthermore, we could distinguish two types of multilocus IBD, relaxed or strict. Relaxed IBD at a segment is a relation between n homologous chromosomes that are IBD at every locus of the segment, each locus being not necessarily of the same ancestral origin as its adjacent loci. Strict IBD requires that in addition the n homologous chromosomes have the same ancestral origin at each locus of the segment.

When considering n homologous chromosomes, one could project on an axis whether or not these chromosomes are IBD for each locus. This axis is here called the IBD axis (see Fig 1). On this axis, we could clearly distinguish IBD tracts and non-IBD tracts. A junction is external if its projection on this axis delimits an IBD and a non-IBD tract, and is internal if its projection is within an IBD or a non-IBD tract. We define a relaxed IBD block as a contiguous

IBD tract delimited by external junctions or tips of chromosomes, without any external junction in it. In addition, strict IBD blocks are also delimited by internal junctions that are within IBD tracts. There is no junction in a strict IBD block. Depending on the definition, there could be a different number of IBD blocks, as can be seen in the example in Fig 1, on which there is either one relaxed IBD block or two strict IBD blocks. Hereafter, we only consider relaxed IBD.

1.2 Modelling choices

In the literature, only two values of n were studied, 2 and the population size N (or $2N$ for diploid populations), although intermediate values of n could be considered as well. When $n = 2$ in a diploid population, some models focus on pairs of homologous chromosomes within individuals, and IBD is then called 'homozygosity by descent' (Franklin, 1977; Stam, 1980), and some on random pairs of homologous chromosomes in a population (Chapman and Thompson, 2003).

For all the definitions presented above, the locations on the chromosome could either be modelled as continuous (Chapman and Thompson, 2003; Ball and Stefanov, 2005; Stam, 1980) or as discrete objects (Bickeböller and Thompson, 1996). In fact, all these cited papers treat the underlying chromosome as a continuum, but some maybe model transitions in IBD state from (discrete) locus to locus – as is natural to do when dealing with actual data at marker loci (anonymous referee, personal communications).

Genome length is measured in Morgan and crossovers are usually supposed to follow the no-interference recombination model of Haldane (Haldane, 1919): at each meiosis, each chromosome of length l (in Morgan) undergoes crossovers, whose number follows a Poisson law of parameter l and whose positions are independent random variables each with a uniform distribution. Therefore, the crossover events follow a Poisson process of rate 1 in the Haldane recombination model. As long as the Haldane recombination model is valid, measuring the genome length in Morgans as in every article cited here, or studying the

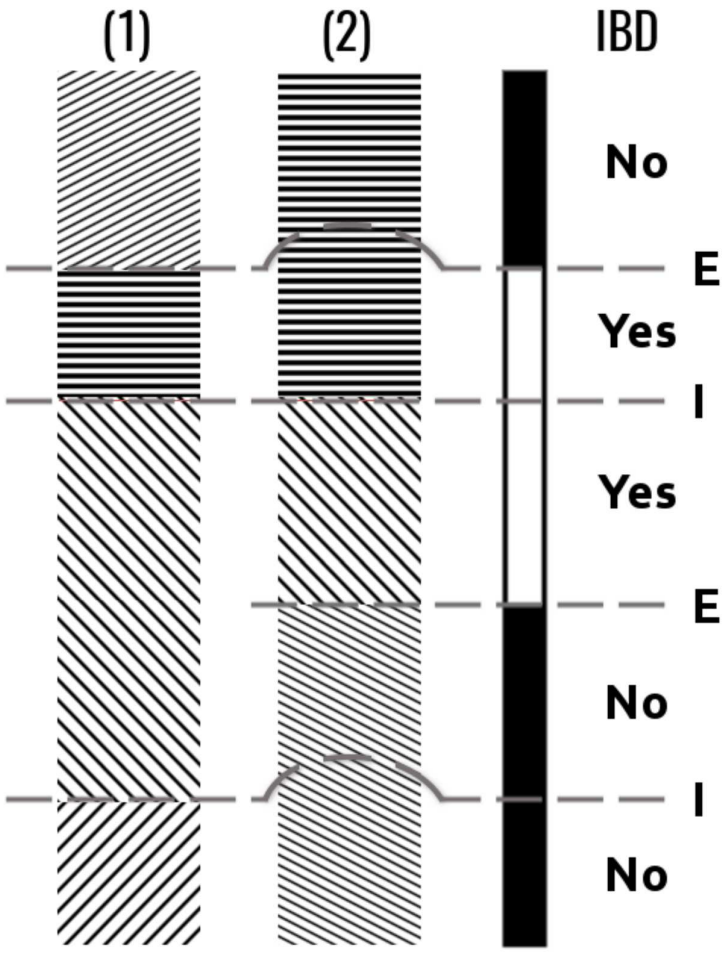

Fig 1: Two homologous chromosomes, labelled '(1)' and '(2)', some generations after founding. Different patterns on the chromosomes represent the different ancestral origins. The third axis, labelled 'IBD', is the IBD axis on which white parts indicate the IBD tracts, and black parts the non-IBD tracts. Each junction is projected on this axis and labelled 'E' if it is an external junction, and 'I' if it is an internal junction. When considering the relaxed IBD, there is only one IBD block, whereas when considering the strict IBD, there are two of them.

consequences of variation in recombination rate along the chromosome (Knief et al., 2017) are strictly equivalent.

One of the major problems of this field is a proper prediction of how IBD evolves over time in a population. There are several ways to quantify IBD in a population, the most important ones, considering n homologous chromosomes, being the number of IBD blocks, the length of one IBD block, and the total length of IBD blocks over these n homologous chromosomes. This paper extends previous studies on the evolution over time of the distributions of these quantities, or of their moments, in stochastic models of population genetics (Palamara et al., 2012; Chapman and Thompson, 2003; Fisher, 1949; Stam, 1980; Browning and Browning, 2002; Martin and Hospital, 2011). The difficulty lies in the accumulation of junctions and the merging of IBD blocks over time. In the next section, we will review two major types of forward models, either based on random walks, or on renewal processes.

2 Models

2.1 Random walk on a hypercube

Considering IBD shared among $n = 2N$ homologous chromosomes inherited from two different founder chromosomes only (denoted 0 and 1), it is possible to derive the true distribution of the relevant quantities of multilocus IBD as follows. One of the relevant quantities we will be focusing on is the total length of IBD blocks over the chromosome, or 'total IBD length'. At each locus, one chromosome is denoted 0 or 1 depending on which founder it originates from. At each locus, the population of n homologous chromosomes is hence a n-tuple of 0's and 1's. Furthermore, we assume the continuous model of a chromosome, so that there is an infinite number of possible positions on a chromosome where a crossover could occur. Therefore, new crossover has a zero probability to occur in a location of another existing crossover. In a process whose states are the n-tuples of 0's and 1's and the time parameter is the map distance along the chromosomes, at most one coordinate of the n-tuple changes

at each position, because of the continuous model. This process may thus be modelled as a realisation of a particular Markov process, namely a continuous-time Markov random walk on the vertices of a n-hypercube. Only two vertices of the hypercube are of interest, (0,...,0) and (1,...,1) which correspond to the states in which the population is IBD. The other vertices are the non-IBD states. Donnelly (Donnelly, 1983) first considered this problem and succeeded to reduce the dimension of the problem by gathering the vertices in what he called orbits, and provided the corresponding transition rate matrix.

Ball and Stefanov (2005) used this theoretical framework to derive the exact characteristic function of the total IBD length among half-sibs, assuming that the number of non-IBD blocks was Poisson distributed. From this work, it is possible to deduce the exact probability of survival of the parental genetic material over one generation. Walters and Cannings (2005) provided a general method for finding the density of the total IBD length, that could be applied to any unilineal relationship, and more specifically they provided the density of total IBD length for a grandparent-grandchild relationship.

Martin and Hospital (2011) considered a particular lineage of recombinant inbred lines of 2 or 4 homologous chromosomes undergoing generations of respectively self-crossings or full-sib matings. In this model, each point of a chromosome is denoted 0 and 1 depending on which chromosome of the previous generation it inherited from (Stefanov, 2000). Considering g generations, and modelling the chromosome as a continuum, the problem could also be modelled as a random walk on a g-hypercube. With this model, Martin and Hospital (2011) studied the distribution of the length of IBD blocks depending on their positions on a semi-infinite chromosome, and showed that the successive blocks are almost independent and that the block at the origin of the chromosome was larger than the others. This counter-intuitive result is mainly due to the non-exponential distribution of block lengths (Martin and Hospital, 2011, see eq. 5 and surrounding text), and could also be observed on finite length chromosomes (data not shown).

Using a random walk, it is possible to derive the distribution of IBD quanti-

ties, but only when assuming very particular pedigrees. In the next section, we will present theoretical results for a more general population model albeit only means have been accurately derived so far.

2.2 Renewal process in a random mating population

In this section, we will study the evolution over time of the relaxed IBD shared among pairs of homologous chromosomes ($n = 2$) in any kind of diploid population descending from a founder population. Without loss of generality, we will hereafter focus on pairs of homologous chromosomes within individuals, or homozygosity by descent, and provide an expression of the expected length of IBD blocks. The length and the number of IBD blocks per chromosome are not independent, and this dependency is very difficult to handle. Therefore, we have tried to develop a workaround by using quantities that are not affected by this dependency.

Let $\mathcal{P}$ denote the set of all possible populations of a stochastic or deterministic population model $\mathcal{M}$. To any population $p \in \mathcal{P}$, the model also assigns a probability $\mathbb{P}(p)$, which is the probability of encountering this population. One population is constituted of individuals, all carrying zero or more IBD blocks, so that a population is both a set of individuals and a set of IBD blocks. In other words, the model also assigns probabilities, indirectly though, to all possible individuals and all possible IBD blocks.

Let us now consider that every population $p \in \mathcal{P}$ has the same number N of individuals. For a given population p, the fact that an individual i is within this population is denoted $i \in I_p$, and similarly, that an IBD block b is carried by an individual within this population is denoted $b \in B_p$. For any individual $i \in I_p$, we denote d_i its total IBD length and k_i the number of IBD blocks it carries. We also denote $m_p = \sum_{i \in I_p} k_i$ the total number of IBD blocks in the population p. For any IBD block $b \in B_p$, we denote l_b its length. Let X be an IBD block randomly drawn from $\cup_{p \in \mathcal{P}} B_p$, and L its length. We are interested in deriving the expected length $\mathbb{E}(L)$ of a randomly drawn IBD block. If $\mathcal{P}^*$ is the set of populations in which there is at least one IBD block, we have:

$$\begin{aligned} \forall p \in \mathcal{P}^*,\ \mathbb{E}(L|\, X \in B_p) &= \frac{\sum_{b \in B_p} l_b}{m_p} \\ &= \frac{\sum_{i \in I_p} d_i}{m_p} \end{aligned} \tag{1}$$

where $X \in B_p$ means that the block X belongs to the population p. The population p was drawn through sampling a block, and is then size-biased: populations do not have the same number of IBD blocks, therefore sampling a block is not an unbiased way of drawing a population. One could then state that:

$$\forall p \in \mathcal{P}^*,\ \mathbb{P}(X \in B_p) = \frac{\mathbb{P}(p) \cdot m_p}{\sum_{q \in \mathcal{P}^*} \mathbb{P}(q) \cdot m_q} \tag{2}$$

where q is a population of $\mathcal{P}^*$, and assuming that $\mathbb{P}(X \in B_p)$ is defined for the population model $\mathcal{M}$, or equivalently that $\sum_{q \in \mathcal{P}^*} \mathbb{P}(q) \cdot m_q$ does not diverge towards infinity. $X \in B_p$ indicates a unique population and the union of these populations, considering all the possible X, is $\mathcal{P}^*$. Therefore, using equations (1) and (2), and the law of total expectation, one could derive that:

$$\begin{aligned} \mathbb{E}(L) &= \mathbb{E}_{\mathcal{P}^*}(\mathbb{E}(L|\, X \in B_p)) \\ &= \sum_{p \in \mathcal{P}^*} \mathbb{P}(X \in B_p) \cdot \mathbb{E}(L|\, X \in B_p) \\ &= \sum_{p \in \mathcal{P}^*} \frac{\mathbb{P}(p) \cdot m_p}{\sum_{q \in \mathcal{P}^*} \mathbb{P}(q) \cdot m_q} \cdot \frac{\sum_{i \in I_p} d_i}{m_p} \\ &= \frac{\sum_{p \in \mathcal{P}^*} \mathbb{P}(p) \cdot \sum_{i \in I_p} d_i}{\sum_{q \in \mathcal{P}^*} \mathbb{P}(q) \cdot \sum_{i \in I_q} k_i} \end{aligned} \tag{3}$$

In parallel, let Y be an individual randomly drawn from $\cup_{p \in \mathcal{P}} I_p$, D its total IBD length and K the number of IBD blocks it carries. One could trivially state that:

$$\forall p \in \mathcal{P},\ \mathbb{P}(Y \in I_p) = \mathbb{P}(p) \tag{4}$$

where $Y \in I_p$ means that the individual Y belongs to the population p. This population p was drawn through sampling an individual, therefore there is no size-bias, because all populations in $\mathcal{P}$ have the same number of individuals N. Also, one could derive that:

$$\mathbb{E}(D|\, Y \in I_p) = \frac{\sum_{i \in I_p} d_i}{N} \tag{5}$$

$$\mathbb{E}(K|\, Y \in I_p) = \frac{\sum_{i \in I_p} k_i}{N} \tag{6}$$

Finally, using all the above, one obtains:

$$\begin{aligned} \mathbb{E}(L) &= \frac{\sum_{p \in \mathcal{P}^*} \mathbb{P}(Y \in I_p) \cdot \sum_{i \in I_p} d_i/N}{\sum_{q \in \mathcal{P}^*} \mathbb{P}(Y \in I_q) \cdot \sum_{i \in I_q} k_i/N} \\ &= \frac{\mathbb{E}(D) - \sum_{p \in \mathcal{P} \setminus \mathcal{P}^*} \mathbb{P}(Y \in I_p) \cdot \mathbb{E}(D|\, Y \in I_p)}{\mathbb{E}(K) - \sum_{q \in \mathcal{P} \setminus \mathcal{P}^*} \mathbb{P}(Y \in I_q) \cdot \mathbb{E}(K|\, Y \in I_q)} \\ &= \frac{\mathbb{E}(D)}{\mathbb{E}(K)} \end{aligned} \tag{7}$$

where $\mathcal{P} \setminus \mathcal{P}^*$ is the set of populations in which there are no IBD blocks, and knowing that $Y \in I_p$ indicates a unique population and that the union of all these populations, considering all the possible Y, is $\mathcal{P}$.

Equation (7),which is the key point of this article, is valid at any time t after founding, for any diploid population model and for any chromosome model (continuous or discrete). The only assumptions are that all populations have the same size at generation t and that $\sum_{q \in \mathcal{P}^*} \mathbb{P}(q) \cdot m_q$ does not diverge towards infinity. In other words, if the population size is only dependent on generation t, then equation (7) is independent of any demographic structure of the population (subdivision in one or several demes, constant population size or not, panmictic or not...), and also of any evolutionary pressure (any kind of selection, any migration rate, recessive deleterious load...): it is up to $\mathbb{E}(D)$ and $\mathbb{E}(K)$ to handle these dependencies. Equation (7) could also be extended to any number n of homologous chromosomes, the only difference being that Y would be a

randomly drawn n-tuple of homologous chromosomes. These n homologous chromosomes should however be all in the same population. Equation (7) is therefore of a very powerful and general use.

Let us now derive the expressions of $\mathbb{E}(D)$ and $\mathbb{E}(K)$ for some population models. Let $\mathbb{E}(H)$ and $\mathbb{E}(Z)$ be respectively the expected non-IBD proportion of a randomly drawn individual (ranging from 0 to 1) and the expected number of external junctions per Morgan within a randomly drawn individual.

In his seminal work, Stam (1980) studied the relaxed IBD in a population and provided an approximation of $\mathbb{E}(H)$ and the exact value of $\mathbb{E}(Z)$. Stam's $\mathbb{E}(Z)$ is so far the only quantity that successfully integrates the accumulation of junctions through time in a whole population. He considered a panmictic monoecious diploid population without selfing and undergoing drift only. The founder population was assumed to be entirely constituted of unrelated and non-inbred individuals (i.e. none of the chromosome pair was IBD). He modelled the chromosomes as continuous objects, and assumed the recombination model of Haldane.

In the second part of his article, Stam (1980) found that the expected length L^* of an IBD block would be expressed as follows:

$$L^* = \frac{1 - \mathbb{E}(H)}{0.5 \cdot \mathbb{E}(Z)} \tag{8}$$

assuming that IBD and non-IBD block lengths were exponentially distributed each with its own parameter. Chapman and Thompson (2003) extended Stam's work and found the same result as equation (8), without assuming exponential distributions of the block lengths. Stam (1980) explicitly assumed stationarity of the IBD process. Though not explicitly assuming stationarity, Chapman and Thompson (2003) used equation (7.3) from Karlin's book (Karlin and Taylor, 1981, p.199), which does assume stationarity of the IBD process. Both of these articles therefore assumed stationarity, implying that the processes 'began indefinitely far in the past' (Karlin and Taylor, 1981, p.199). The x-axis of processes described in Karlin and Taylor (1981) was time, whereas the x-axis of processes studied here is the genetic map. So strictly speaking, assuming stationarity

amounts to assuming that in equation (8) the chromosome length was infinite.

If the chromosome length is assumed to be infinite, we get $\mathbb{E}(D) = 1 - \mathbb{E}(H)$ and $\mathbb{E}(K) = 0.5 \cdot \mathbb{E}(Z)$, so that our equation (7) is equivalent to equation (8). If the chromosome is however of finite length l, we use the results from Fisher (1949) to obtain the following:

$$\mathbb{E}(D) = l \cdot (1 - \mathbb{E}(H)) \quad (9)$$

$$\mathbb{E}(K) = 0.5 \cdot l \cdot \mathbb{E}(Z) + (1 - \mathbb{E}(H)) \quad (10)$$

Equation (10) corresponds to half of the number of IBD block edges, i.e. half of the number of external junctions over l Morgans plus half of the number of chromosome tips for which a fraction $1 - \mathbb{E}(H)$ is IBD. Injecting equations (9) and (10) into our equation (7), we obtain that for a chromosome of finite length l:

$$\mathbb{E}(L) = \frac{l \cdot (1 - \mathbb{E}(H))}{0.5 \cdot l \cdot \mathbb{E}(Z) + (1 - \mathbb{E}(H))} \quad (11)$$

One may wish to use the moments, of L in statistical inferences from population data, and for instance develop a neutrality test. Let us consider a pseudo-dataset obtained from simulations of the same population model as in Stam (1980). When simulating R replicates, for one generation, it is possible to measure the mean length of IBD blocks in this dataset in three ways:

$$L_{AR} = \frac{\sum_{r=1}^{R} \sum_{i=1}^{N} \sum_{j=1}^{k_{r,i}} l_{r,i,j}}{\sum_{r=1}^{R} \sum_{i=1}^{N} k_{r,i}}$$

$$L_{PW} = \frac{1}{R} \sum_{r=1}^{R} \frac{\sum_{i=1}^{N} \sum_{j=1}^{k_{r,i}} l_{r,i,j}}{\sum_{i=1}^{N} k_{r,i}} = \frac{1}{R} \sum_{r=1}^{R} L_{PW,r}$$

$$L_{IW} = \frac{1}{R} \sum_{r=1}^{R} \frac{1}{N} \sum_{i=1}^{N} \frac{\sum_{j=1}^{k_{r,i}} l_{r,i,j}}{k_{r,i}} = \frac{1}{R} \sum_{r=1}^{R} \frac{1}{N} \sum_{i=1}^{N} L_{IW,r,i}$$

where $k_{r,i}$ is the number of IBD blocks in the individual i of the replicate r and $l_{r,i,j}$ is the length of the block j in the individual i of the replicate r. L_{AR} is a measure over all the replicates and therefore we have only one value for a whole dataset. L_{PW} is the mean over the replicates of $L_{PW,r}$ that is a population-wise measure for which we have one value per population. L_{IW} is the mean over all the individuals in all the replicates of $L_{IW,r,i}$ that is an individual-wise measure for which we have one value per individual and a whole distribution per population.

On Fig 2 that shows all the different measures and prediction, we could see that L_{AR} is very close to $\mathbb{E}(L)$ of equation (11), and it is indeed easy to prove mathematically why the former tends towards the latter when the number of replicates tends towards infinity. We have therefore developed a formula, $\mathbb{E}(L)$ of equation (11), to very well predict L_{AR}, as shown on Fig 2. We could also see that these measures are different, because they are indeed all the mean lengths of IBD blocks randomly drawn, but from different samplings: L_{AR} is the mean length of an IBD block drawn from the whole pseudo-dataset; L_{PW} is of a block drawn from a randomly drawn population of the dataset; and L_{IW} is of a block drawn from a randomly drawn individual of the dataset. Since the number of IBD blocks is different in each population and each individual, these samplings, and so these measures, are different and size-biased. Similarly, we could see that the asymptotic value of L_{AR}, that is $\mathbb{E}(L)$ of equation (11), is a lower bound of L_{PW} and L_{IW}: we then have a theoretical formulation for what appears to be a lower bound of L_{PW} and L_{IW}. This relation is yet to be mathematically proven.

3 Discussion

In this paper, we have reviewed two types of forward models commonly used to study theoretically the evolution of IBD blocks of chromosomes in a population, and have shown how these models are complementary. Models based on a random walk on a hypercube are very powerful to provide exact formula about

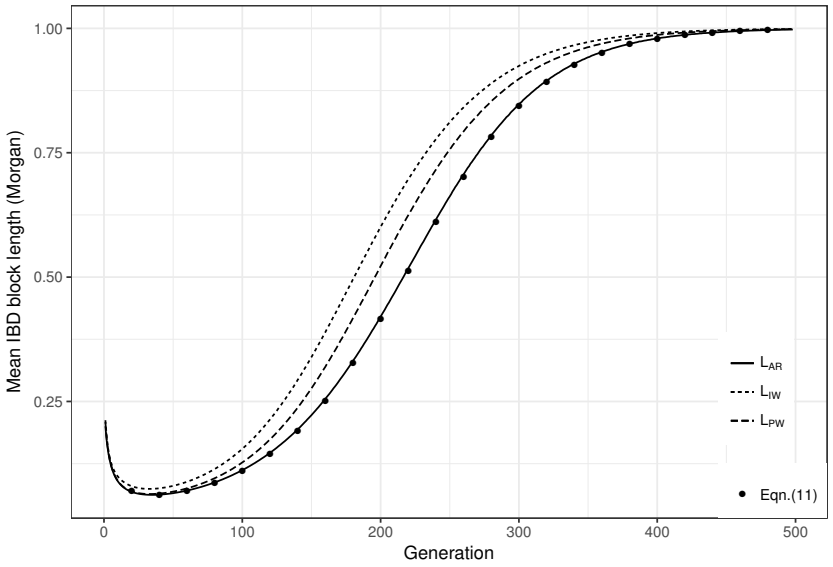

Fig 2: Comparing the different measures L_{AR}, L_{PW} and L_{IW} in lines and the prediction of equation (11) in dots. These values were obtained from simulations of a population of $N = 20$ diploid individuals, with a chromosome length of $l = 1$ Morgan, over 500 generations. 1,000,000 replicates were simulated.

the distribution of the total IBD length, but are only available for some very particular pedigrees. On the other hand, models based on a renewal process are very powerful to consider more general population models, but only means of IBD quantities have been obtained so far. We have provided a general formula for the mean IBD block length with equation (7), that is independent of the demographic structure of the population or any evolutionary pressure. It is moreover the exact asymptotic value of L_{AR}, and an asymptotic lower bound of L_{PW} and L_{IW}.

When studying real data, one should be aware of the difference between the aforementioned measures (L_{AR}, L_{PW} and L_{IW}) before developing the appropriate statistical test. If IBD blocks are sampled from one or several populations without any constraint, the appropriate measure will be L_{AR} and the corresponding prediction $\mathbb{E}(L)$. If IBD blocks are sampled from one or several populations, but drawing the same number of IBD blocks from each population, the appropriate measure will be L_{PW}. Finally, if IBD blocks are sampled

from one or several populations, but drawing the same number of IBD blocks from each individual, the appropriate measure will be L_{IW}. When there is no replicate in real data, there is no practical difference between L_{AR} and L_{PW}. Their asymptotical distributions are not the same however, so that one could develop two different tests for the same measure, and pick the most appropriate one depending on the sampling policy. When there are replicates, L_{AR} and L_{PW} are indeed different. However, apart from the sampling policy, choosing between L_{AR} and L_{PW} is arbitrary. Further studies more thoroughly describing the distributions of L_{AR} and L_{PW} should help to make this choice no more arbitrary. Finally, exact theoretical formulations of L_{PW} and L_{IW} are yet to be discovered, and therefore, further work should also focus on completing this theoretical framework to make the study of any kind of real population datasets possible.

Acknowledgements

The authors are grateful to A.Lambert, S.Boitard and two anonymous reviewers for their thorough reading and their relevant comments on the paper.

Conflict of interest

There is no conflict of interest.

Chapitre 3

Chromosome painting

Dans cette partie – et dans l'article qui suit (non encore publié), nous considèrerons une approche différente des modèles de population : jusque-là nous avons considéré les modèles sous une approche dite "forward" : comme l'axe du temps est orienté du passé vers le futur, les propriétés que nous avons précédemment énoncées permettent de générer une population future à partir d'une population présente. Cette approche intuitive a pour principal intérêt de pouvoir générer tous les types de population, mais l'analyser et la simuler computationnellement restent relativement lourds, surtout si seul l'état final nous intéresse. Si les états transitoires ne nous intéressent pas, il existe une approche bien plus économique en terme de calculs qui permet de rendre compte des états finaux et initiaux : l'approche dite "backward". Dans une telle approche, seuls nous intéressent les individus qui sont effectivement des ancêtres des individus actuellement présents, et ainsi diminuer fortement le nombre d'individus considérés et par là même la complexité du problème. Nous allons donc brièvement présenter l'approche backward et la théorie du coalescent associée, pour ensuite présenter l'article qui la traite. L'article en

question n'a pas encore été publié et est en cours de rédaction, les résultats présentés ici sont donc préliminaires.

3.1 Les approches backwards et la théorie du coalescent

En approche backward, on ne considère pas le temps du passé vers le présent, mais dans le sens inverse, du présent vers le passé : on remonte le temps à partir des individus présents aux ancêtres desdits individus. On trace ainsi l'arbre généalogique, ou plutôt phylogénétique, mais à partir des gènes et non des filiations.

En terme de modélisation, la principale différence est que l'on ne formule pas de schéma de reproduction, mais plutôt un schéma de coalescence éventuellement accompagné d'un schéma de méiose. Le but de ces modèles n'est pas d'étudier l'évolution d'une population mais de connaitre la structure de l'arbre phylogénétique des individus présents. Le schéma de coalescence est destiné à générer itérativement les nœuds de l'arbre phylogénétique, et plus précisément permet de générer un individu à partir de deux voire plus ; on a donc de moins en moins d'individu à chaque itération. Ce schéma est donc l'exact contraire du schéma de reproduction, ce qui est exactement ce que l'on veut lorsque l'on considère le temps du présent vers le passé. C'est d'ailleurs ce schéma qui garantit que la phylogénie considérée soit sous forme d'un arbre.

Le schéma de méiose est le schéma qui détermine comment obtenir deux branches de l'arbre à partir d'une seule, et est l'exact contraire du schéma de méiose détaillé dans les parties précédentes (notamment avec le modèle de

recombinaison de Haldane). Ce schéma complexifie grandement le modèle, notamment en transformant l'arbre en graphe, que l'on appelle graphe de recombinaison ancestral. Bien que complexe, il est nécessaire de considérer ce schéma si l'on veut avoir une approche multilocus. Nous nous intéresserons ici à la composition d'une t-population ancestrale, dite à stationnarité, en partant d'une certaine population du temps présent.

3.2 Les clusters et les blocs

Dans un modèle de population en génération, sans autre pression évolutive que la dérive, au bout d'un certain temps la population se fixe, c'est-à-dire que tous les individus de ladite population sont identiques. Considérons un chromosome parmi ces individus de la population fixée, ledit chromosome portera différents blocs d'origines ancestrales différentes, et par souci de simplicité nous appellerons les origines ancestrales différentes les couleurs ; en d'autres termes, ce chromosome porte des blocs de couleurs différentes. C'est en cela que nous parlons de *chromosome painting*. Nous aimerions connaitre comment les couleurs sont réparties sur ce chromosome fixé.

Reformulons ce problème en terme d'approche backward : on considère un seul chromosome du présent pour remonter dans le passé aussi loin que l'on veut. En remontant un temps donné dans le passé, on trouvera un certain nombre d'ancêtres dans ce graphe (c'est une t-population d'ancêtres). Si on colorie tous ces ancêtres d'une couleur différente chacun, le chromosome présent, mis à jour en conséquence, se révèle être une mosaïque de couleurs. Cette mosaïque a la même distribution asymptotique que la distribution de couleurs sur un chromosome fixé, mentionné ci-dessus : ces deux problèmes

sont équivalents. Nous remonterons donc dans le passé jusqu'à stationnarité, et étudier la distribution des couleurs. La disposition des couleurs, et donc des clusters, nous intéresse ici parce qu'elle nous informe sur l'état final d'une population, et peut potentiellement être une mesure de la conservation du matériel génétique.

Nous appellerons dans cette partie un bloc un morceau contigü de chromosome uniformément coloré, et un cluster l'ensemble des blocs d'une couleur. La stationnarité est mesurée en terme de nombre de clusters. Nous avons dans un premier temps implémenté un programme qui simule un modèle de population en backward, en considérant un individu initial unicolore, avec un taux de coalescence de 1 (ce qui correspond à un taux de $\binom{L}{2}$ pour L lignées) et un taux de recombinaison égal à la somme des longueurs de couverture de tous les clusters (où la longueur de couverture d'un cluster est la longueur entre les deux points les plus éloignés du cluster). Ce programme a également été implémenté en C++ (compilé avec GCC v5.1), et nous a permis d'inférer numériquement la distribution de la longueur des clusters (qui est la longueur totale des blocs qui composent le cluster) situés en bordure de chromosome : une distribution exponentielle de paramètre $ln(R)/R$, où R est la "longueur" du chromosome. La raison d'être des guillemets dans "longueur" est la suivante : le modèle de recombinaison considéré n'est pas celui de Haldane, mais une version simplifiée. En effet, ici nous considérons qu'à chaque recombinaison au plus un crossover réparti uniformément sur le chromosome peut subvenir avec une probabilité ρ. En réalité, $R = \rho \times N$ où N est la taille de population. Nous ne devrions parler de la longueur de chromosome que dans le cas du modèle de Haldane ou dans des modèles similaires, où la relation entre les locus peut effectivement être formulée sous

forme de distance. Dans notre cas, nous supposons approcher le modèle de Haldane avec ce schéma encore plus simple, et nous parlerons par abus de langage de longueur (sans unité).

Dans un deuxième temps, nous avons implémenté un deuxième programme pour simuler un modèle en approche forward d'une population de type Wright-Fisher, de taille de population constante, composée d'individus haploïdes à deux parents, avec comme pressions évolutives la sélection et la dérive, et comme schéma de reproduction la panmixie (pondérée par la sélection) et le modèle de recombinaison simplifié décrit ci-dessus. Chaque individu portant l'allèle sélectionné a un taux de $1 + s$, où s est le taux de sélection, de participer aux accouplements, contrairement aux autres individus qui avaient un taux de 1. Nous avons implémenté ce programme en C++ (compilé avec GCC v5.1) en forward et non en backward à cause des difficultés supplémentaires qu'allait poser cette implémentation. Nous n'avons pas pu dériver de prédictions théoriques sur les clusters sous la pression de sélection que nous avons ici introduite, mais nous avons mené des analyses numériques pour détecter malgré tout à travers les mesures les différents types de sélection. Nous avons de ce fait implémenté différents types de sélection (localisé à un locus ou épistatique) selon des scénarios différents (hard ou soft sweep). Dans le cas d'une sélection localisée à un seul locus, nous avons réussi à détecter la sélection dans les deux types de scénarios testés.

Nous n'avons pas à proprement parlé d'extension pour cet article, ce travail n'étant pas encore publié et terminé. Il nous reste à mener les analyses dans un cas de sélection épistatique spécifique, c'est-à-dire une sélection d'une ampleur différente lorsque deux allèles dits sélectionnés sont sur le même chromosome, et d'un effet moindre s'ils sont seuls. Les sélections épistatiques

ne requièrent en rien que les deux locus soient situés sur le même chromosome, nous l'avons juste supposé dans cette partie par souci de simplicité.

Chromosome painting and ancestral recombination graph

M. Tiret, V. Miró Pina, E. Schertzer, A. Lambert, F. Hospital

Abstract

We study an idealized population where the genome of each individual at time 0 is painted uniformly in a different color. By the blending effect of recombination, the genomes of descending individuals look like mosaics of colors, where each segment of the same color is called an identical-by-descent (IBD) segment. As soon as the same mosaic genome is carried by all individuals it is not altered by recombination any further. We gather some new and existing theoretical predictions on the time to fixation, number of colors and number of segments of the fixed genome under neutrality. We show numerical simulations of this mosaic genome and use them to validate predictions and characterize the empirical distribution of colors along the genome. Next, we study alternative scenarios where one allele (or more) in the stationary mosaic background suddenly becomes advantageous and the genomes carrying it are instantly repainted. We compare the empirical distribution of colors along the genome in various such scenarios (hard or soft sweep, possible epistasis) and identify regions of parameter space for which this distribution significantly deviates from neutrality. These findings could give new insights into the detection of selection and epistasis in experimental

evolution.

1 Introduction

Population genetics is a theoretical field that focuses on how genetic material is transmitted over time. We know that what are transmitted from one generation to another are neighboring loci altogether, that is blocks of chromosome whose boudaries are either chromosome tips or crossover cutpoints. In light of this, we cannot say that loci are transmitted independently as if they were on different chromosomes, even if these loci are neutral. Former models of population genetics however assumed that loci were transmitted independently, since this assumption, yet simplifying, fitted well the scarcer maps then available. Nowadays, with the advent of Next Generation Sequencing (NGS) techniques, considering that loci are transmitted independently results in a huge loss of information, therefore developing tools to overcome this problem, to integrate multilocus information in a genome scan became essential. It has been recently shown that analyses of Extended Haplotype Homozygosity (Sabeti et al., 2002; Ohashi et al., 2004), Runs of Homozygosity (McQuillan et al., 2008; Pemberton et al., 2012; Szpiech et al., 2013; Curik et al., 2014; Kardos et al., 2017), or Long Range Haplotype test add value to classic genome scans, especially in detecting genetic signatures of selection or inferring geographic ancestry using haplotype blocks.

Multilocus tools are indeed very powerful, but integrating the dependency between loci highly increases the complexity of the genome scans. In the experiment by Teotonio et al. (personal communication) for instance, initial individuals are unrelated and non inbred, so that they are all different re-

garding to identity by descent (IBD) – could they be colored with pairwise distinct colors. After some generations, the individuals look like mosaics of colors, and actually constituted of singly-colored blocks of chromosomes. However, because of the complexity of the problem little is known about these blocks, or the sets of blocks of the same color – that hereafter we will call clusters.

Using both backward and forward simulations we have numerically derived the distribution of the length of a cluster (the sum length of all the blocks of one cluster) at the boundaries of the chromosome under neutrality, providing hereby a statistical neutrality test. On the other hand, scanning over genomes (obtained from simulations) with a sliding window and measuring the number of clusters in each window, we have shown that it is possible to detect different patterns of selection, either hard or soft sweep, possibly with epistasis.

Unlike haplotypes, knowing the ancestral population as in Teotonio ([ref]) is essential to infer clusters from real data. Since this information is rarely available, our work here is therefore mainly theoretical. We believe that in the futur more and more data like Teotonio's will be available so that our predictions could be applied. In this paper, we will first present the Wright-Fisher population model we used for forward and backward simulations, and deepen the concept of clusters. Then, using these models, we have run simulations under neutrality and under hard or soft sweep scenarios to derive numerically and empirically properties of the clusters.

1.1 A Wright-Fisher model with recombination

Consider a neutral Wright-Fisher haploid population where N individuals carry one single linear chromosome (that can be seen as a continuous segment of length 1). Let us consider the following reproduction mechanism: each individual in generation $t+1$ chooses two parents (from generation t). With probability $1-\rho$, there is no recombination and the individual inherits a perfect copy of the chromosome from one of the parents (chosen at random). With probability ρ, there is a crossing-over and the offspring inherits a chromosome that is a mixture of the two parental chromosomes (the position of the crossing-over is chosen uniformly at random).

Let us assume that the founder population (at time 0) is unrelated and not inbred, so that all individuals can be considered different regarding to identity by descent (IBD) – or could they be colored with pairwise distinct colors. As time proceeds forward, the founder chromosomes are broken apart by crossovers over generations, so that chromosomes are mosaics of singly-colored blocks. Assuming drift as the only evolutionary pressure, this population reaches fixation in finite time, i.e. after a finite number of generations, all the individuals carry the same mosaic of colors (that cannot be altered any further by recombination). The color partition of this fixed chromosome and its properties are our main interest in this paper.

The color partition looks like a mosaic of singly-colored blocks. More specifically, let us call a 'block' a maximal connected set of points (called sites) carrying the same color (i.e., a maximal singly-colored interval) and let us call a 'cluster' the subset of blocks of the same color. All the sites in a same cluster have been inherited from the same individual in the founder population, so they are identical-by-descent (IBD). The length of a cluster

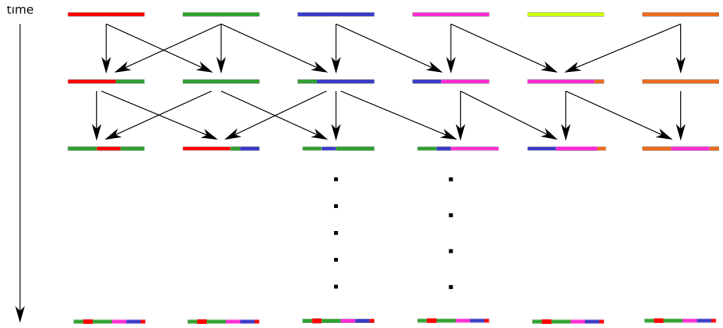

Figure 1: Wright-Fisher model with recombination

would be the total length of its blocks. Our main two measures throughout this paper will be the number and the length of the clusters, and we will try to provide the distributions of these measures.

1.2 Ancestral recombination graph and backward simulations

The model described above is difficult to analyse mathematically, and simulations are computationally intensive. This is why we introduce the Ancestral Recombination Graph (ARG), which is a generalization of the coalescent that describes the genealogy of each site of the chromosome. The ARG is an equivalent backwards in time to our model, in the large population limit and will allow us to perform more efficient simulations.

First of all, we are going to take a large population limit. To be precise, we consider that the probability of recombination ρ scales with N, in such a way that:

$$N\rho_N \underset{N\to\infty}{\longrightarrow} R$$

where R is a constant. Assume time is rescaled by N and let $N \to \infty$. We

are going to define a model that would be equivalent to our Wright-Fisher model with recombination in an infinite population, backwards in time. The only parameter of this model is R, that would correspond to the size of the chromosome in units of recombination in the new timescale.

The idea behind the ARG is to follow backwards in time the ancestral lineage of each site in the chromosome. Let us start by considering two sites that are at distance d, with $0 < d < R$, on the same chromosome sampled at present time. We would like to know whether t time backwards, the ancestors of these two sites were together in the same individual (state 1) or in two different individuals (state 2). Actually as time proceeds backwards, these two lineages jump from the state 1 to the state 2 (recombination seen backwards in time) at rate d and from the state 2 to the state 1 at rate 1 (coalescence). It is then simple to compute the probability that the two lineages were in the identical state t time units backwards. In particular as t goes towards infinity, this probability converges to $1/(\rho d + 1)$. This corresponds to the probability that these two sites are inherited from the same individual in the founder generation, i.e., the probability that they are in the same color cluster.

Now, consider a whole chromosome sampled in the present population. As time proceeds backwards, and before this chromosome experiences its first recombination event, all the sites in the chromosome are carried by the same ancestral lineage. When the first recombination event takes places, this lineage splits into two: one lineage corresponding to all sites on the right-hand side of the cut-point and another corresponding to the sites on the left-hand side. These two lineages can further coalesce and also each of them can split into two new lineages. Each splitting (recombination) event

is associated to the position of the cut-point on the chromosome; the sites that were carried by the same lineage are now carried by one of the two new lineages depending on if they are on the left or the right of the cut-point. More precisely, each pair of lineages coalesces at rate 1 (so, if there are L lineages, the total coalescence rate is $\binom{L}{2}$). Each lineage splits into two at rate $cov(\ell)$, where $cov(\ell)$ is the coverage length of the cluster (set of sites) associated to this path (ℓ). The coverage length of a cluster is the length between the leftmost and the rightmost points of the cluster. The position of the associated cut-point is chosen uniformly at random between these two points. Starting from one chromosome, one can see the ARG as a process with values in the partitions of the chromosome: each lineage at time t corresponds to a cluster, that is a set of sites of the same color. The stationary distribution of the ARG is the same as the distribution of the partition that is fixed in the Wright-Fisher model with recombination (when $N \to \infty$) (see Figure 1.2).

We performed simulations of the ARG, for different values of R (10,000 replicates each) using the programming language C++ (GNU 5.1), and analysed the results with R (R version 3.2.3). We considered that this process reached its equilibrium state when the number of clusters reached a stationary state empirically defined (data not shown). We do not have any prediction for this equilibrium. Therefore, we simulated the process until the equilibrium was reached and characterized the color partition at equilibrium.

1.3 Selection model

Detecting selection through its multilocus consequences is the core principle behind EHH or ROH. We have looked if clusters were affected equivalently by

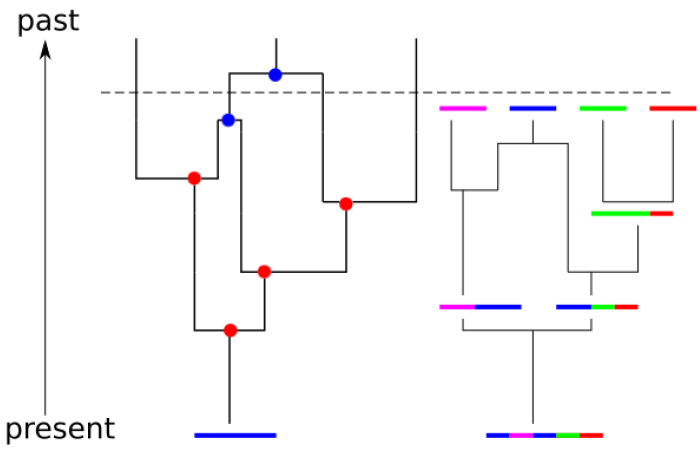

Figure 2: The left side corresponds to a realization of the ARG. Blue dots correspond to coalescence events and red dots correspond to recombination events (and are associated to a position in the chromosome). In the right side, we see how the ARG at time t (represented by the dashed line) corresponds to the color partition of a sampled chromosome in the present population.

different patterns of selection. Although backward simulations are computationally faster, including selection in ARG simulations is really complicated, needing further assumptions. Therefore, we had to perform forward simulations, with finite population sizes. Because the forward simulations are computationally slower and heavier, only small values of R could be tested.

We have included selection in the Wright-Fisher model aforementioned in the following way: we first wait for fixation under neutrality (i.e. only drift), then start a selection scenario. At fixation, one individual (hard sweep) or several individuals (soft sweep) are randomly drawn and entirely recolored uniformly: these individuals have a same new color. From this generation on, these individuals are considered selected and have a higher fitness. In the simple model, we considered that an individual has a fitness of $1 + s$ if it carries this new color at the middle of the chromosome, s being the

selection coefficient, and a fitness of 1 otherwise. In the model with epistasis, we considered that an individual has a fitness of $1 + s$ if it carries this new color both at the third and two third of the chromosome length, and a fitness of 1 otherwise. Again, we wait until fixation, before proceeding to a genome scan.

In this model we have tried to detect selection through the measure of the number of clusters per sliding window, that is a measure on a part (which length is fixed) of the chromosome, then shifting this window by a (also fixed) decay. This decay is smaller than the windows size, so that the sliding windows are overlapping. Since there is no state of art justifying the value of the length and the decay values of sliding windows, we will consider different values of these. Therefore, when we say that we could detect selection we mean that there is at least a window size and a decay value for which the genetic profile has a particular behavior around the selected (and known) locus; reciprocally, when we say that we could not detect selection, we mean that we have not found any value of window size and decay that makes it possible to identify any particular behavior. Posing that the chromosome length is 1, the sizes and the decays of sliding windows are expressed relatively to chromosome length as a percentage.

2 Results

2.1 The neutral case

In this section, we gather together some new and previous theoretical results on the number and the length of the clusters that allow us to describe the color partition at equilibrium, for an infinite population, in the long chromo-

some limit (when $R \to \infty$).

First of all, let $\mathcal{L}^R(0)$ the total length of the cluster covering the origin (the leftmost cluster in the chromosome). Some of the co-authors have shown (personal communications) that, in the long chromosome limit, $\mathcal{L}^R(0)$ follows an exponential distribution of parameter $\log(R)$. By symmetry, the length of the cluster covering R (the rightmost cluster) follows the same distribution. Similarly, let $x \in]0, R[$ a position on the chromosome and consider $\mathcal{L}^R(x)$, the length of the cluster covering x (the set of sites that are of the same color as x). It has been shown that when R is large enough, the distribution of $\mathcal{L}^R(x)/\log(R)$ can be approximated by a Gamma distribution, $\Gamma(2, 1)$.

From these results, the mean length of a cluster (covering a given position), in a chromosome of length R is $\log(R)$, so N_R, the number of clusters in a chromosome of length R, should be of the order $R/\log(R)$. In fact, Wiuf and Hein (1997) showed, using simulations that, when R is large, $N_R \simeq C\frac{R}{\log(R)}$, where $C = 1.28$.

We performed numerical simulations in order to study the empirical distribution of the cluster lengths in a given chromosome (i.e. the distribution of lengths of the clusters on a chromosome at equilibrium). We used the backwards algorithm described in Section 1.2. We tested values of R in a range from 2000 to 50000 and the number of simulations performed was 500. As in Wiuf and Hein (1997), we found that $N_R \simeq 1.28\frac{R}{\log(R)}$. For each replicate, we have found that the empirical distribution of the cluster lengths (renormalised by $\log(R)$) fitted, as expected, an exponential distribution (Kolmogorov-Smirnov test with p-values always lower than 0.004). Therefore, we fitted each replicate data with exponential distributions (using Scipy's function 'fit'), to find a maximum likelihood estimator of λ, the

parameter of the exponential distribution. We found a mean λ of 0.77, with a standard deviation of 0.02. This value is very close to 1/1.28. See Figure 3 for a summary of the results. To conclude, we have showed that, the distribution of the cluster lengths in a given chromosome follows an exponential distribution of parameter $0.77 \log(R)$ (see Figure 4).

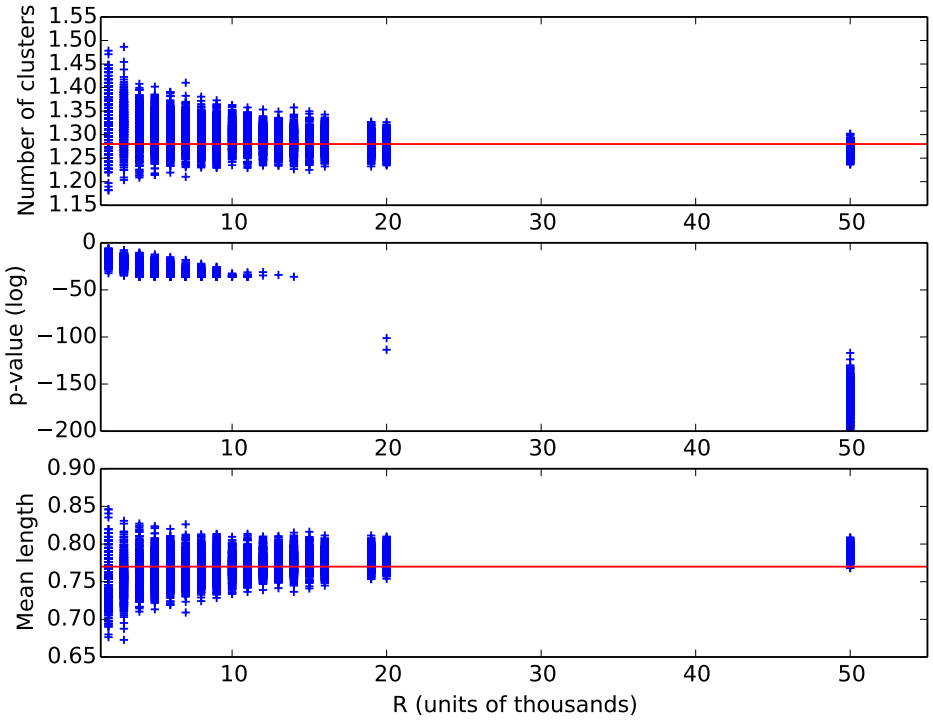

Figure 3: The fist figure represents the number of clusters, renormalised by $R/\log(R)$ for each replicate. The red line corresponds to 1.28. The second figure represents the p-values of the Kolmogorov-Smirnov test (comparison with an exponential distribution) for each replicate. The third figure represents the maximum likelihood estimator of the parameter of the exponential distribution for each replicate. The red line represents the mean, 0.77.

Our result shows that, in a given chromosome, the mean length of a

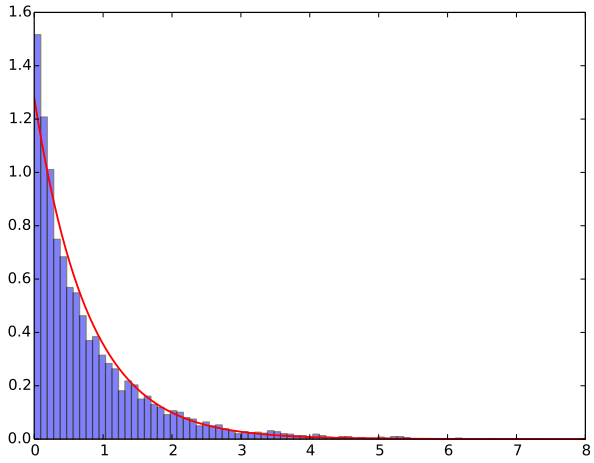

Figure 4: Empirical distribution of the cluster lengths. The blue histogram represents the distribution of the cluster lengths on a chromosome, renormalised by $\log(R)$. Parameters of the simulation: $R = 50000$. The red curve represents the probability density function (corrected with the histogram bar max length), for an exponential random variable of parameter $1/1.28$. The p-value of the Kolmogorov-Smirnov test is 2.10^{-16}.

cluster is $0.77 \log(R)$. In fact, the length of the cluster covering a given site ($\mathcal{L}^R(x)$ or $\mathcal{L}^R(0)$) is, on average, larger than the length of a randomly chosen cluster. This is because of size biasing: larger clusters have a better chance of covering a given position, so $\mathbb{E}(\mathcal{L}^R(x))$ and $\mathbb{E}(\mathcal{L}^R(0))$ are larger than the mean length of a randomly drawn cluster.

2.2 Selection and different scenarios

We have studied the consequences of several patterns of selection on clusters, and especially the number of clusters. The selection could either be modelized as monolocus or epistatic (as described above). We have only worked here with a forward approach and then with a small population size (1,000 haploid

individuals) because of the computational limits.

On Figure 5, we could see that the number of clusters is effectively sensitive to selection, and that it decays drastically around the selected locus in the case of a hard sweep monolocus selection. However, we are here interested to see whether this measure, the number of clusters, could tell more about selection that other methods could; and in this case, seeing Figure 6, any other method could have detected a large chromosome block in the middle. We have then tried to explore range of parameters that avoid such a simple detection.

We have to this end simulated smaller selection rate ($s = 0.05$), and we could see on Figure 7 that we could still detect selection in the middle of the chromosome by a slight decrease of the number of clusters. We could also see on Figure 8 that this is not due to a larger block in the middle of the chromosome.

Hard sweep is however usually easy to detect, unlike soft sweep selection pattern. Here we have considered that during a soft sweep only 10% of the individuals undergo selection. We then have simulated such a pattern, in a monolocus case. We could see on Figures 9 and 10 that in the case of soft sweep, as in the case of hard sweep, detecting selection is easy when the selection rate is high ($s = 0.2$), but has the same "problem" than the equivalent hard sweep case. We then have simulated a smaller selection rate of 0.05, and we could see on Figures 11 and 12 that the number of clusters makes it possible to detect small selection in a soft sweep scenario.

The same analysis on epistatic pattern of selection (as described above) was not that successful (see Figure 13). In other words, we have not found yet the parameter ranges in which the number of clusters detect selection

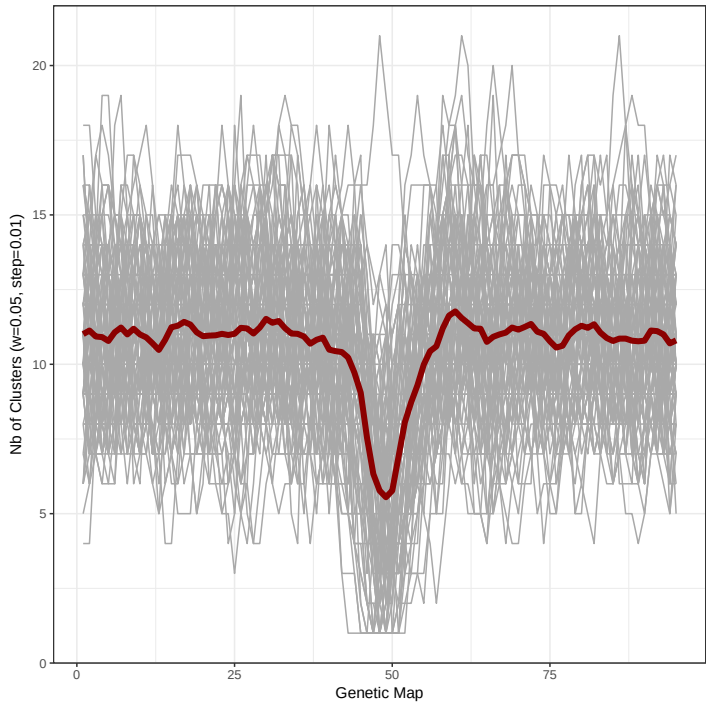

Figure 5: Number of clusters per overlapping sliding window of size 0.05 and a decay (or step) of 0.01 in a 1,000 sized population undergoing hard sweep and monolocus selection. The gray lines are the number of clusters for each simulation (1,000 simulations), and the bold red line is the mean value of these simulations. The selection rate is high ($s = 0.2$) so that we are sure to observe an effect, if the number of clusters is sensitive to selection.

without an easily detectable larger block of chromosome.

3 Conclusion and Perspectives

In this article we have tried to develop the concept of clusters in chromosome painting, and also developed some associated measures, such as the number of clusters over a sliding window. We have theoretically and numerically derived the distribution of the cluster length at the boudaries of the chromosome in the neutral case, that could be used for implementing any neutral multilocus

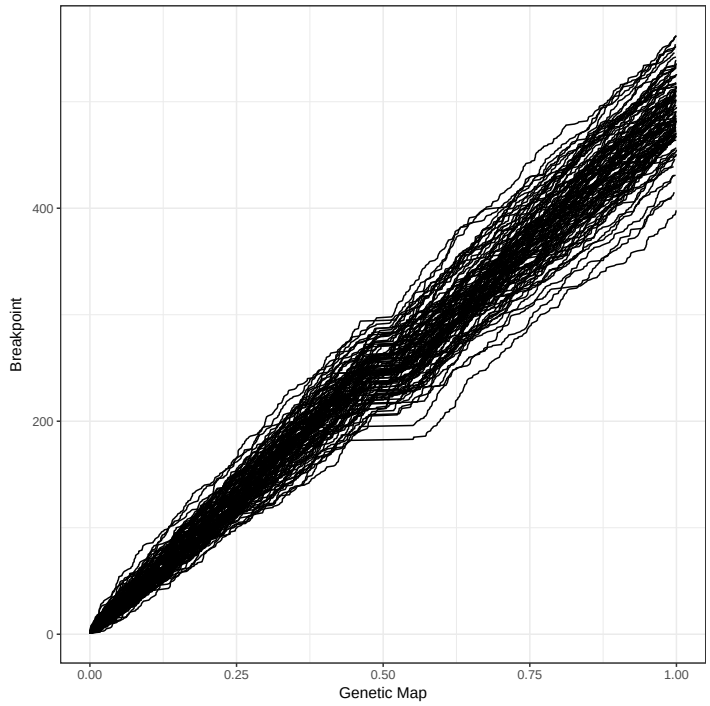

Figure 6: Plotting for all the simulations of a 1,000 sized population undergoing hard sweep and monolocus selection ($s = 0.2$) the cumulative number of blocks over the chromosome. We could see that there is a platform in the middle of the chromosome, indicating a large unique block and consequently, this is a selection pattern easy to detect with any other selection detecting method.

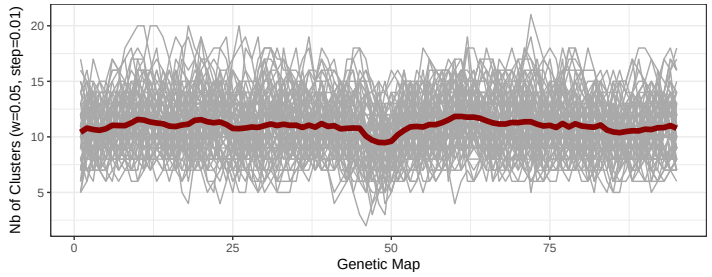

Figure 7: Number of clusters per overlapping sliding window of size 0.05 and a decay (or step) of 0.01 in a 1,000 sized population undergoing hard sweep and monolocus selection. The gray lines are the number of clusters for each simulation (1,000 simulations), and the bold red line is the mean value of these simulations. The selection rate is small ($s = 0.05$).

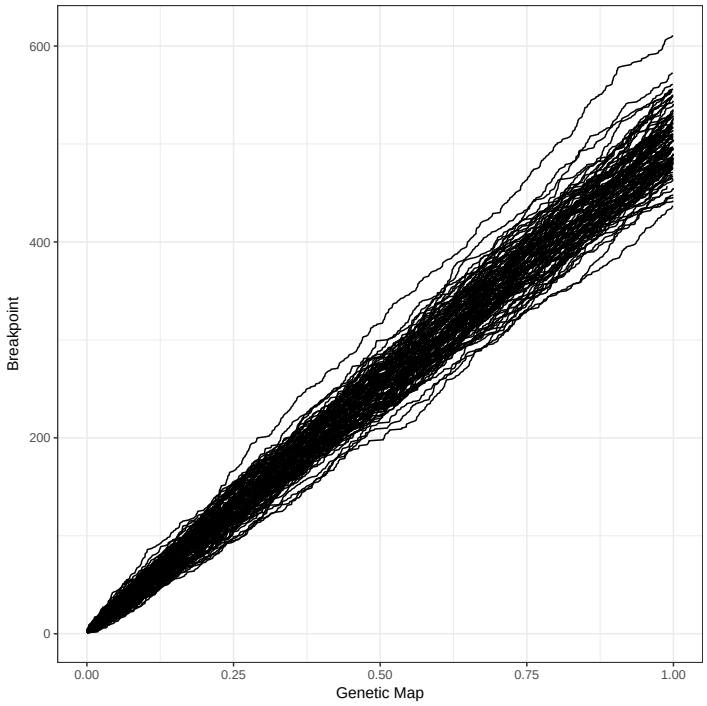

Figure 8: Plotting for all the simulations of a 1,000 sized population undergoing hard sweep and monolocus selection ($s = 0.05$) the cumulative number of blocks over the chromosome. The cumulative number of blocks follows a line without any platform, indicating that there is no specially larger block anywhere.

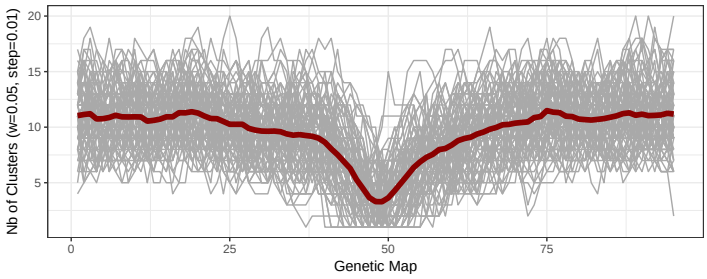

Figure 9: Number of clusters per overlapping sliding window of size 0.05 and a decay (or step) of 0.01 in a 1,000 sized population undergoing soft sweep (of 10%) and monolocus selection. The gray lines are the number of clusters for each simulation (1,000 simulations), and the bold red line is the mean value of these simulations. The selection rate is high ($s = 0.2$).

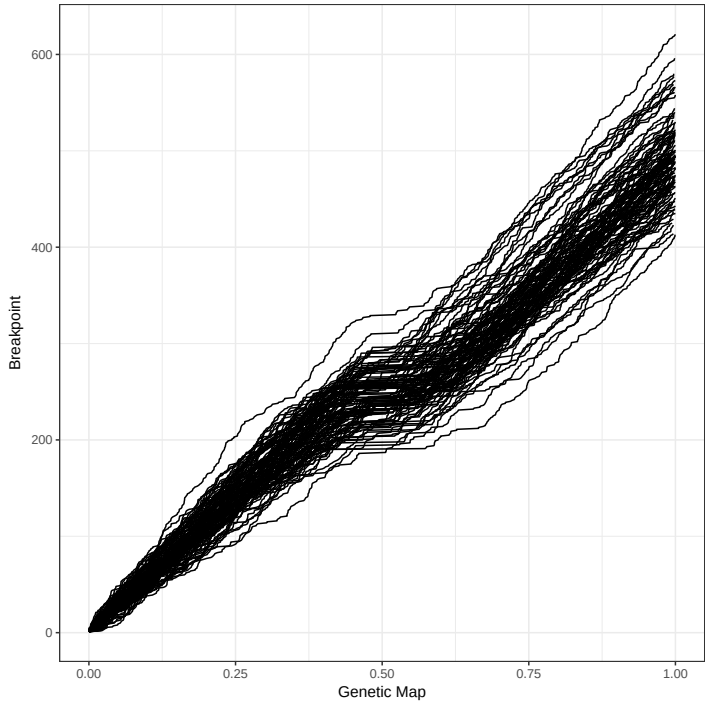

Figure 10: Plotting for all the simulations of a 1,000 sized population undergoing soft sweep (of 10%) and monolocus selection ($s = 0.2$) the cumulative number of blocks over the chromosome. We could see that there is a platform in the middle of the chromosome, indicating a large unique block and consequently, this is a selection pattern easy to detect with any other selection detecting method.

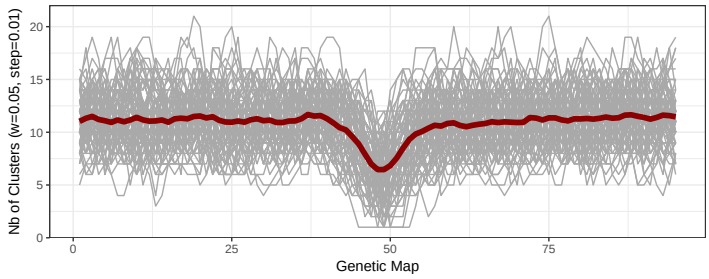

Figure 11: Number of clusters per overlapping sliding window of size 0.05 and a decay (or step) of 0.01 in a 1,000 sized population undergoing soft sweep (of 10%) and monolocus selection. The gray lines are the number of clusters for each simulation (1,000 simulations), and the bold red line is the mean value of these simulations. The selection rate is small ($s = 0.05$).

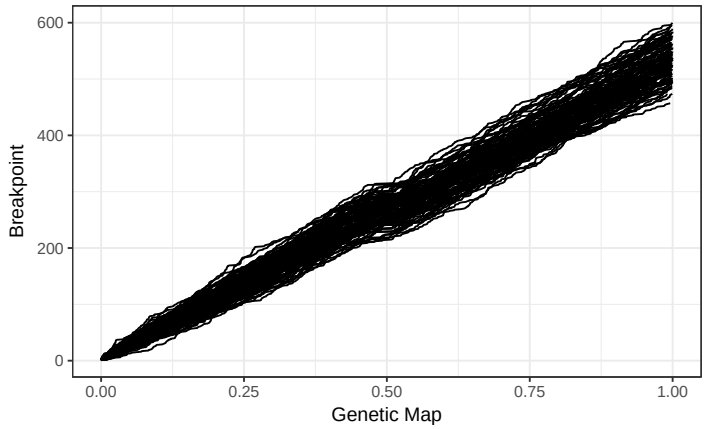

Figure 12: Plotting for all the simulations of a 1,000 sized population undergoing soft sweep (of 10%) and monolocus selection ($s = 0.05$) the cumulative number of blocks over the chromosome. the cumulative number of blocks follows a line with a tiny platform, indicating that there is a larger block.

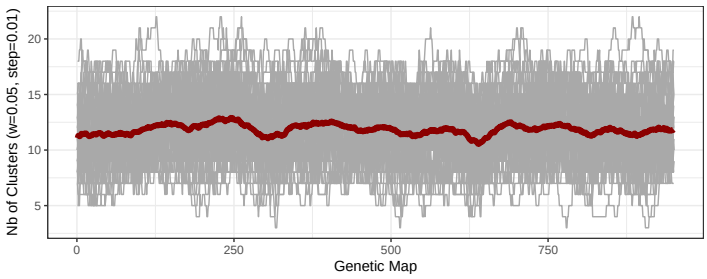

Figure 13: Number of clusters per overlapping sliding window of size 0.05 and a decay (or step) of 0.01 in a 1,000 sized population undergoing soft sweep (of 10%) and epistatic selection. The gray lines are the number of clusters for each simulation (1,000 simulations), and the bold red line is the mean value of these simulations. The selection rate is small ($s = 0.05$).

test. We also have numerically and graphically deduced that studying the number of clusters is a powerful tool to detect selection in different patterns, even for usually considered difficult pattern to detect, such as soft sweep scenario.

Epistatic selection could not be detected successfully as clearly as the monolocus selection pattern. Epistasis as we modeled here is a very particular case of epistasis, and could be modeled in thousand of other ways: indeed, considering several loci increases the number of parameters. In further studies, working on different epistatic scenarios will eventually make it possible to use the measure we have developed here, the number of clusters, and then easily integrate the multilocus dimension in genome scan analysis on newly available higly dense data, on which multilocus interaction could not be ignored.

Chapitre 4

Les autocorrélogrammes

Dans ce projet court (qui n'est pas sous le format d'un article), nous étudions l'objet particulier qu'est l'autocorrélogramme, une autre façon de concevoir les mesures multilocus. Imaginons que l'on mesure pour chaque locus une statistique – un profil génétique initial. Pour une distance donnée en Morgan, il existe un certain nombre de paires de locus éloignés d'une telle distance, et donc un certain nombre de paires de statistiques. Soit E_d l'ensemble de ces paires pour une distance d. L'autocorrélation de ce profil est une fonction qui à une distance en Morgan associe une valeur comme suit :

$$A(d) = corr(E_d(1), E_d(2)) \tag{4.1}$$

où $A(d)$ est l'autocorrélation d'une statistique pour la distance d en Morgan, *corr* est la corrélation statistique, et $E_d(1)$ et $E_d(2)$ les ensembles respectivement des premiers et deuxièmes éléments de chaque paire de E_d. Ainsi, l'autocorrélogramme est une fonction qui à une distance donnée associe la corrélation entre les statistiques mesurées à des locus éloignés de ladite dis-

tance. À noter que plus d augmente, si le chromosome est modélisé comme un objet fini (ce qui est ici le cas), moins E_d comprend d'éléments et plus l'autocorrélation est sensible au bruit.

L'autocorrélogramme est une mesure différente de celles développées précédemment, dans le sens où celui-ci produit un profil génétique : cette mesure n'est pas juste une valeur, mais une fonction, un ensemble de valeurs. Ce n'est donc plus une valeur mais le tracé du graphe en entier qui est à interpréter.

Nous considèrerons un modèle de population de type Wright-Fisher, avec une taille de population N constante, composée d'individus diploïdes, durant un certain nombre de générations non chevauchantes, et avec comme pressions évolutives la dérive et la sélection (selon un taux s, sur le locus du milieu de chromosome, avec une fitness avantageuse de $1 + s$ et de 1 sinon). Les chromosomes seront modélisés comme des vecteurs de $n = 1000$ Single Nucleotide Polymorphisms (ou SNPs), chacun distant de 0,1 cM, le chromosome mesurant ainsi 1 Morgan. Chaque chromosome de la population initiale est unique (chaque allèle à chaque locus est à une fréquence de $1/2N$). Nous mesurerons l'autocorrélogramme, à certaines générations noté g, de l'hétérozygotie, c'est-à-dire que le profil génétique initial est un profil d'hétérozygotie. Nous définissons l'hétérozygotie sur un locus comme suit :

$$He = 1 - \sum_{i=1}^{K} p_i^2 \tag{4.2}$$

où He est l'hétérozygotie, K le nombre d'allèles au locus considéré et p_i la fréquence de l'allèle i. Ce modèle a été simulé avec un programme implémenté en C++ (compilé avec GCC v5.1).

Nous avons ici simulé des populations jusqu'à fixation ou perte de l'allèle

sélectionné, pour différente valeur de s. Sur la Figure 4.1, nous pouvons voir le scan de l'hétérozygotie et l'autocorrélation associée. Idem sur la Figure 4.2 et la Figure 4.3. Le scan est nettement plus en V lorsque l'allèle s'est effectivement fixé, mais il est difficile de voir un schéma se dégager sur l'autocorrélogramme. Le bruit étant fort, il serait donc intéressant de considérer un nombre important de réplicats.

Nous avons donc simulé un grand nombre de réplicats R (ici, $R = 10,000$) pour obtenir l'autocorrélogramme moyen en fonction de s. Puisque nous voulons voir l'effet de la sélection avec le moins de bruit possible, nous n'avons gardé que les populations dans lesquelles la sélection a été effective, c'est-à-dire les populations qui ont effectivement fixé leur allèle sélectionné, pour $g = 40$ et $N = 50$ (voir la Figure 4.4).

Nous voyons sur la Figure 4.4 que pour différentes valeurs de s nous obtenons différents autocorrélogrammes. Nous aimerions maintenant une statistique résumé par autocorrélogramme pour pouvoir déterminer s'il y a eu sélection ou non, et d'étudier la distribution pour en faire un test. Par ailleurs, nous voudrions maintenant ne plus avoir à filtrer les populations en fonction de leur état final (si elles ont fixé ou non leur allèle) pour étudier l'influence des bruits.

Nous avons donc simulé des populations avec $N = 50$, $g = 60$ et $R = 300$, sans filtrer l'état final. Nous avons tracé sur la Figure 4.5 l'autocorrélogramme moyen pour différentes valeurs de s. Sur la Figure 4.6 nous avons tracé l'autocorrélogramme moyen, le premier quartile et le troisième pour $s = 0.2$. Nous observons une différence forte entre les autocorrélations pour les distances inférieures à 0,5 Morgan et celles supérieures à 0,5 Morgan, différence que l'on observe pas sans sélection : ceci est juste dû à la position du

locus sélectionné au milieu du chromosome. Le bruit d'une envergure bien plus grande à droite de l'autocorrélogramme est aussi tout à fait normal, ceci est dû au fait que le nombre de paires de statistiques diminue au fur et à mesure que les distances entre paires de statistiques augmentent sur un chromosome de longueur finie, comme évoqué précédemment.

Il semble que l'on puisse prendre comme statistique pour résumer un autocorrélogramme l'abscisse auquel l'autocorrélogramme s'annule pour la première fois (il serait aussi possible de prendre une autre ordonnée que 0). Nous avons tracé la distribution de x_0, le premier abscisse auquel l'autocorrélogramme s'annule (voir la Figure 4.7 et Figure 4.8). Les distributions sont nettement différentes en fonction de la valeur de s, ce qui permet même d'envisager l'estimation de la valeur de s à partir de l'autocorrélogramme (ici avec x_0).

Pour conclure brièvement, nous avons essayé d'illustrer ici une approche méthodologique purement numérique dans un cadre multilocus, avec un objet particulier qu'est l'autocorrélogramme. Nous avons réussi à établir des différences de distributions fortes entre les différentes valeurs de taux de sélection. Nous n'avons ici pas cherché à dériver des prédictions d'autocorrélogramme, mais l'intérêt fondamental de cette mesure est qu'il serait facile de le faire étant donné l'immense étendue de la théorie physique sur le traitement des signaux bruités, et le commun qu'est l'autocorrélogramme.

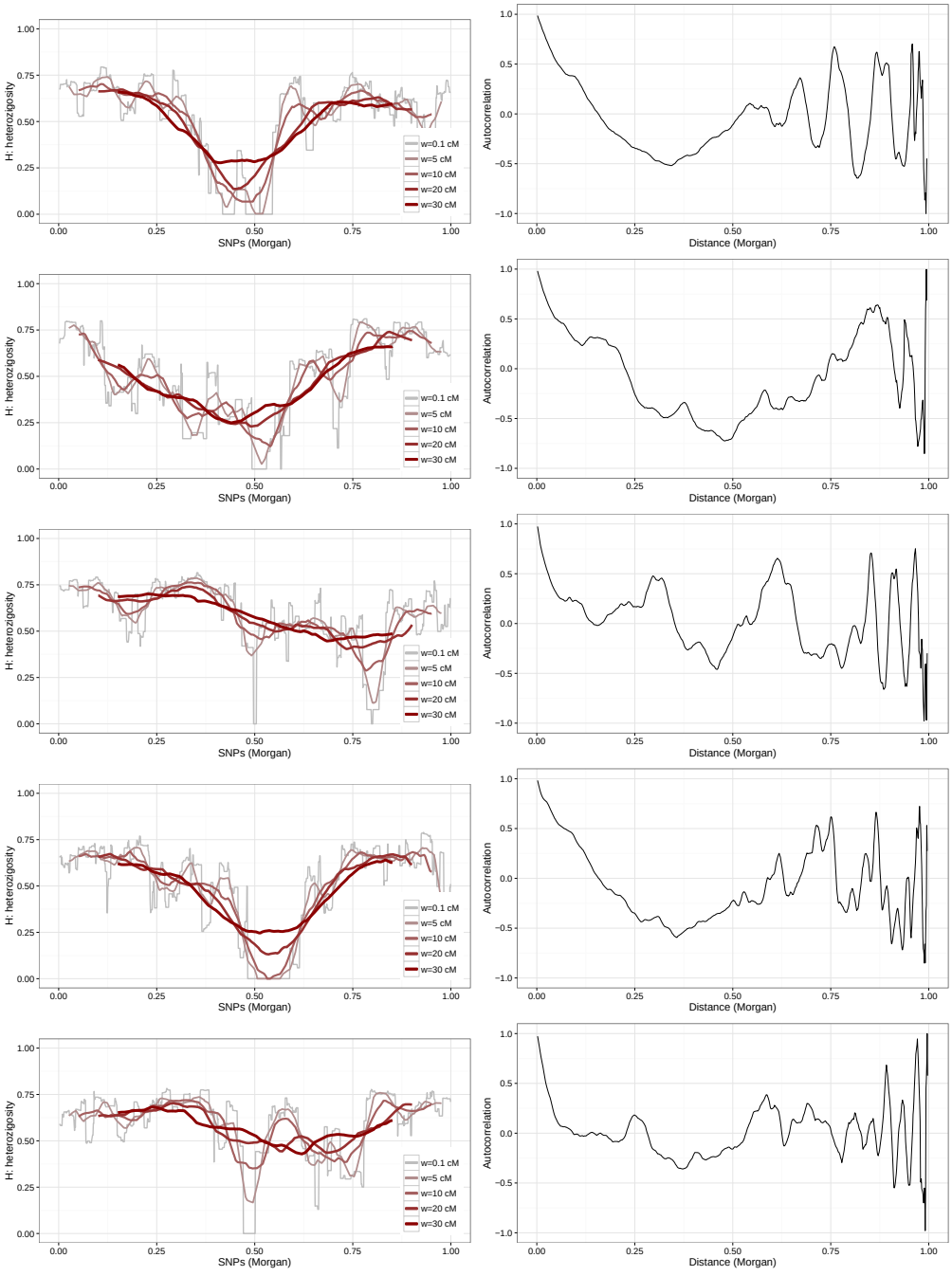

FIGURE 4.1 – Cinq réplicats de population : $g = 10$, $N = 50$, $s = 0.2$. Ces populations ont fixé leur allèle sélectionné. À gauche le scan pour l'hétérozygotie, et à droite, l'autocorrélogramme correspondant.

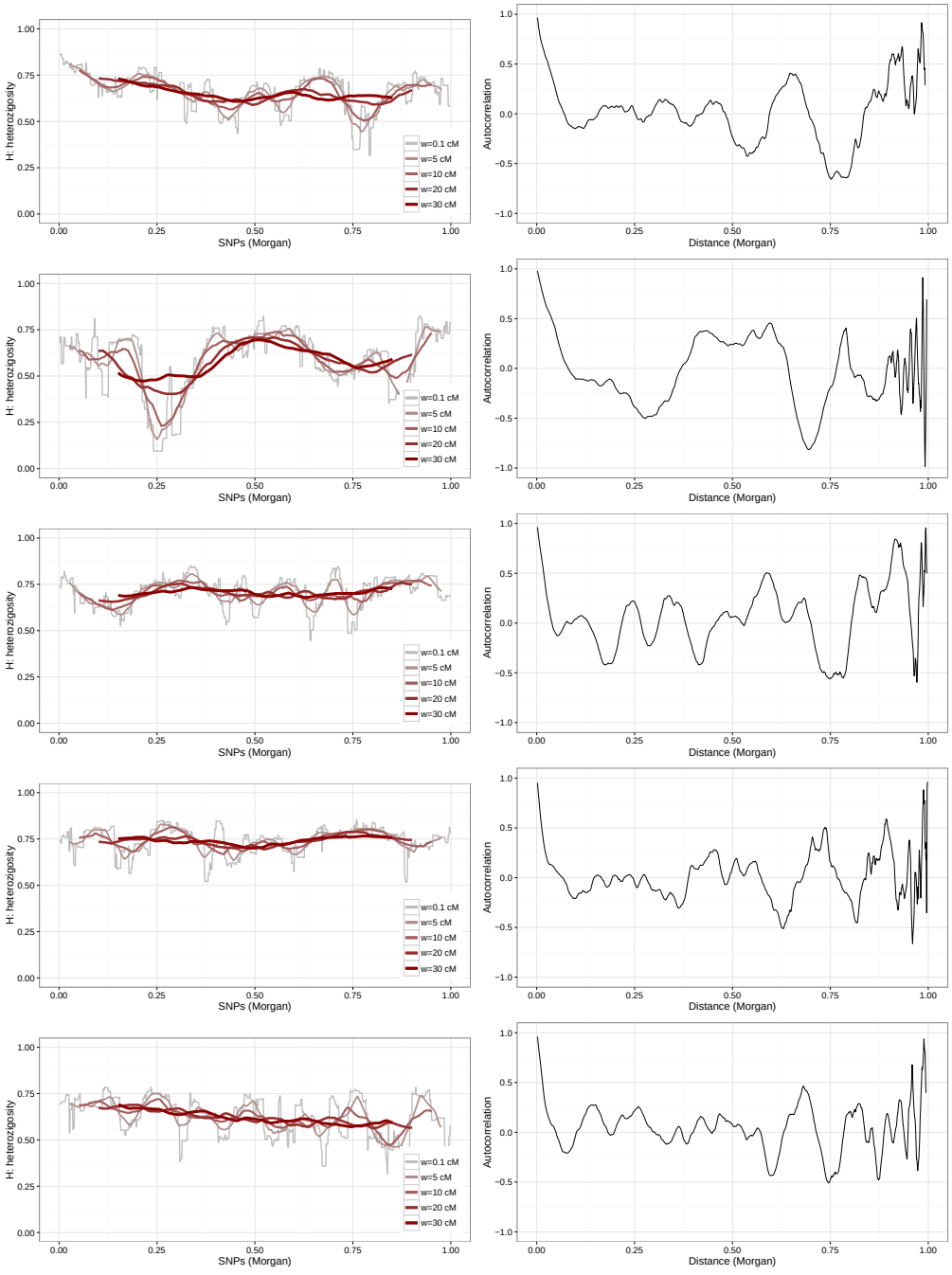

FIGURE 4.2 – Cinq réplicats de population : $g = 10$, $N = 50$, $s = 0.2$. Ces populations ont perdu leur allèle sélectionné. À gauche le scan pour l'hétérozygotie, et à droite, l'autocorrélogramme correspondant.

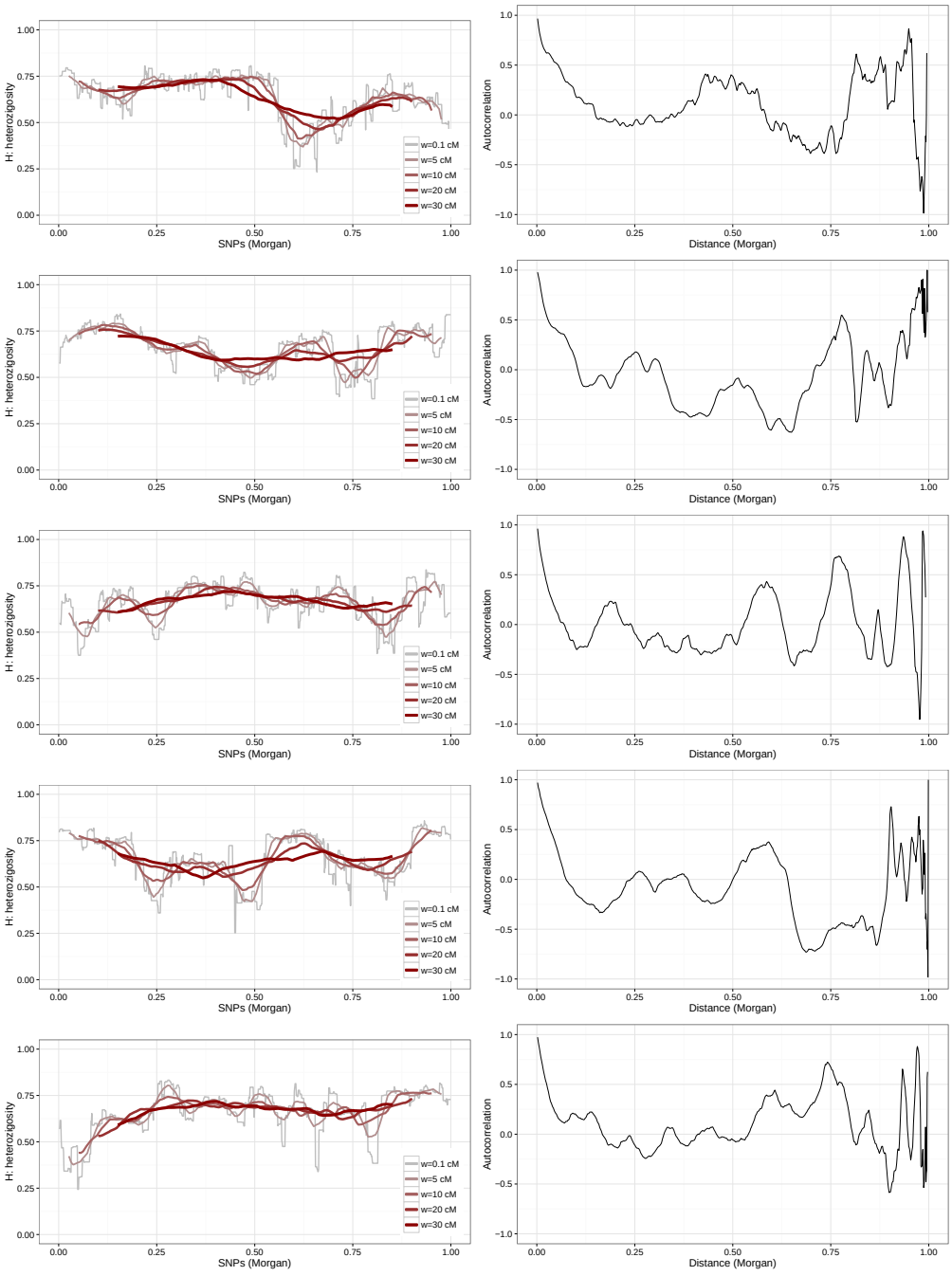

FIGURE 4.3 – Cinq réplicats de population : $g = 10$, $N = 50$, $s = 0$. À gauche le scan pour l'hétérozygotie, et à droite, l'autocorrélogramme correspondant.

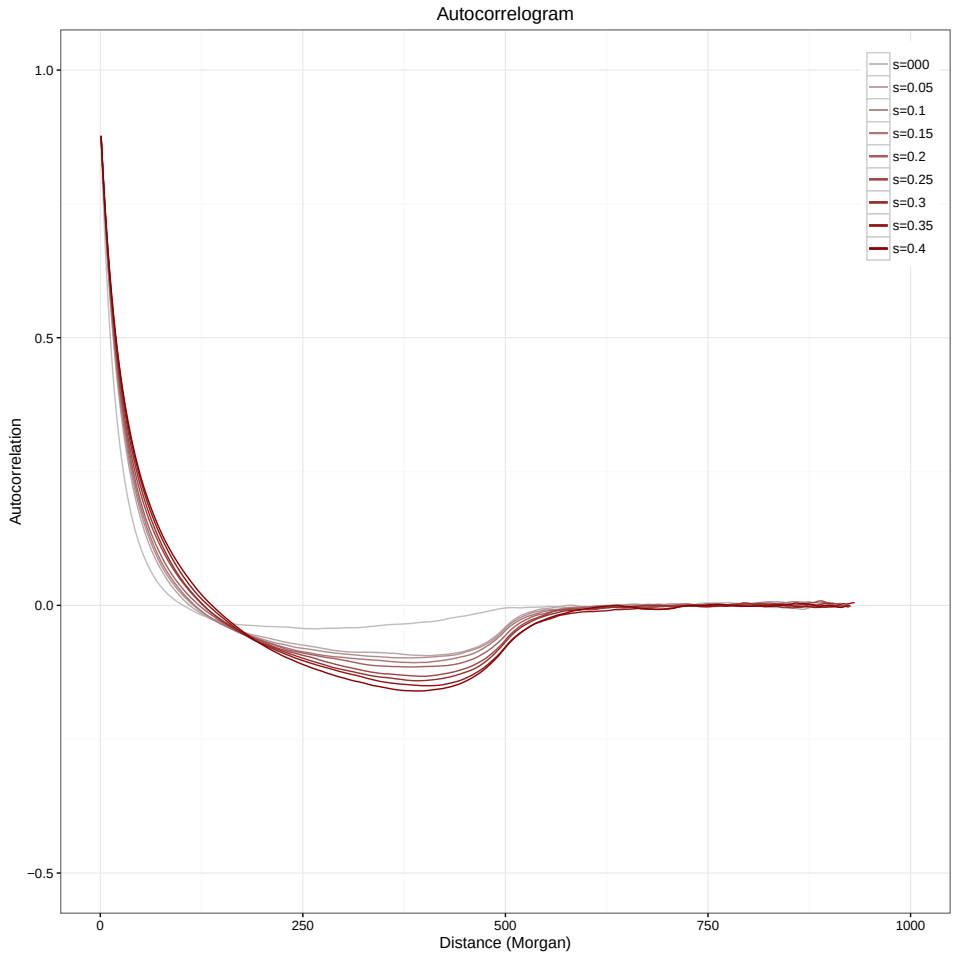

FIGURE 4.4 – Autocorrélogramme moyen sur 10,000 réplicats, pour différentes valeurs de s, $g = 40$, $N = 50$, dans des populations qui ont fixé leur allèle sélectionné. Plus la courbe est foncée, plus s est grand.

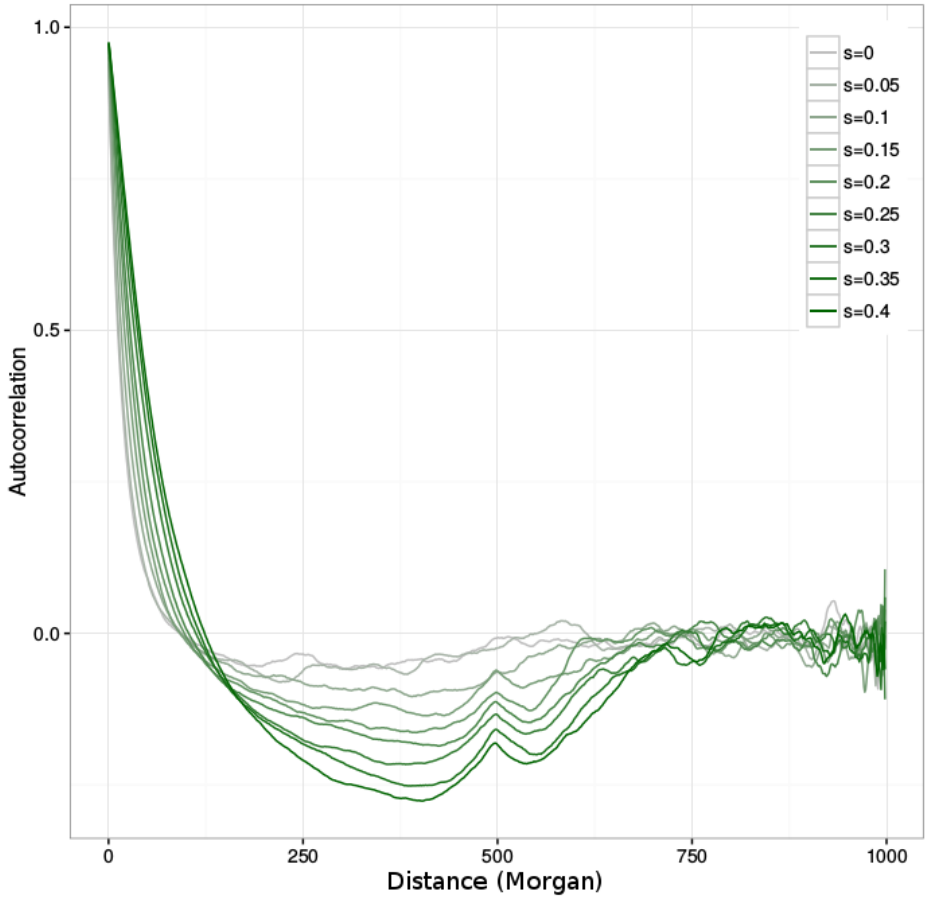

FIGURE 4.5 – Autocorrélogramme moyen avec toutes les populations (fixées ou perdues). Avec comme paramètres : $N = 50$, $g = 60$, $R = 300$, pour différentes valeurs de s (plus la courbe est foncées, plus s est grand).

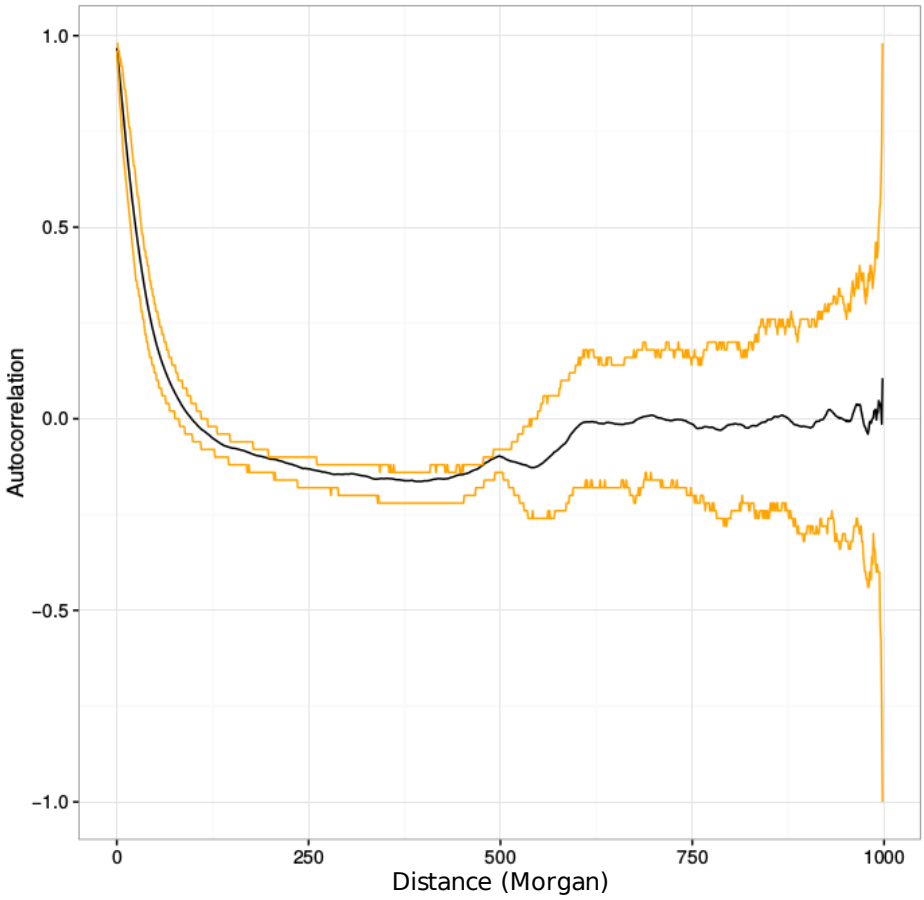

FIGURE 4.6 – Autocorrélogramme moyen avec toutes les populations (fixées ou perdues). Avec comme paramètres : $N = 50$, $g = 60$, $R = 300$, $s = 0.2$. La courbe noire est l'autocorrélation moyenne, et les courbes oranges sont les premier et troisième quartiles.

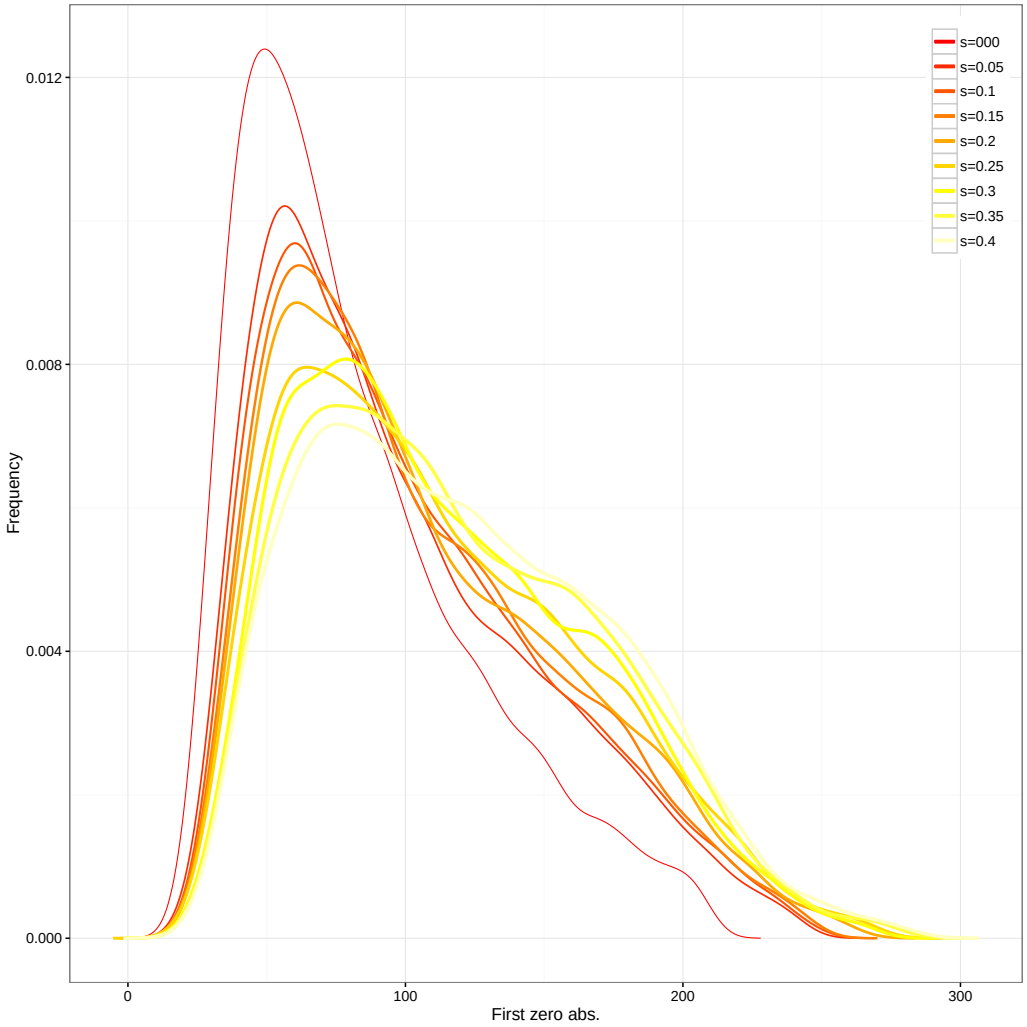

FIGURE 4.7 – Distribution de x_0 selon différentes valeurs de s, avec comme paramètres : $N = 50$, $g = 60$, $R = 10000$.

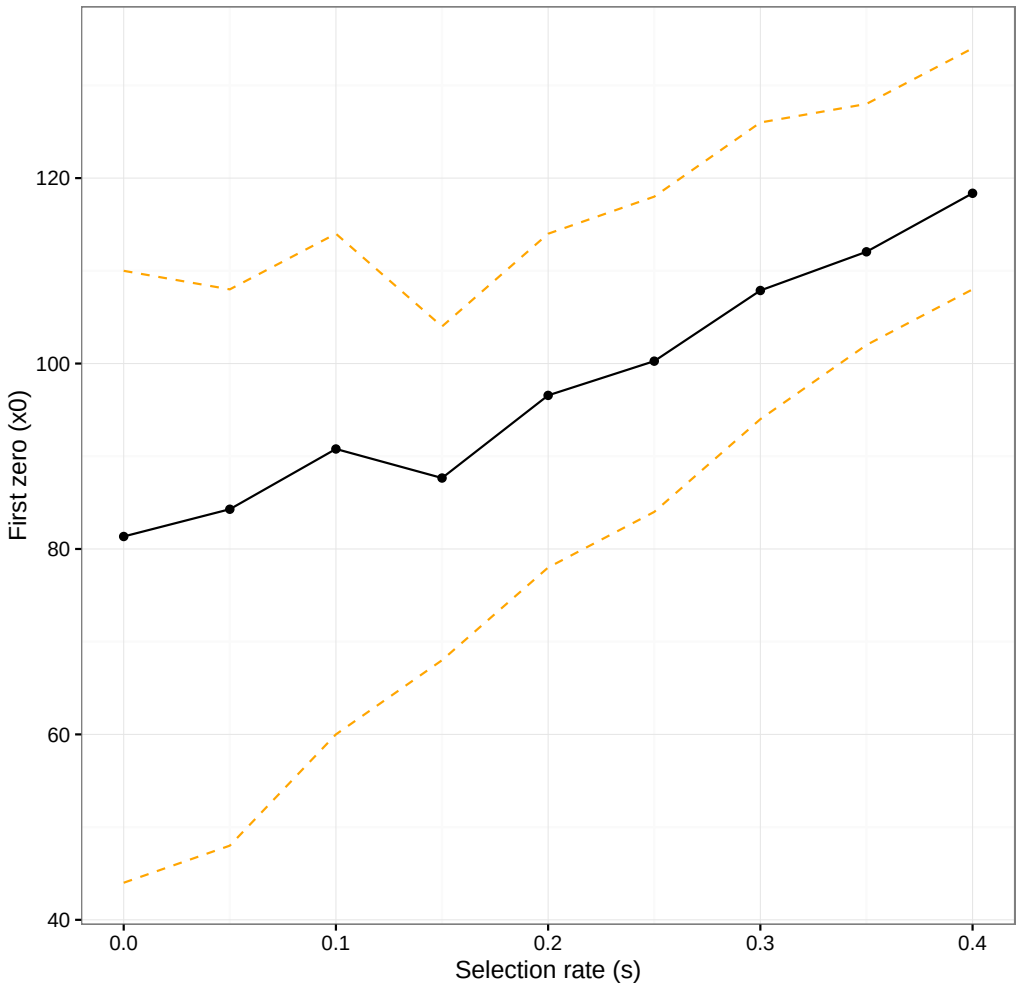

FIGURE 4.8 – Valeur moyenne de x_0 en fonction de s, avec comme paramètres : $N = 50$, $g = 60$, $R = 300$. La courbe noire est la valeur moyenne et les courbes oranges sont les premier et troisième quartiles.

Chapitre 5

Conclusions et perspectives

Dans cette thèse nous avons essayé d'illustrer ce qu'est la génétique des populations multilocus, en précisant les fondements de cette discipline et en présentant plusieurs travaux de modélisation et de méthodologie. Nous avons essayé de concentrer nos efforts sur l'aspect multilocus des transmissions génétiques, c'est-à-dire la transmission en "bloc", et ce à travers l'IBD et les clusters. Les autocorrélogrammes ont permis d'illustrer un projet de méthodologie plus avancé et des retombées qu'une telle étude pourrait avoir. Le travail de cette thèse n'est d'ailleurs pas terminé, et soulève encore plusieurs questions : comment développer une mesure multilocus, et quelle en est la méthodologie ? Nous n'avons ici fait qu'illustrer des méthodologies, sans vraiment la connaitre et sans vraiment pouvoir la formuler. Les projets ont chacun visé à montrer par l'exemple ce qu'est un développement de mesure multilocus. En considérant que collectionner les exemples de développement de mesure multilocus permette au bout du compte de connaitre la méthodologie des mesures multilocus, nous définissons en quelque sorte en extension ladite méthodologie, plutôt qu'en compréhension (voir le vocabulaire mathé-

matique utilisé dans le premier chapitre). Si tel est notre effort, cette thèse n'en est que le début.

Le deuxième chapitre a été l'occasion de développer une mesure multilocus basée sur l'identité par descendance, qui est une étude de cas très classique. Ce chapitre a été l'occasion de tirer pleinement profit du travail de modélisation effectué dans le premier chapitre, et de l'exploiter mathématiquement. Nous avons développé une prédiction mathématique de la taille moyenne des blocs IBD, mais en dépit de sa généralité, cette prédiction concerne uniquement le premier moment de la distribution qui nous intéresse. Une telle mesure quantifie en quelque sorte la transmission multilocus au cours du temps, et ouvre la voie, une fois l'IBD et la transmission multilocus clairement mises en lien, de son utilisation dans la pratique. Par exemple, dans une analyse génomique il arrive que l'on filtre un SNP s'il est en fort déséquilibre de liaison avec un autre SNP voisin (technique appelée LD pruning), en d'autres termes si les deux SNPs ont été transmis vraisemblablement ensemble – transmission multilocus donc. Ainsi, dans cet exemple, nous pourrions, à travers les informations qu'apporterait l'IBD sur la transmission multilocus, adapter le LD pruning en définissant, notamment, ce que sont des "SNPs voisins" en terme de génétique des populations multilocus : en fonction de l'histoire de la population, deux SNPs peuvent être considérés comme voisins, ou non. Ce chapitre purement théorique illustre le fait qu'il est possible de prédire mathématiquement dans un cas simple une mesure multilocus concernant la transmission génétique.

Le troisième chapitre a été l'occasion d'illustrer un travail de modélisation et de méthodologie menées conjointement, sous une approche particulière dite backward. Cette façon différente de procéder apporte d'autres questions et

d'autres méthodes qui permettent de rendre compte de toute la diversité des approches possibles. Nous avons construit un objet, le cluster, et nous avons pu prédire mathématiquement le nombre de clusters dans un régime stationnaire, en bordure de chromosome dans un cas dit neutre. Nous avons cherché à connaitre, de façon plus appliquée, comment cette mesure se comporte en présence de sélection sous différentes formes. Nous avons pu conclure que cette mesure permet de détecter une sélection là où il est généralement difficile de la détecter. Ce chapitre a donc été l'occasion de mener un travail méthodologique sur un objet développé juste en amont, objet alliant simplicité et multilocus.

Enfin, dans le quatrième chapitre nous avons mené un travail purement méthodologique sur les autocorrélogrammes : nous avons décrit, sans développer, la mesure pour l'étudier sous certains angles. Nous avons pu voir que l'autocorrélogramme est un objet simple à manipuler, et qui permet de détecter facilement une sélection. Il ne serait pas étonnant qu'une telle mesure permettent de détecter des sélections épistatiques, mais la mise en œuvre d'un tel travail méthodologique est très compliquée, à cause de l'augmentation du nombre de paramètres (l'emplacement des locus, les comportements conjoints...).

À travers ces trois projets, nous avons essayé d'illustrer la diversité des méthodologies et des développements, tout en suivant un fil directeur commun aux trois projet qu'est le concept de locus voisins, d'allèles voisins et d'haplotypes – ces trois concepts étant fortement intriqués. Chaque mesure exploite l'un des concepts, et ces concepts sont au cœur du concept de transmission génétique "en bloc". Nous avons décidé de caractériser cette transmission à travers des mesures multilocus car il nous a semblé que l'aspect "en

bloc" a beaucoup trop souvent été négligé au profit d'approches monolocus ou considérant des locus dits indépendants – ce qui est absurde puisqu'il ne peut y avoir de locus indépendants selon notre définition –, alors que cette transmission "en bloc" nous semblait être bien plus naturaliste et réaliste que de considérer des locus indépendants. En outre, nous nous sommes assurés que ces projets soient tous ancrés sur la même base théorique que nous avons essayé de développer, dans le premier chapitre, d'*une* génétique des populations, pour "parler" la même théorie dans tous les projets mais sous différents angles.

Enfin, nous pouvons nous rendre compte qu'il y a une question que nous n'avons pas tout au long de notre thèse abordée : celle de savoir en quoi rendre compte de l'aspect multilocus est souhaitable, comparativement à d'autres aspects ; pourquoi faire cet effort supplémentaire, et dans cette orientation en particulier ? Nous n'avons justifié l'intérêt des mesures multilocus qu'à travers l'existence de données multilocus. Nous n'avons pas remis en question l'intérêt de recueillir de telles données en premier lieu, et pourquoi elles étaient intéressantes à expliquer, à rendre compte : nous avons supposé, en accord avec le domaine de la génétique des populations qu'il est intrinsèquement intéressant d'expliquer et de prédire les données multilocus. Nous pensons qu'il serait intéressant dans une biologie de plus en plus fragmentée de s'interroger sur la priorité des aspects à expliquer, et de savoir à quel point on souhaite collecter des données multilocus plutôt que des données, par exemple, épigénétiques. La transmission multilocus nous semble plus réaliste qu'une transmission non multilocus, mais pas plus que tous les autres aspects que nous avons écartés dans le premier chapitre sans véritable raison apparente. Nous ne disons pas que les données multilocus sont inintéressantes,

mais qu'il ne serait pas absurde de se questionner en quoi elles sont intéressantes : permettent-elles une meilleure prédiction ? Sont-elles plus faciles à exploiter, à expliquer ? Pour répondre à de telles questions, et pour confirmer ou infirmer l'intérêt d'utiliser de telles données, il ne semble y avoir, au bout du compte, qu'une seule solution : evisager l'interdisciplinarité.

Bibliographie

Albrechtsen A., Sand Korneliussen T., Moltke I., van Overseem Hansen T., Nielsen F. C. et Nielsen R. (2009), 'Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium', *Genet. Epidemiol.* **33**(3), 266–274.

Auton A., Bryc K., Boyko A. R., Lohmueller K. E., Novembre J., Reynolds A. et al. (2009), 'Global distribution of genomic diversity underscores rich complex history of continental human populations', *Genome Res.* **19**(5), 795–803.

Baird S. J. E. (1995), 'A Simulation Study of Multilocus Clines', *Evolution* **49**(6), 1038–1045.

Ball F. et Stefanov V. T. (2005), 'Evaluation of identity-by-descent probabilities for half-sibs on continuous genome', *Mathematical Biosciences* **196**(2), 215–225.

Beckmann L., Fischer C., Obreiter M., Rabes M. et Chang-Claude J. (2005), 'Haplotype-sharing analysis using Mantel statistics for combined genetic effects', *BMC Genetics* **6**(1), 1–5.

Begun D. J., Holloway A. K., Stevens K., Hillier L. W., Poh Y.-P., Hahn

M. W. et al. (2007), 'Population Genomics : Whole-Genome Analysis of Polymorphism and Divergence in Drosophila simulans', *PLOS Biology* **5**(11), e310.

Behjati S. et Tarpey P. S. (2013), 'What is next generation sequencing ?', *Archives of Disease in Childhood. Education and Practice Edition* **98**(6), 236–238.

Bersaglieri T., Sabeti P. C., Patterson N., Vanderploeg T., Schaffner S. F., Drake J. A. et al. (2004), 'Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene', *The American Journal of Human Genetics* **74**(6), 1111–1120.

Bickeböller H. et Thompson E. A. (1996*a*), 'Distribution of Genome Shared IBD by Half-Sibs : Approximation by the Poisson Clumping Heuristic', *Theoretical Population Biology* **50**(1), 66–90.

Bickeböller H. et Thompson E. A. (1996*b*), 'The Probability Distribution of the Amount of an Individual's Genome Surviving to the Following Generation', *Genetics* **143**(2), 1043–1049.

Blue E. M., Cheung C. Y., Glazner C. G., Conomos M. P., Lewis S. M., Sverdlov S. et al. (2014), 'Identity-by-descent graphs offer a flexible framework for imputation and both linkage and association analyses', *BMC Proceedings* **8**(Suppl 1), S19.

Boitard S. et Loisel P. (2007), 'Probability distribution of haplotype frequencies under the two-locus Wright–Fisher model by diffusion approximation', *Theoretical Population Biology* **71**(3), 380–391.

Bonhomme M., Chevalet C., Servin B., Boitard S., Abdallah J., Blott S. et al. (2010), 'Detecting Selection in Population Trees : The Lewontin and Krakauer Test Extended', *Genetics* **186**(1), 241–262.

Bosse M., Megens H.-J., Madsen O., Paudel Y., Frantz L. A. F., Schook L. B. et al. (2012), 'Regions of Homozygosity in the Porcine Genome : Consequence of Demography and the Recombination Landscape', *PLOS Genet* **8**(11), e1003100.

Browning B. L. et Browning S. R. (2007), 'Efficient multilocus association testing for whole genome association studies using localized haplotype clustering', *Genet. Epidemiol.* **31**(5), 365–375.

Browning B. L. et Browning S. R. (2011), 'A Fast, Powerful Method for Detecting Identity by Descent', *The American Journal of Human Genetics* **88**(2), 173–182.

Browning B. L. et Browning S. R. (2013*a*), 'Detecting Identity by Descent and Estimating Genotype Error Rates in Sequence Data', *The American Journal of Human Genetics* **93**(5), 840–851.

Browning B. L. et Browning S. R. (2013*b*), 'Improving the Accuracy and Efficiency of Identity-by-Descent Detection in Population Data', *Genetics* **194**(2), 459–471.

Browning S. et Browning B. L. (2002), 'On Reducing the Statespace of Hidden Markov Models for the Identity by Descent Process', *Theoretical Population Biology* **62**(1), 1–8.

Browning S. R. (2008), 'Estimation of Pairwise Identity by Descent From

Dense Genetic Marker Data in a Population Sample of Haplotypes', *Genetics* **178**(4), 2123–2132.

Browning S. R. et Browning B. L. (2012), 'Identity by Descent Between Distant Relatives : Detection and Applications', *Annual Review of Genetics* **46**(1), 617–633.

Browning S. R. et Browning B. L. (2015), 'Accurate Non-parametric Estimation of Recent Effective Population Size from Segments of Identity by Descent', *The American Journal of Human Genetics* **97**(3), 404–418.

Browning S. R. et Thompson E. A. (2012), 'Detecting Rare Variant Associations by Identity-by-Descent Mapping in Case-Control Studies', *Genetics* **190**(4), 1521–1531.

Caballero A. et Toro M. A. (2000), 'Interrelations between effective population size and other pedigree tools for the management of conserved populations', *Genetics Research* **75**(03), 331–343.

Cannings C. (2003), 'The Identity by Descent Process along the Chromosome', *Human Heredity* **56**(1-3), 126–130.

Carmi S., Palamara P. F., Vacic V., Lencz T., Darvasi A. et Pe'er I. (2013), 'The Variance of Identity-by-Descent Sharing in the Wright–Fisher Model', *Genetics* **193**(3), 911–928.

Carmi S., Wilton P. R., Wakeley J. et Pe'er I. (2014), 'A renewal theory approach to IBD sharing', *Theoretical population biology* **97**, 35–48.

Chapman N. H. et Thompson E. A. (2002*a*), 'The effect of population history

on the lengths of ancestral chromosome segments', *Genetics* **162**(1), 449–458.

Chapman N. H. et Thompson E. A. (2002*b*), 'The Effect of Population History on the Lengths of Ancestral Chromosome Segments', *Genetics* **162**(1), 449–458.

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

Charlier C., Farnir F., Berzi P., Vanmanshoven P., Brouwers B., Vromans H. et al. (1996), 'Identity-by-descent mapping of recessive traits in livestock : application to map the bovine syndactyly locus to chromosome 15.', *Genome Res.* **6**(7), 580–589.

Chen X., Ma Z.-M. et Wang Y. (2014), 'Markov jump processes in modeling coalescent with recombination', *Ann. Statist.* **42**(4), 1361–1393.

Cheng R., Ma J. Z., Wright F. A., Lin S., Gao X., Wang D. et al. (2003), 'Nonparametric Disequilibrium Mapping of Functional Sites Using Haplotypes of Multiple Tightly Linked Single-Nucleotide Polymorphism Markers', *Genetics* **164**(3), 1175–1187.

Clark A. G. (2004), 'The role of haplotypes in candidate gene studies', *Genetic Epidemiology* **27**(4), 321–333.

Cobbs G. (1978), 'Renewal Process Approach to the Theory of Genetic Linkage : Case of No Chromatid Interference', *Genetics* **89**(3), 563–581.

Coppieters F., Van Schil K., Bauwens M., Verdin H., De Jaegher A., Syx D. et al. (2014), 'Identity-by-descent-guided mutation analysis and exome sequencing in consanguineous families reveals unusual clinical and molecular findings in retinal dystrophy', *Genet Med* **16**(9), 671–680.

Cox D. R. (1962), *Renewal Theory*, Methuen.

Curik I., Ferenčaković M. et Sölkner J. (2014), 'Inbreeding and runs of homozygosity : A possible solution to an old problem', *Livestock Science* **166**, 26–34.

Debouzie D. (1999), 'La notion de population en dynamique et génétique des populations', *Natures Sciences Sociétés* **7**(4), 19–26.

Donnelly K. P. (1983), 'The probability that related individuals share some section of genome identical by descent', *Theoretical Population Biology* **23**(1), 34–63.

Double M. C., Peakall R., Beck N. R., Cockburn A. et Sorenson M. (2005), 'Dispersal, philopatry, and infidelity : dissecting local genetic structure in superb fairy-wrens (malurus cyaneus)', *Evolution* **59**(3), 625–635.

Drummond A. J., Suchard M. A., Xie D. et Rambaut A. (2012), 'Bayesian phylogenetics with BEAUti and the BEAST 1.7', *Molecular biology and evolution* **29**(8), 1969–1973.

Durrant C. et Morris A. P. (2005), 'Linkage disequilibrium mapping via cladistic analysis of phase-unknown genotypes and inferred haplotypes in the Genetic Analysis Workshop 14 simulated data', *BMC Genetics* **6**(1), 1–5.

Dutheil J. Y., Ganapathy G., Hobolth A., Mailund T., Uyenoyama M. K. et Schierup M. H. (2009), 'Ancestral Population Genomics : The Coalescent Hidden Markov Model Approach', *Genetics* **183**(1), 259–274.

Eriksson J., Larson G., Gunnarsson U., Bed'hom B., Tixier-Boichard M., Strömstedt L. et al. (2008), 'Identification of the Yellow Skin Gene Reveals a Hybrid Origin of the Domestic Chicken', *PLOS Genet* **4**(2), e1000010.

Eyre-Walker A. (2000), 'Do mitochondria recombine in humans ?', *Philosophical Transactions of the Royal Society B : Biological Sciences* **355**(1403), 1573–1580.

Fagny M., Patin E., Enard D., Barreiro L. B., Quintana-Murci L. et Laval G. (2014), 'Exploring the Occurrence of Classic Selective Sweeps in Humans Using Whole-Genome Sequencing Data Sets', *Mol Biol Evol* **31**(7), 1850–1868.

Falush D., Stephens M. et Pritchard J. K. (2003), 'Inference of Population Structure Using Multilocus Genotype Data : Linked Loci and Correlated Allele Frequencies', *Genetics* **164**(4), 1567–1587.

Fariello M. I., Boitard S., Naya H., SanCristobal M. et Servin B. (2013), 'Detecting Signatures of Selection Through Haplotype Differentiation Among Hierarchically Structured Populations', *Genetics* **193**(3), 929–941.

Ferrer-Admetlla A., Leuenberger C., Jensen J. D. et Wegmann D. (2016), 'An Approximate Markov Model for the Wright Fisher Diffusion and Its Application to Time Series Data', *Genetics* **203**(2), 831–846.

Ferrer-Admetlla A., Liang M., Korneliussen T. et Nielsen R. (2014), 'On De-

tecting Incomplete Soft or Hard Selective Sweeps Using Haplotype Structure', *Mol Biol Evol* **31**(5), 1275–1291.

Fisher R. A. (1949), 'The theory of inbreeding.', pp. viii + 120 pp.

Fisher R. A. (1954), 'A Fuller Theory of "Junctions" in Inbreeding'.

Fisher R. A. (1959), 'An Algebraically Exact Examination of Junction Formation and Transmission in Parent-offspring Inbreeding'.

Franklin I. R. (1977), 'The distribution of the proportion of the genome which is homozygous by descent in inbred individuals', *Theoretical Population Biology* **11**(1), 60–80.

Freedman A. H., Schweizer R. M., Vecchyo D. O.-D., Han E., Davis B. W., Gronau I. et al. (2016), 'Demographically-Based Evaluation of Genomic Regions under Selection in Domestic Dogs', *PLOS Genet* **12**(3), e1005851.

Gabriel S. B. (2002), 'The Structure of Haplotype Blocks in the Human Genome', *Science* **296**(5576), 2225–2229.

Garte S. (2003), 'Locus-specific genetic diversity between human populations : An analysis of the literature', *Am. J. Hum. Biol.* **15**(6), 814–823.

Gauvin H., Moreau C., Lefebvre J.-F., Laprise C., Vézina H., Labuda D. et al. (2014), 'Genome-wide patterns of identity-by-descent sharing in the French Canadian founder population', *Eur J Hum Genet* **22**(6), 814–821.

Ghosh A. P. (2010), 'Backward and forward equations for diffusion processes', *Wiley Encyclopedia of Operations Research and Management Science (EORMS), Hoboken, NJ : Wiley* .

Gibson F. et Froguel P. (2004), 'Genetics of the APM1 Locus and Its Contribution to Type 2 Diabetes Susceptibility in French Caucasians', *Diabetes* **53**(11), 2977–2983.

Gittleman J. L. et Kot M. (1990), 'Adaptation : Statistics and a Null Model for Estimating Phylogenetic Effects', *Syst Biol* **39**(3), 227–241.

Golding G. B. et Strobeck C. (1980), 'Linkage disequilibrium in a finite population that is partially selfing', *Genetics* **94**(3), 777–789.

Goldringer I. et Bataillon T. (2004), 'On the Distribution of Temporal Variations in Allele Frequency', *Genetics* **168**(1), 563–568.

Gompert Z., Parchman T. L. et Buerkle C. A. (2012), 'Genomics of isolation in hybrids', *Philosophical Transactions of the Royal Society of London B : Biological Sciences* **367**(1587), 439–450.

González-Tortuero E., Rusek J., Maayan I., Petrusek A., Piálek L., Laurent S. et al. (2016), 'Genetic diversity of two Daphnia-infecting microsporidian parasites, based on sequence variation in the internal transcribed spacer region', *Parasites & Vectors* **9**, 293.

Gutiérrez S., González-Cerón L., Montoya A., Sandoval M. A., Tórres M. E. et Cerritos R. (2016), 'Genetic structure of Plasmodium vivax in Nicaragua, a country in the control phase, based on the carboxyl terminal region of the merozoite surface protein-1', *Infection, Genetics and Evolution* **40**, 324–330.

Hahn M. W. (2006), 'Accurate Inference and Estimation in Population Genomics', *Mol Biol Evol* **23**(5), 911–918.

Hahn M. W. (2008), 'Toward a Selection Theory of Molecular Evolution', *Evolution* **62**(2), 255–265.

Haldane J. (1919), 'The combination of linkage values, and the calculation of distances between the loci of linked factors', *Genetics* **8**, 299–309.

Haldane J. B. S. et Waddington C. H. (1931), 'Inbreeding and linkage', *Genetics* **16**(4), 357.

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

Harris K., Sheehan S., Kamm J. A. et Song Y. S. (2014), Decoding coalescent hidden Markov models in linear time, *in* 'Research in Computational Molecular Biology', Springer, pp. 100–114.

Hartl D. L., Clark A. G. et Clark A. G. (1997), *Principles of population genetics*, Vol. 116, Sinauer associates Sunderland.

Hayes B. J., Goddard M. E. et others (2001), 'Prediction of total genetic value using genome-wide dense marker maps', *Genetics* **157**(4), 1819–1829.

Hayes B. J., Visscher P. M. et Goddard M. E. (2009), 'Increased accuracy of artificial selection by using the realized relationship matrix', *Genetics Research* **91**(01), 47.

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

Hedrick P. W. et Lacy R. C. (2015), 'Measuring Relatedness between Inbred Individuals', *J Hered* **106**(1), 20–25.

Hellmig S., Mascheretti S., Renz J., Frenzel H., Jelschen F., Rehbein J. et al. (2005), 'Haplotype analysis of the CD11 gene cluster in patients with chronic Helicobacter pylori infection and gastric ulcer disease', *Tissue Antigens* **65**(3), 271–274.

Hill W. G. et Hernández-Sánchez J. (2007), 'Prediction of Multilocus Identity-by-Descent', *Genetics* **176**(4), 2307–2315.

Hill W. et Weir B. (2011), 'Variation in actual relationship as a consequence of Mendelian sampling and linkage', *Genetics research* **93**(1), 47–64.

Hohenlohe P. A., Phillips P. C. et Cresko W. A. (2010), 'USING POPULATION GENOMICS TO DETECT SELECTION IN NATURAL POPULATIONS : KEY CONCEPTS AND METHODOLOGICAL CONSIDERATIONS', *Int J Plant Sci* **171**(9), 1059–1071.

Hu X.-S., Yeh F. C. et Wang Z. (2011*a*), 'Structural Genomics : Correlation Blocks, Population Structure, and Genome Architecture', *Curr Genomics* **12**(1), 55–70.

Hu X.-S., Yeh F. C. et Wang Z. (2011*b*), 'Structural Genomics : Correlation Blocks, Population Structure, and Genome Architecture', *Curr Genomics* **12**(1), 55–70.

Innan H., Padhukasahasram B. et Nordborg M. (2003), 'The Pattern of Polymorphism on Human Chromosome 21', *Genome Res* **13**(6a), 1158–1168.

Iyengar V. K., Reeve H. K. et Eisner T. (2002), 'Paternal inheritance of a female moth's mating preference', *Nature* **419**(6909), 830–832.

Johnston S. E., Bérénos C., Slate J. et Pemberton J. M. (2016), 'Conserved Genetic Architecture Underlying Individual Recombination Rate Variation in a Wild Population of Soay Sheep (Ovis aries)', *Genetics* **203**(1), 583–598.

Kaplan N. L., Darden T. et Hudson R. R. (1988), 'The coalescent process in models with selection.', *Genetics* **120**(3), 819–829.

Kaplan N. L., Hudson R. R. et Langley C. H. (1989), 'The "hitchhiking effect" revisited.', *Genetics* **123**(4), 887–899.

Kardos M., Qvarnström A. et Ellegren H. (2017), 'Inferring Individual Inbreeding and Demographic History from Segments of Identity by Descent in Ficedula Flycatcher Genome Sequences', *Genetics* p. genetics.116.198861.

Karlin S. et Taylor H. E. (1981), *A Second Course in Stochastic Processes*, Elsevier.

Kim Y. et Nielsen R. (2004), 'Linkage Disequilibrium as a Signature of Selective Sweeps', *Genetics* **167**(3), 1513–1524.

Kim Y. et Stephan W. (2002), 'Detecting a Local Signature of Genetic Hitchhiking Along a Recombining Chromosome', *Genetics* **160**(2), 765–777.

Knief U., Kempenaers B. et Forstmeier W. (2017), 'Meiotic recombination shapes precision of pedigree- and marker-based estimates of inbreeding', *Heredity* **118**(3), 239–248.

Lercher M. J., Urrutia A. O. et Hurst L. D. (2002), 'Clustering of housekeeping genes provides a unified model of gene order in the human genome', *Nat Genet* **31**(2), 180–183.

Leutenegger A.-L., Prum B., Génin E., Verny C., Lemainque A., Clerget-Darpoux F. et al. (2003), 'Estimation of the Inbreeding Coefficient through Use of Genomic Data', *The American Journal of Human Genetics* **73**(3), 516–523.

Levy A. A. et Feldman M. (2004), 'Genetic and epigenetic reprogramming of the wheat genome upon allopolyploidization', *Biological Journal of the Linnean Society* **82**(4), 607–613.

Lewontin R. C. et Krakauer J. (1973), 'Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms', *Genetics* **74**(1), 175–195.

Li C., Wang X., Cai H., Fu Y., Luan Y., Wang W. et al. (2016), 'Molecular microevolution and epigenetic patterns of the long non-coding gene H19 show its potential function in pig domestication and breed divergence', *BMC Evolutionary Biology* **16**, 87.

Li X., Jian Y., Xie C., Wu J., Xu Y. et Zou C. (2017), 'Fast diffusion of domesticated maize to temperate zones', *Scientific Reports* **7**.

Li Y., Sung W.-K. et Liu J. J. (2007), 'Association Mapping via Regularized Regression Analysis of Single-Nucleotide–Polymorphism Haplotypes in Variable-Sized Sliding Windows', *The American Journal of Human Genetics* **80**(4), 705–715.

Liang M. et Nielsen R. (2014), 'The Lengths of Admixture Tracts', *Genetics* **197**(3), 953–967.

Librado P. et Rozas J. (2009), 'DnaSP v5 : a software for comprehensive analysis of DNA polymorphism data', *Bioinformatics* **25**(11), 1451–1452.

Lin S., Chakravarti A. et Cutler D. J. (2004), 'Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies', *Nat Genet* **36**(11), 1181–1188.

Lo C.-L., Lossie A. C., Liang T., Liu Y., Xuei X., Lumeng L. et al. (2016), 'High Resolution Genomic Scans Reveal Genetic Architecture Controlling Alcohol Preference in Bidirectionally Selected Rat Model', *PLOS Genet* **12**(8), e1006178.

Mantel N. (1967), 'The Detection of Disease Clustering and a Generalized Regression Approach', *Cancer Res* **27**(2 Part 1), 209–220.

Mardis E. R. (2008), 'The impact of next-generation sequencing technology on genetics', *Trends in Genetics* **24**(3), 133–141.

Markianos K., Bischoff E., Mitri C., Guelbeogo W. M., Gneme A., Eiglmeier K. et al. (2016), 'Genetic Structure of a Local Population of the Anopheles gambiae Complex in Burkina Faso', *PLOS ONE* **11**(1), e0145308.

Martin O. C. (2006), 'Two- and Three-Locus Tests for Linkage Analysis Using Recombinant Inbred Lines', *Genetics* **173**(1), 451–459.

Martin O. C. et Hospital F. (2011), 'Distribution of Parental Genome Blocks in Recombinant Inbred Lines', *Genetics* **189**(2), 645–654.

Mathias R. A., Gao P., Goldstein J. L., Wilson A. F., Pugh E. W., Furbert-Harris P. et al. (2006), 'A graphical assessment of p-values from sliding window haplotype tests of association to identify asthma susceptibility loci on chromosome 11q', *BMC Genetics* **7**(1), 38.

McQuillan R., Leutenegger A.-L., Abdel-Rahman R., Franklin C. S., Pericic M., Barac-Lauc L. et al. (2008), 'Runs of Homozygosity in European Populations', *The American Journal of Human Genetics* **83**(3), 359–372.

McVean G. A. T. et Cardin N. J. (2005), 'Approximating the coalescent with recombination', *Philosophical Transactions of the Royal Society B : Biological Sciences* **360**(1459), 1387–1393.

Meng Z., Zaykin D. V., Xu C.-F., Wagner M. et Ehm M. G. (2003), 'Selection of Genetic Markers for Association Analyses, Using Linkage Disequilibrium and Haplotypes', *The American Journal of Human Genetics* **73**(1), 115–130.

Meuwissen T. M. H. et Goddard M. E. (2001), 'Prediction of identity by descent probabilities from marker-haplotypes', *Genetics Selection Evolution* **33**(6), 605–634.

Meyer R. S., Choi J. Y., Sanches M., Plessis A., Flowers J. M., Amas J. et al. (2016), 'Domestication history and geographical adaptation inferred from a SNP map of African rice', *Nat Genet* **advance online publication**.

Mitra S. (1975), 'On Nei and Roychoudhury's Sampling Variances of Heterozygosity and Genetic Distance', *Genetics* **80**(1), 223–226.

Mondal M., Casals F., Xu T., Dall'Olio G. M., Pybus M., Netea M. G. et al. (2016), 'Genomic analysis of Andamanese provides insights into ancient human migration into Asia and adaptation', *Nat Genet* **advance online publication**.

Moore R. C. et Stevens M. H. H. (2008), 'Local Patterns of Nucleotide Poly-

morphism Are Highly Variable in the Selfing Species Arabidopsis thaliana', *J Mol Evol* **66**(2), 116.

Mueller J. C. et Andreoli C. (2004), 'Plotting haplotype-specific linkage disequilibrium patterns by extended haplotype homozygosity', *Bioinformatics* **20**(5), 786–787.

Muranty H., Jorge V., Bastien C., Lepoittevin C., Bouffier L. et Sanchez L. (2014), 'Potential for marker-assisted selection for forest tree breeding : lessons from 20 years of MAS in crops', *Tree Genetics & Genomes* **10**(6), 1491–1510.

Nakajima T., Wooding S., Sakagami T., Emi M., Tokunaga K., Tamiya G. et al. (2004), 'Natural Selection and Population History in the Human Angiotensinogen Gene (AGT) : 736 Complete AGT Sequences in Chromosomes from Around the World', *The American Journal of Human Genetics* **74**(5), 898–916.

Nei M. et Roychoudhury A. K. (1974), 'Sampling variances of heterozygosity and genetic distance', *Genetics* **76**(2), 379–390.

Nielsen R. (2001), 'Statistical tests of selective neutrality in the age of genomics', *Heredity* **86**(6), 641–647.

Ohashi J., Naka I., Patarapotikul J., Hananantachai H., Brittenham G., Looareesuwan S. et al. (2004), 'Extended Linkage Disequilibrium Surrounding the Hemoglobin E Variant Due to Malarial Selection', *The American Journal of Human Genetics* **74**(6), 1198–1208.

Oleksyk T. K., Zhao K., Vega F. M. D. L., Gilbert D. A., O'Brien S. J.

et Smith M. W. (2008), 'Identifying Selected Regions from Heterozygosity and Divergence Using a Light-Coverage Genomic Dataset from Two Human Populations', *PLOS ONE* **3**(3), e1712.

Paape T., Zhou P., Branca A., Briskine R., Young N. et Tiffin P. (2012), 'Fine scale population recombination rates, hotspots and correlates of recombination in the Medicago truncatula genome', *Genome Biol Evol* p. evs046.

Palamara P. F. (2014), 'Population genetics of identity by descent', *arXiv preprint arXiv :1403.4987* .

Palamara P. F., Lencz T., Darvasi A. et Pe'er I. (2012), 'Length Distributions of Identity by Descent Reveal Fine-Scale Demographic History', *The American Journal of Human Genetics* **91**(5), 809–822.

Park D. S., Baran Y., Hormozdiari F., Eng C., Torgerson D. G., Burchard E. G. et al. (2015), 'PIGS : improved estimates of identity-by-descent probabilities by probabilistic IBD graph sampling', *BMC Bioinformatics* **16**(5), 1–12.

Paschou P., Feng Y., Pakstis A. J., Speed W. C., DeMille M. M., Kidd J. R. et al. (2004), 'Indications of Linkage and Association of Gilles de la Tourette Syndrome in Two Independent Family Samples : 17q25 Is a Putative Susceptibility Region', *The American Journal of Human Genetics* **75**(4), 545–560.

Pemberton T., Absher D., Feldman M., Myers R., Rosenberg N. et Li J. (2012), 'Genomic Patterns of Homozygosity in Worldwide Human Populations', *The American Journal of Human Genetics* **91**(2), 275–292.

Pritchard J. K., Stephens M. et Donnelly P. (2000), 'Inference of Population Structure Using Multilocus Genotype Data', *Genetics* **155**(2), 945–959.

Rodolphe F., Martin J. et Della-Chiesa E. (2008), 'Theoretical description of chromosome architecture after multiple back-crossing', *Theoretical Population Biology* **73**(2), 289–299.

Rogers A. R. (2014), 'How population growth affects linkage disequilibrium', *Genetics* **197**(4), 1329–1341.

Ross J. A., Koboldt D. C., Staisch J. E., Chamberlin H. M., Gupta B. P., Miller R. D. et al. (2011), 'Caenorhabditis briggsae Recombinant Inbred Line Genotypes Reveal Inter-Strain Incompatibility and the Evolution of Recombination', *PLOS Genetics* **7**(7), e1002174.

Rousset F. (2002), 'Inbreeding and relatedness coefficients : what do they measure ?', *Heredity* **88**(5), 371–380.

Rozas J. et Rozas R. (1999), 'DnaSP version 3 : an integrated program for molecular population genetics and molecular evolution analysis.', *Bioinformatics* **15**(2), 174–175.

Sabeti P. C., Reich D. E., Higgins J. M., Levine H. Z. P., Richter D. J., Schaffner S. F. et al. (2002), 'Detecting recent positive selection in the human genome from haplotype structure', *Nature* **419**(6909), 832.

Sahasrabudhe R. M., Lott P., Ruiz-Ponte C., Teixeira M. et Carvajal-Carmona L. G. (2014), 'Abstract 1281 : Identification of novel susceptibility genes in familial gastric cancer using next generation sequencing and identity-by-descent mapping', *Cancer Res* **74**(19 Supplement), 1281–1281.

Santiago E. et Caballero A. (2005), 'Variation After a Selective Sweep in a Subdivided Population', *Genetics* **169**(1), 475–483.

Schlötterer C. (2002), 'A Microsatellite-Based Multilocus Screen for the Identification of Local Selective Sweeps', *Genetics* **160**(2), 753–763.

Schlötterer C. (2003), 'Hitchhiking mapping – functional genomics from the population genetics perspective', *Trends in Genetics* **19**(1), 32–38.

Schlötterer C. et Dieringer D. (2013), *A Novel Test Statistic for the Identification of Local Selective Sweeps Based on Microsatellite Gene Diversity*, Landes Bioscience.

Schmid K. et Yang Z. (2008), 'The Trouble with Sliding Windows and the Selective Pressure in BRCA1', *PLOS ONE* **3**(11), e3746.

Schork N. J. (1993), 'Extended multipoint identity-by-descent analysis of human quantitative traits : efficiency, power, and modeling considerations.', *Am J Hum Genet* **53**(6), 1306–1319.

Sellars W. et al. (1956), 'Empiricism and the philosophy of mind', *Minnesota studies in the philosophy of science* **1**(19), 253–329.

Shah N., Hirakawa H., Kusakabe S., Sandal N., Stougaard J., Schierup M. H. et al. (2016), 'High-resolution genetic maps of Lotus japonicus and L. burttii based on re-sequencing of recombinant inbred lines', *DNA Res* p. dsw033.

Skipper L., Wilkes K., Toft M., Baker M., Lincoln S., Hulihan M. et al. (2004), 'Linkage Disequilibrium and Association of MAPT H1 in Parkinson Disease', *The American Journal of Human Genetics* **75**(4), 669–677.

Smirnov D., Bruzel A., Morley M. et Cheung V. G. (2004), 'Direct IBD mapping : identical-by-descent mapping without genotyping', *Genomics* **83**(2), 335–345.

Smouse P. E. et Peakall R. (1999), 'Spatial autocorrelation analysis of individual multiallele and multilocus genetic structure', *Heredity* **82**(5), 561–573.

Spielman R. S., McGinnis R. E. et Ewens W. J. (1993), 'Transmission test for linkage disequilibrium : the insulin gene region and insulin-dependent diabetes mellitus (IDDM).', *American journal of human genetics* **52**(3), 506.

Stainton J., Charlesworth B., Haley C., Kranis A., Watson K. et Wiener P. (2016), 'Use of high-density SNP data to identify patterns of diversity and signatures of selection in broiler chickens', *J. Anim. Breed. Genet.* pp. n/a–n/a.

Stam P. (1980), 'The distribution of the fraction of the genome identical by descent in finite random mating populations', *Genetics Research* **35**(02), 131–155.

Staples J., Qiao D., Cho M., Silverman E., Nickerson D. et Below J. (2014), 'PRIMUS : Rapid Reconstruction of Pedigrees from Genome-wide Estimates of Identity by Descent', *The American Journal of Human Genetics* **95**(5), 553–564.

Stefanov V. T. (2000), 'Distribution of genome shared identical by descent by two individuals in grandparent-type relationship.', *Genetics* **156**(3), 1403–1410.

Stinchcombe J. R. et Hoekstra H. E. (2007), 'Combining population geno-

mics and quantitative genetics : finding the genes underlying ecologically important traits', *Heredity* **100**(2), 158–170.

Stölting K. N., Nipper R., Lindtke D., Caseys C., Waeber S., Castiglione S. et al. (2013), 'Genomic scan for single nucleotide polymorphisms reveals patterns of divergence and gene flow between ecologically divergent species', *Molecular Ecology* **22**(3), 842–855.

Sved J. A. (2011), 'The covariance of heterozygosity as a measure of linkage disequilibrium between blocks of linked and unlinked sites in Hapmap', *Genetics Research* **93**(04), 285–290.

Szpiech Z., Xu J., Pemberton T., Peng W., Zöllner S., Rosenberg N. et al. (2013), 'Long Runs of Homozygosity Are Enriched for Deleterious Variation', *Am J Hum Genet* **93**(1), 90–102.

Te Meerman G. J., Van Der Meulen M. A. et Sandkuijl L. A. (1995), 'Perspectives of identity by descent (IBD) mapping in founder populations', *Clinical & Experimental Allergy* **25**, 97–102.

Thompson E. A. (2008), 'The IBD process along four chromosomes', *Theoretical population biology* **73**(3), 369–373.

Thompson E. A. (2013), 'Identity by Descent : Variation in Meiosis, Across Genomes, and in Populations', *Genetics* **194**(2), 301–326.

Thompson E. A. et Chapman N. H. (2002), Haplotype Blocks in Small Populations, *in* S. Istrail, M. Waterman et A. Clark, eds, 'Computational Methods for SNPs and Haplotype Inference', number 2983 *in* 'Lecture Notes in Computer Science', Springer Berlin Heidelberg, pp. 74–83. DOI : 10.1007/978-3-540-24719-7_6.

Thouzeau V. (2017), 'Inférer l'histoire des populations humaines à partir des diversités génétiques et linguistiques', *http ://www.theses.fr/s148585* .

Tiret M. et Hospital F. (2017), 'Blocks of chromosomes identical by descent in a population : Models and predictions', *PLOS ONE* **12**(11), e0187416.

Visscher P. M., Medland S. E., Ferreira M. A. R., Morley K. I., Zhu G., Cornes B. K. et al. (2006), 'Assumption-Free Estimation of Heritability from Genome-Wide Identity-by-Descent Sharing between Full Siblings', *PLOS Genet* **2**(3), e41.

Vitalis R. et Couvet D. (2001), 'Estimation of Effective Population Size and Migration Rate From One- and Two-Locus Identity Measures', *Genetics* **157**(2), 911–925.

Walters K. et Cannings C. (2005), 'The probability density of the total IBD length over a single autosome in unilineal relationships', *Theoretical Population Biology* **68**(1), 55–63.

Wang X., Long Y., Wang N., Zou J., Ding G., Broadley M. R. et al. (2017), 'Breeding histories and selection criteria for oilseed rape in Europe and China identified by genome wide pedigree dissection', *Scientific Reports* **7**.

Waples R. S. (1989), 'A generalized approach for estimating effective population size from temporal changes in allele frequency.', *Genetics* **121**(2), 379–391.

Ward B. J. et van Oosterhout C. (2016), 'hybridcheck : software for the rapid detection, visualization and dating of recombinant regions in genome sequence data', *Mol Ecol Resour* **16**(2), 534–539.

Weir B. S., Avery P. J. et Hill W. G. (1980), 'Effect of mating structure on variation in inbreeding', *Theoretical Population Biology* **18**(3), 396–429.

Weir B. S., Cardon L. R., Anderson A. D., Nielsen D. M. et Hill W. G. (2005), 'Measures of human population structure show heterogeneity among genomic regions', *Genome Res.* **15**(11), 1468–1476.

Weir B. S. et Hill W. G. (1980), 'Effect of mating structure on variation in linkage disequilibrium', *Genetics* **95**(2), 477–488.

White B. J., Hahn M. W., Pombi M., Cassone B. J., Lobo N. F., Simard F. et al. (2007), 'Localization of Candidate Regions Maintaining a Common Polymorphic Inversion (2la) in Anopheles gambiae', *PLOS Genetics* **3**(12), e217.

Williamson E. G. et Slatkin M. (1999), 'Using maximum likelihood to estimate population size from temporal changes in allele frequencies.', *Genetics* **152**(2), 755–761.

Wiuf C. et Hein J. (1997), 'On the Number of Ancestors to a DNA Sequence', *Genetics* **147**(3), 1459–1468.

Wright S. (1921*a*), 'Systems of Mating. I. the Biometric Relations between Parent and Offspring', *Genetics* **6**(2), 111–123.

Wright S. (1921*b*), 'Systems of mating. IV. The effects of selection', *Genetics* **6**(2), 162.

Wright S. (1921*c*), 'Systems of mating. V. General considerations', *Genetics* **6**(2), 167.

Xu, Wiesch et Meyers (1998), 'Genetics of complex human diseases : genome screening, association studies and fine mapping', *Clinical & Experimental Allergy* **28**, 1–5.

Zhao H., Pfeiffer R. et Gail M. H. (2003), 'Haplotype analysis in population genetics and association studies', *Pharmacogenomics* **4**(2), 171–178.

Zheng X. et Weir B. S. (2016), 'Eigenanalysis of SNP data with an identity by descent interpretation', *Theoretical Population Biology* **107**, 65–76.

Ødegård J. et Meuwissen T. H. (2014), 'Identity-by-descent genomic selection using selective and sparse genotyping', *Genetics Selection Evolution* **46**, 3.

université PARIS-SACLAY
ÉCOLE DOCTORALE
Agriculture, alimentation, biologie, environnement, santé (ABIES)

Titre : Approche multilocus du génome dans les modèles de génétique des populations

Mots clés : approche multilocus, génétique des populations, modélisation, IBD, ARG, autocorrélogramme

Résumé : La génétique des populations est l'étude de l'évolution des fréquences alléliques au sein d'une population et de l'influence des pressions évolutives sur ces fréquences. Au sein de cette discipline, des modèles de population et des mesures génétiques sont développés pour pouvoir expliquer et prédire les données génétiques. Toutefois, au fur et à mesure des avancées technologiques, de nouveaux types de données sont disponibles, et il devient primordial de développer de nouveaux modèles et de nouvelles mesures pour pouvoir expliquer ces nouvelles données génétiques, plus denses et plus riches en marqueurs génétiques grâce à l'avènement de techniques comme la Next Generation Sequencing. Pour ce faire, nous proposons dans cette thèse de développer de nouvelles mesures avec une approche dite multilocus, qui considère le génome comme un tout plutôt que comme un agglomérat de locus indépendants. Dans un premier temps, nous avons tenté de construire une base théorique de l'approche multilocus en génétique des populations. Ensuite, nous avons illustré une telle approche à travers l'étude de l'identité par descendance, des graphes de recombinaison ancestraux et des autocorrélogrammes dans les modèles de génétique des populations. À travers ces différentes études de cas, nous avons tenté d'identifier les principaux enjeux et questions que soulève la génétique des populations multilocus.

Title : Multilocus Approach of the genome in the population genetics models

Keywords : multilocus approach, population genetics, modelling, IBD, ARG, autocorrelogram

Abstract : Population genetics is the study of the evolution of allelic frequencies within a population and the influence of evolutionary pressures on these frequencies. Within this field, one could develop population models and measures to explain and predict genetic data. However, as technologie evolves new types of data are available, and it becomes essential to develop new models and new measures to reflect these new genetic marker data, increasingly richer and denser thanks to the advent of new techniques such as the Next Generation Sequencing. To this end, we propose in this thesis to develop new measures with the so-called multilocus approach, which considers the genome as a whole rather than an agglomerate of independent loci. We have first tried to build a theoretical basis for the multilocus approach in population genetics. Then, we have illustrated this multilocus approach with the case studies of identity by descent, ancestral recombination graphs and autocorrelograms in population genetics models. Through these different studies, we tried to identify the main issues and questions that the multilocus population genetics raises.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France