Modélisation de l'impact de la sélection naturelle et culturelle sur la diversité génétique : cas de la transmission du succès reproducteur et des réseaux de gènes

Jean-Tristan Brandenburg

UNIVERSITE PARIS-SUD 11

ÉCOLE DOCTORALE : « *Sciences du Végétal : du gène à l'écosystème* »

Laboratoire Ecologie, Systématique et Evolution (UMR 8079)
Laboratoire Eco-Anthropologie et Ethnobiologie (UMR 7206)

DISCIPLINE : *Génétique des populations*

THÈSE DE DOCTORAT SUR TRAVAUX

Soutenance prévue le 19/12/2011

par

Jean-Tristan Brandenburg

Modélisation de l'impact des sélections naturelle et culturelle sur la diversité génétique : cas de la transmission du succès reproducteur et des réseaux de gènes

Directeur de thèse : Frédéric Austerlitz — Chargé de recherche (CNRS)
Co-directeur de thèse : Bruno Toupance — Maître de conférences (Université Paris Diderot)

Composition du jury :

Rapporteurs :	Emmanuelle Génin	Directrice de Recherche (INSERM)
	Xavier Vekemans	Professeur (Université de Lille 1)
Examinateurs :	Christine Dillmann	Professeure (Université Paris-Sud)
	Maud Tenaillon	Chargée de recherche (CNRS)
	Bruno Toupance	Maître de conférences (Université Paris Diderot)
	Frédéric Austerlitz	Chargé de recherche (CNRS)

Remerciements

Je tiens tout d'abord à remercier mes deux encadrants, sans qui ce travail n'aurait pu voir le jour. Bruno, Fred, merci pour votre encadrement, votre soutien, vos conseils et votre patience.

Je remercie tout particulièrement ceux qui ont suivi mon travail et m'ont conseillé : Michael Blum, Pauline Garnier-Géré, Denis Mestivier et Maud Tenaillon, comme les deux directeurs de l'école doctorale qui m'ont suivi pendant mon parcours : Michel Dron et Jacqui Shykoff.

Je remercie les personnes avec lesquelles j'ai eu l'occasion et le plaisir de collaborer : je pense à Bénédicte Rhoné, Evelyne Heyer, Michaela Leonardi, Marina Ciullo, Anne-Louise Leutenegger et Catherine Bourgain.

Je remercie, par ordre alphabétique, Aurélie D., Bénédicte R., Cindy A. et Marine L. et. Maya L. pour la relecture de certaines parties de mon manuscrit.

J'ai eu l'occasion de réaliser mon monitorat au sein de l'Université Paris-Sud, sous l'encadrement de Domenica. J'en profite pour la remercier de ses conseils.

Je tiens à remercier Christine Dillmann, Maud Tenaillon d'avoir accepté d'être membres de mon jury et plus particulièrement Emmanuelle Génin et Xavier Vekemans d'être les rapporteurs de mon travail.

Le doctorat ne se résume pas qu'à une expérience professionnelle se valorisant par l'écriture d'articles et d'un manuscrit de thèse. Il s'agit tout autant d'une période de maturation de l'esprit, catalysée par des rencontres et des échanges.

Mon doctorat s'est déroulé entre le laboratoire de Génétique des populations humaines rattaché au département Hommes, Natures, Sociétés (HNS) du Muséum national d'Histoire Naturelle (MNHN) et le laboratoire Ecologie, Systématique et Evolution (ESE) de l'Université de Paris Sud.

En 2007, j'ai intégré pour la première fois, grâce à Bruno, le laboratoire de génétique des populations humaines du Musée de l'Homme dans le cadre d'un stage. En cette occasion, j'ai fait mes premiers pas dans le monde de la recherche et ai pu côtoyer les personnes qui ont accompagné mes premiers pas dans la recherche et tout pendant mon doctorat. Je pense à Begona, Franz, Laure, Paul, Myriam, Nancy, Noémie, Priscille, Patrick P, Renaud, Samuel, Sophie, Sylvain, et bien sûr mes encadrants.

Le musée de l'Homme est une structure qui m'a amené à côtoyer avec plaisir d'autres départements. Au sein de ceux-ci, j'ai eu la joie de connaître : Liliana – que je remercie pour ses organisations d'événements comme pour ses conseils – Philippe et Véronique des collections ; Sylvie en ethnomusicologie ; Julien, Laurent, Patrick S.… chez les préhistoriens ; et enfin les gardiens, dont Christophe que j'ai côtoyé en tant que gardien. Je remercie toutes ces personnes pour leur accueil, leurs conseils et les moments agréables passés ensemble.

Il y a plus de deux ans maintenant, le laboratoire a déménagé du Trocadéro au Jardin des Plantes, où nous avons rejoint physiquement les autres équipes du département HNS. Cette nouvelle localisation m'a permis de découvrir au sein du département des personnalités qui m'ont ouvert aux sciences humaines et sociales. Je pense aux membres de la « cellule psy », structure de partage de connaissance, d'entraide et de conseil composée de doctorants et jeunes docteurs de plusieurs disciplines rattachés au département HNS : Agnès, Aline, Alix, Anne-Claire, Aurélie, Béatrice S., Carla, Cindy, Elise, Emeric, Erik, Florence R., Hermine, Manon, Marine R., Maya, Nicolas M., Noémie, Rihlat, Sandrine, Romain. Cette cellule nous a permis, à notre humble niveau, d'instaurer un dialogue entre différentes disciplines, de partager des connaissances comme de bons moments. Nombre de ses membres sont par ailleurs devenus des amis.

J'ai une pensée pour d'autres membres de notre département : Farida, Florence L., Jeanne L.D., Richard, Serge, Taoues…

Notre arrivée au Jardin des Plantes m'a donné la chance de côtoyer des membres d'autres départements du Muséum dont : Alan, Alix C., Anne(s), Camilia, Cat, Hélène, Jeanne R., Marine L., Nicolas D., Noëlie, Pierre F., Stéphane, Vincent, Yann et tous ceux que j'oublie. Merci pour votre soutien et pour les moments conviviaux passés ensemble.

Et bien sûr, je remercie Evelyne Heyer et Serge Bahuchet (dit « Le grand Manitou ») de m'avoir accepté dans leur unité.

Une autre partie de mon doctorat eu lieu dans le laboratoire de l'ESE, sur le campus d'Orsay, où j'ai également eu l'occasion d'échanger autant scientifiquement qu'amicalement avec de nombreuses personnes. Je pense à ma stagiaire Nathalie, mais aussi aux membres de l'ESE : Alexis, Aliénor, Alodie, Amanda, Amandine, Béatrice A., Benjamin, Charles, Claire, David, Estelle, Jonathan L., Gwendal, Hervé, Julien, Laetitia, Ludwig, Marianne, Mickael, Philippe, Pierre G., Puri, et bien sûr Jacqui Shykoff et Paul Leadley, que je remercie pour leur accueil au sein leur unité.

En deuxième année de doctorat, je me suis investi dans le monde associatif, à travers l'association « Agir pour les Doctorants et jeunes DOCteurs » de l'Université Paris-Sud. ADDOC m'a fait découvrir la « Confédération des Jeunes Chercheurs », me faisant prendre conscience de la nécessité de valoriser le doctorat comme expérience professionnelle. Merci à tous les membres de ces deux associations, et aussi à ceux de l'AMBI, pour tous les projets que nous avons pu concrétiser ensemble.

Enfin, j'aimerais remercier l'ensemble de ma famille surtout mes parents, mes frères, ma sœur comme mes amis, pour leur soutien.

Table des matières

Introduction Générale

L'évolution des populations est un sujet passionnant mais complexe. Comprendre les processus qui ont permis d'arriver à la diversité génétique actuelle est un défi que la génétique des populations a décidé de relever. Celle-ci prend sa source dans certaines découvertes comme les lois de Mendel d'une part et la théorie de l'Évolution et de la sélection naturelle par Charles Darwin (1859) d'autre part. Ces découvertes ont été synthétisées dans la théorie synthétique de l'évolution entre 1930 et 1950 par Theodosius Dobzhansky (1937), Ronald Fisher (1930), J.B.S. Haldane (1932), Sewall Wright (1931), Julian Huxley (1942), Ernst Mayr (1942), Bernhard Rensch, (1947) George Gaylord Simpson (1944) et George Ledyard Stebbins (1950), en y incluant la dérive, les migrations et la mutation. Les quatre forces évolutives, à savoir la sélection (naturelle, artificielle ou culturelle), la dérive, la migration et la mutation, conditionnent en grande partie la diversité génétique observée aujourd'hui dans les populations.

Expliquer les variations génétiques et leurs conséquences dans une population relève cependant d'une problématique qui est loin d'être achevée, car les quatre forces énoncées ci-dessus relèvent de processus complexes. De ce fait, pour comprendre l'évolution des fréquences des variants génétiques au sein des populations, un cadre théorique a été très vite mis en place sur la base de modèles mathématiques. Un des premiers modèles développé en 1908 indépendamment par Godfrey Harold Hardy (1908) et Wilhelm Weinberg (1908) permettait, dans une population de taille infinie, panmictique, sans migration, sans sélection, sans mutation et à générations non chevauchantes, de prédire l'équilibre des fréquences des allèles et des génotypes au sein d'une population en utilisant un modèle probabiliste. Mais au vu de l'ensemble des hypothèses du modèle, il ne peut s'appliquer que dans quelques cas restreints.

L'autre approche qui a été introduite par Sewall Wright et Ronald Fisher (Fisher 1930; Wright 1931) est de modéliser l'évolution d'une population de taille finie au cours des générations successives, selon une approche probabiliste. Ce modèle individu-centré considère une population soumise à des mutations, de taille finie, panmictique, sans migration et sans sélection à générations non chevauchantes. Il a permis entre autres de mettre en évidence l'action de la dérive génétique sur l'évolution des fréquences alléliques,

notamment dans le cadre de la théorie neutraliste de Motoo Kimura (1968; 1983). Il est à noter qu'un modèle à générations chevauchantes reposant sur des principes similaires a été développé par Patrick Moran (1958).

Ultérieurement, c'est en étudiant les propriétés des généalogies de gènes qu'a été développée une approche ascendante du modèle (c'est-à-dire en remontant le temps). Cette approche, dénommée théorie de la coalescence, a été formalisée par John Kingman (1982a; 1982b). Elle se focalise directement sur un échantillon de copies de gène de la population actuelle. Elle étudie les lois de probabilité de la topologie de l'histoire de ces copies de gène (appelé arbre de coalescence ou coalescent) et de la longueur des branches de celui-ci. Notons qu'un travail similaire avait été effectué indépendamment par Tajima (1983). Le modèle de Wright-Fisher demande un temps de simulation long puisqu'il est nécessaire de simuler tous les individus de la population durant un grand nombre de générations, afin d'atteindre l'équilibre. A l'inverse, le modèle de coalescence permet de simuler rapidement les généalogies d'un échantillon de *n* copies de gène données, en tirant les temps de coalescence selon leur loi de probabilité et en ne se focalisant que sur les lignées reliant les *n* copies de gène à leur ancêtre commun. A condition qu'on se restreigne à des locus neutres, les mutations peuvent ensuite être rajoutées indépendamment le long des branches de l'arbre selon un modèle mutationnel choisi. Grâce à des extensions théoriques, ce procédé de simulation permet de considérer non seulement des populations de tailles constantes mais aussi des populations soumises à des processus démographiques comme par exemple des croissances de populations ou des goulots d'étranglement (Griffiths & Tavare 1994; Slatkin & Hudson 1991). Il permet aussi de simuler plusieurs locus liés génétiquement avec des taux de recombinaison variables (Griffiths & Marjoram 1996).

Ces modèles ont donc permis de bien décrire les impacts des processus démographiques et de la recombinaison dans les populations. Mais le modèle de coalescence ne s'adapte pas à des cas plus complexes comme ceux qui découlent de la présence de la sélection, mis à part dans des cas particuliers de sélection faible sur un seul locus (Neuhauser & Krone 1997), où les développements analytiques sont très complexes. Les modèles individu-centrés permettent à l'inverse de simuler des populations soumises à toutes les formes envisageables de sélection naturelle ou culturelle.

Pour en revenir à la sélection naturelle, Darwin considérait que la sélection naturelle était le moteur de l'évolution des espèces. La sélection peut être définie comme la survie et la

reproduction différentielle (positive ou négative) des individus grâce à leur phénotype dans un environnement donné. Ainsi la présence d'individus portant des phénotypes avantageux dans une population va entraîner une différence de reproduction entre eux et les autres individus de la population. La diversité d'un caractère peut être expliquée par un nombre réduit de gènes (on parle alors de caractère à déterminisme simple) ou par un grand nombre de gènes qui vont interagir ensemble (on parle alors de caractère à déterminisme complexe). Dans ce dernier cas autant les différences de niveau d'expression des gènes que l'état allélique de ces gènes vont conditionner les phénotypes chez un individu donné. Par son action, la sélection modifie la diversité génétique des populations. Lorsque les caractères sont à déterminisme simple, la diversité va être modifiée aux locus responsables de la diversité du caractère (voir par exemple Fisher, Fisher 1930). L'impact de la sélection sur la diversité des gènes impliqués dans un caractère à déterminisme complexe a été moins analysée dans un cadre théorique même si on relève quelques études qui l'analysent à l'aide de modèles additifs (Latta 1998; Le Corre & Kremer 2003) (cf. introduction de la partie II pour plus de détails sur ces modèles).

La sélection entraîne une différence de succès reproducteur entre les individus liée à des facteurs transmis de manière héréditaire entre les générations. Or, dans certaines populations humaines et animales, une corrélation entre le succès de reproduction d'un individu et celui de ses parents a été observée (Heyer et al. 2005). Le phénomène peut avoir une part génétique et relèverait donc dans ce cas-là de la transmission d'un ou plusieurs caractères à déterminisme complexe impliqués dans la valeur sélective. La transmission du succès reproducteur peut avoir aussi une part culturelle. On parle alors de sélection culturelle (Cavalli-Sforza & Feldman 1981) lorsque des caractères transmis culturellement confèrent un avantage à ceux qui les portent. La question de la part respective des sélections naturelle et culturelle est un débat qui a commencé à la fin du XIX^{e} siècle et qui est toujours d'actualité (cf. la revue de Murphy 1999). Les facteurs culturels influencent aussi la diversité génétique dans les populations de certains mammifères comme certaines espèces de baleines (Frère et al. 2010), de dauphins (Whitehead 1998), de guépards (Kelly 2001) et surtout dans un certain nombre de populations humaines (Austerlitz & Heyer 1998; Blum et al. 2006). De nombreux phénomènes culturels peuvent modifier la diversité génétique d'une population comme la polygynie (Neel 1970), les migrations sexe spécifiques (Segurel et al. 2008), les différences de règles de descendance (cognatique, matrilinéaire, patrilinéaire Kumar et al. 2006), l'hétérogénéité et la transmission culturelle du succès reproducteur (Sibert et al. 2002). Considérer les impacts de ces forces sur la diversité génétique n'est pas intuitif. Il est

nécessaire de passer par une modélisation stochastique, ce qui a été fait notamment en ce qui concerne les impacts de la transmission du succès reproducteur sur la diversité génétique (Austerlitz & Heyer 1998; Blum et al. 2006; Sibert et al. 2002). Ces études sont présentées plus en détail dans l'introduction de la partie I.

Le travail présenté dans cette thèse consiste à analyser les impacts de certains phénomènes complexes de sélection naturelle ou culturelle sur la diversité génétique, principalement à l'aide d'outils de simulation informatique. Notre travail s'effectue à travers deux exemples : la transmission intergénérationnelle du succès reproducteur et l'évolution des phénotypes à déterminisme complexe, en l'occurrence des phénotypes codés par des réseaux de gènes.

Dans une première partie, nous analyserons l'impact de la transmission du succès reproducteur sur les arbres de coalescence et la diversité génétique à partir d'un modèle individu-centré (Partie I.B), en nous intéressant plus particulièrement aux effets de l'interaction entre ce phénomène et d'autres phénomènes comme l'hétérogénéité du succès reproducteur ou le choix non-aléatoire du conjoint. Nous comparerons aussi les cas où la transmission du succès reproducteur se fait de manière biparentale ou de manière uniparentale (c'est-à-dire uniquement de mères en filles ou de pères en fils). De plus, nous présenterons une méthodologie permettant de détecter ce phénomène à partir des données génétiques. Nous appliquerons cette méthodologie sur des données de polymorphisme génétique des chromosomes autosomaux présents dans la base de données HapMap (Partie I.C.i) ou du chromosome Y sur des données d'Asie centrale (Partie I.C.ii). Pour finir, nous présenterons les effets de la transmission du succès reproducteur non plus sur des généalogies de gènes mais sur des généalogies d'individus, telles qu'elles peuvent être retracées en démographie historique pour des populations disposant de registres d'état civil disponibles. Nous en déduirons des méthodes de détection de ce type de sélection culturelle utilisables dans ces généalogies et les appliquerons aux données généalogiques de la population du Cilento en Italie (Partie I.D).

Dans une seconde partie, nous décrirons l'impact de la sélection sur des caractères à déterminisme complexe. Plus précisément, nous nous intéresserons à des caractères codés par des réseaux de gènes, c'est-à-dire un ensemble de gènes dont le produit régule l'expression des autres gènes ainsi que leur propre expression. A partir d'un modèle individu centré, nous

analyserons l’impact de la sélection tant au niveau du phénotype, que des interactions entre les gènes impliqués dans celui-ci et de la diversité de ceux-ci (partie II.B).

Partie I :

Transmission du succès reproducteur :

impacts et détection

I.A Introduction

Le travail présenté dans cette partie a un double objectif : analyser l'influence de la transmission du succès reproducteur à différents niveaux (génétique ou généalogique) et inversement détecter le phénomène à partir de données mesurables au niveau des populations (polymorphisme génétique ou généalogies ascendantes d'individus). On parle de transmission du succès reproducteur quand les individus ont un succès reproducteur d'autant plus élevé que celui de leur parent était lui-même élevé.

Définir le phénomène

Dès le XIX^{e} siècle, Pearson et *al.* (1899) ont constaté une relation entre les tailles de fratrie des individus et leur nombre d'enfants dans la population anglaise impliquant que les individus transmettaient à leurs enfants leur capacité de se reproduire. En d'autres termes, le nombre de frères et de sœurs d'un individu est corrélé positivement avec son nombre de fils et de filles. Ce phénomène ne peut être observé que s'il y a une variabilité et une transmission du succès reproducteur. Pour comprendre les déterminants de la transmission du succès reproducteur, il est donc nécessaire de comprendre ce qui entraîne une différence de reproduction entre deux individus et pourquoi cette différence se maintient dans les générations suivantes au sein de la population. Les causes de la transmission du succès reproducteur peuvent être génétiques, environnementales ou culturelles.

Transmission culturelle du succès reproducteur

La transmission du succès reproducteur est culturelle lorsque des facteurs culturels ou sociaux entraînent des différences de reproduction entre deux individus et que ces facteurs sont transmis aux générations suivantes, ceci pouvant être dû à plusieurs causes.

Ce sera par exemple le cas dans une population structurée où les individus établis (« cœur » de la population) vont avoir plus d'enfants et se reproduire davantage dans la population que les individus qui arrivent dans la population (« frange » de la population) (Heyer 1993), ce phénomène perdurant au cours des générations. Dans ce cas, la notion de succès reproducteur est étendue et on parle de succès reproducteur « utile », c'est-à-dire que l'on ne comptabilise que les enfants qui restent dans la population et se reproduisent en son sein (dit « enfants utiles », Heyer 1999). Dans ce cas particulier, c'est ce succès reproducteur utile qui est transmis, les enfants issus de la frange de la population émigrant plus que ceux du cœur de la population, et participant de ce fait moins à la reproduction dans la population. Des

raisons socio-économiques ou un accès différentiel aux ressources peuvent expliquer ces différences migratoires comme dans la vallée de Valserine dans le Jura (Heyer 1993) ou dans la population du Saguenay au Québec (Austerlitz & Heyer 1998; Gagnon & Heyer 2001).

Un autre exemple est le rang social dans une population. Le succès reproducteur sera transmis culturellement s'il dépend de ce rang social et que ce rang est transmis entre générations. Ceci pourrait être le cas chez les femmes maoris (Murray-McIntosh et al. 1998) ou chez les descendants mâles de Gengis Khan (Zerjal et al. 2003).

Plus généralement, une différence de succès reproducteur peut être due à des différences de statuts sociaux, de ressources, à la polygamie (Neel 1970) ou à des différentiels de migration.

Nous avons présenté ci-dessus quelques exemples concernant des populations humaines, mais une transmission culturelle du succès reproducteur n'est pas uniquement constatée chez cette espèce. En effet, plusieurs exemples de possible transmission du succès reproducteur ont été relevés chez les animaux. Par exemple chez les baleines matrilinéaires, une différence d'apprentissage, d'occupation des petits et de stratégie de recherche de nourriture pourrait en être à l'origine (Whitehead 1998). Aussi dans les populations de chimpanzés une différence d'apprentissage pourrait être à l'origine d'une transmission du succès reproducteur (Lonsdorf et al. 2004).

Notons finalement que la transmission du succès reproducteur peut être liée aux règles de descendance propres à la population : une hérédité liée à des comportements liés à la mère comme dans des populations de chasseurs cueilleurs (Blum et al. 2006) ou dans les populations de baleine (Whitehead 1998) (matrilinéaire), des comportements liés au père (patrilinéaire) ou des deux parents comme ce qui est observé au Saguenay Lac Saint Jean au Québec (Austerlitz & Heyer 1998) (biparental).

Transmission génétique du succès reproducteur

Une autre origine possible de la transmission du succès reproducteur peut être génétique, ceci impliquant que des différences dans le patrimoine génétique des individus entraînent des succès reproducteurs différents. Ceci implique la présence de polymorphismes qui modifient le succès reproducteur entre les individus et le maintien des différences de succès reproducteur entre les générations.

L’apparition d’un allèle avantageux avant qu’il ne se fixe (dans le cas d'une sélection directionnelle) ou d’un allèle délétère avant qu’il ne disparaisse (dans le cas d'une sélection purificatrice) pourrait entraîner cette variabilité de reproduction entre les individus. Cependant cette variabilité devrait être épuisée par la sélection naturelle. Cela ne sera cependant pas le cas si la valeur sélective est liée à des caractères quantitatifs codés par de nombreux locus, car dans ce cas la mutation peut recréer en permanence de la variabilité à ces locus et de ce fait maintenir une variance héritable dans le succès reproducteur. Par ailleurs si la sélection opère sur de nombreux caractères simultanément, elle ne peut plus être efficace pour éliminer la variabilité à tous ces caractères (Orr 2000) ce qui permet de maintenir une part de variance héritable pour ces caractères liés à la valeur sélective dans les populations.

Par exemple dans la population huttérite nord-américaine, où une transmission du succès reproducteur a été constatée, Pluzhnikov et *al.* (2007) considèrent que l’origine de cette transmission est génétique car aucun élément culturel ne permet d’expliquer la reproduction entre les individus. Par ailleurs, dans la population danoise, une forte composante biologique de la variation du succès reproducteur a été montrée en comparant des jumeaux monozygotes et dizygotes (Kohler et al. 1999; Kohler et al. 2006).

Le débat entre les causes génétiques et culturelles de ce phénomène est un débat qui a commencé au XIX^{e} siècle et reste toujours d’actualité. Il a bien été résumé par Murphy (1999).

Modéliser le phénomène

L’une des utilisations d’un modèle est premièrement de prédire les possibles impacts d’un phénomène et de les confronter à des données réelles ; aussi il permet de déterminer plus spécifiquement l’impact d’un paramètre particulier sur une population et les possibles interactions entre les paramètres. De nombreux modèles stochastiques et individu-centrés neutres ont été développés en génétique des populations.

L’un des modèles de base les plus connus est le modèle neutre de Wright-Fisher (Fisher 1930; Wright 1931) qui considère des générations séparées avec une taille de population constante où tous les individus d’une génération donnée ont la même probabilité d’être parent. À l’inverse, le modèle de Moran considère non plus des générations séparées mais des générations chevauchantes puisqu’à chaque pas de temps un individu de la population est aléatoirement choisi pour mourir alors qu'un autre individu choisi au hasard se reproduit et donne naissance à un descendant qui remplace l'individu supprimé. Pour finir on

peut également citer le modèle de Cannings (1974), où à chaque génération *k* individus sont remplacés par *k* autres après s'être reproduit ou pas. Dans ce cas, il s'agit d'un cas généralisant les deux précédents puisque si *k* = 1, il s'agit du modèle de Moran et si *k* est égal à la taille de la population le modèle est celui de Wright-Fisher (Cannings 1974).

En étudiant les propriétés des arbres des gènes produits dans le cadre du modèle de Wright-Fisher, Kingman a développé une approche ascendante du modèle en ne considérant qu'un échantillon restreint de l'ensemble des copies de gène dans la population et a décrit les caractéristiques probabilistes de l'arbre généalogique de ces copies de gène (appelé arbre de coalescence), en ce qui concerne notamment sa topologie et ses longueurs de branches (liées aux temps de coalescence) (Kingman 1982a; Kingman 1982b). Bien que ce type d'approche ascendante ait été également développé par Tajima (Tajima 1983), c'est la formalisation de Kingman qui s'est révélée la plus féconde en génétique des populations dans la mesure où elle permet de modéliser rapidement des histoires de gènes de population pour des scénarios démographiques complexes. Néanmoins, à l'heure actuelle, sauf dans des cas extrêmement particuliers, la théorie de la coalescence est incapable d'inclure de manière générale les forces de sélection et en particulier la transmission du succès reproducteur.

Dans tous les exemples de modèles individu-centrés cités ci-dessus, chaque individu a la même probabilité de se reproduire. Pour modéliser la transmission du succès reproducteur, la propension d'un individu à se reproduire doit dépendre du nombre d'enfants de ses parents, c'est-à-dire de son nombre de frères et sœurs.

Par exemple, Austerlitz et Heyer (1998) considérait une population sexuée à générations non chevauchantes. À chaque génération discrète, le nombre d'enfants d'un couple est tiré dans une loi de Poisson ou une loi géométrique dont la moyenne dépend d'un paramètre qui définit le niveau de transmission du succès reproducteur et du taux de croissance de la population. Il permet ainsi de modéliser une hétérogénéité plus ou moins forte du succès reproducteur (la loi géométrique ayant une variance sensiblement plus élevée que la loi de Poisson), ainsi qu'un niveau plus ou moins fort de transmission du succès reproducteur. L'inconvénient de ce modèle est que la taille de population fluctue et peut être difficilement maîtrisée pour des simulations à long-terme, même sans croissance de population (Austerlitz & Heyer 1998).

Aussi, un second modèle a été développé par Sibert et *al.* (2002), il porte sur une population haploïde à générations discrètes où la probabilité pour chaque individu d'être

choisi comme parent dépend de la taille de sa fratrie. C'est le modèle que nous avons étendu au cas diploïde. Il est détaillé dans la partie I.B.

Prédire les effets du phénomène

La transmission du succès reproducteur modifie autant les paramètres démographiques que la diversité génétique. Au niveau démographique, le phénomène entraîne l'apparition d'une corrélation entre le nombre de descendants d'un individu et de celui de ses parents (c'est-à-dire son nombre de frères et sœurs) alors que cette corrélation intergénérationnelle est inexistante dans une population standard de Wright-Fisher où chaque individu a la même probabilité de se reproduire quelle que soit son histoire parentale. Par ailleurs, la transmission du succès reproducteur entraîne mécaniquement une augmentation de la variance du nombre d'enfants puisque certains individus se reproduisent plus que d'autres (Austerlitz & Heyer 1998; Sibert et al. 2002).

Sur la diversité génétique, l'impact de la transmission du succès reproducteur peut être mis en évidence soit sur les lignées de gènes (arbres de coalescence) soit sur la structure allélique de la population. Au final, les deux sont intrinsèquement liés, puisque la distribution des fréquences alléliques dépend de la forme de l'arbre de coalescence et du processus mutationnel qui place les mutations le long des lignées de cet arbre de coalescence (cf. par exemple Hudson 1990). Au niveau des fréquences alléliques, les patterns de fréquences observés lors de la transmission du succès reproducteur sont caractérisés par un excès d'allèles rares et d'allèles très fréquents, alors que les allèles en fréquence intermédiaire deviennent plus rares (Sibert et al. 2002). Par ailleurs, la diversité génétique est diminuée et le déséquilibre de liaison est augmenté par le phénomène (Austerlitz & Heyer 2000).

En ce qui concerne les arbres de coalescence, qui sont les arbres généalogiques d'un échantillon de gènes pris à un locus donné dans la population (cf. Figure I.A-1), les lois de probabilité régissant la forme et la longueur des branches de ces arbres ont été calculées pour une population neutre de Wright-Fisher (Kingman 1982a; Kingman 1982b). Lorsqu'on compare la forme d'un arbre de Kingman à un arbre de coalescence de gènes échantillonnés dans une population soumise à une transmission du succès reproducteur, on observe trois modifications majeures dans les arbres (Blum et al. 2006; Sibert et al. 2002) : (i) une réduction de la taille de l'arbre entraînant une réduction de la diversité dans les séquences et un effectif efficace réduit à même taille de population (Sibert et al. 2002), (ii) des branches externes plus longues que les branches internes : arbre en forme d'étoile (Figure I.A-2.b),

entraînant un excès d'allèles rares (Sibert et al. 2002), et (iii) un déséquilibre des arbres (cf. Figure I.A-2.d), à savoir que si l'on prend un nœud dans l'arbre, le nombre de descendants se situant sur un de ses deux sous nœuds n'est pas égal au nombre de descendants se situant sur l'autre sous nœud (Blum et al. 2006; Sibert et al. 2002)

Feuilles : Séquences échantillonnées dans la population

Nœuds : Ancêtre commun à deux lignées

Branches externes : Branches reliant un nœud à une feuille

Branches internes : Branches reliant 2 nœuds.

Ancêtre commun le plus récent de l'ensemble des feuilles échantillonnées (en anglais MRCA : Most Recent Common Ancestor).

TMRCA : Temps de l'ancêtre commun le plus récent

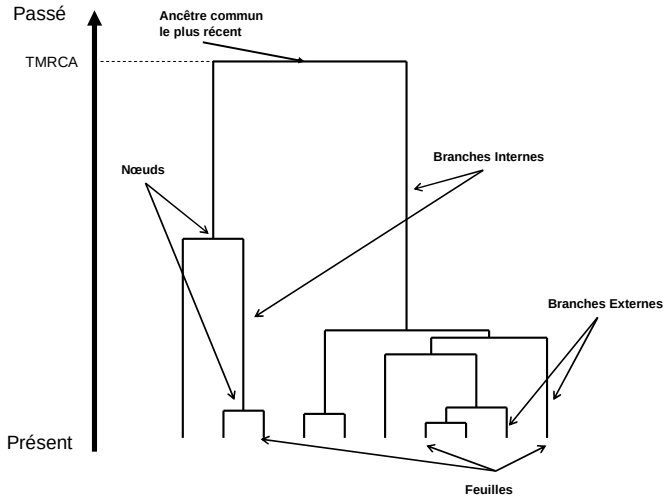


Figure I.A-1 Exemple et descriptif d'un arbre généalogique de copies de gènes (ou coalescent) constitué de 10 feuilles, 9 nœuds

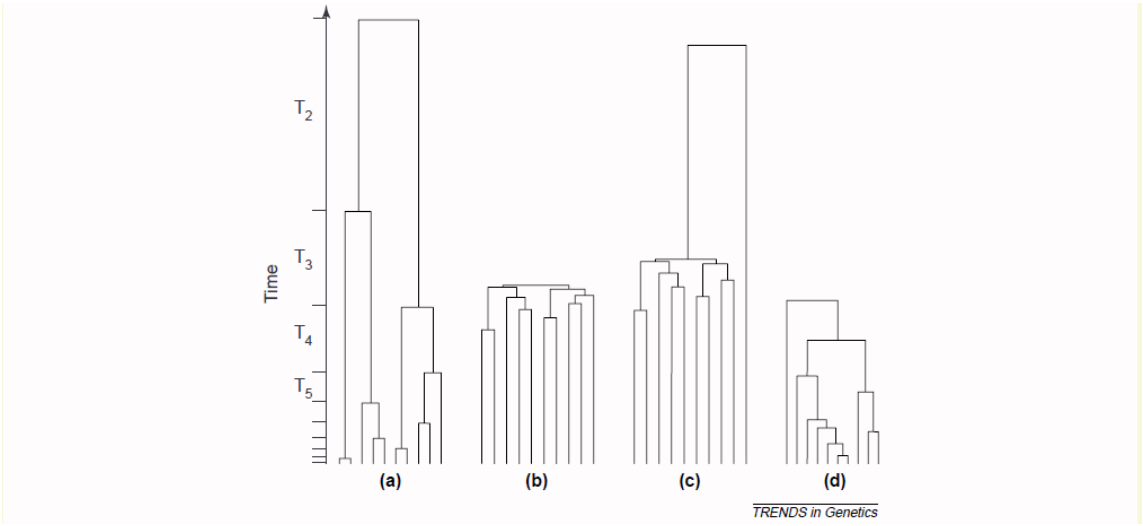


Figure I.A-2 Exemple d'arbres de coalescence. Arbre typique pour un locus neutre dans une population de grande taille Stationnaire (a), un locus neutre dans une population en expansion (b), un locus sous sélection directionnelle avant fixation (c) et un locus neutre dans une population sous transmission du succès reproducteur. Les temps de coalescence « T_i » de l'arbre (a) sont représentés sur l'axe vertical ; par exemple T_4 correspond au temps nécessaire pour passer de 4 à 3 lignées. Extrait de Heyer et *al.* (2005)

D'autres phénomènes peuvent avoir des effets similaires à ceux décrits ci-dessus. Un arbre de coalescence en forme d'étoile (cf. Figure I.A-2.b) est par exemple aussi observé lors d'un événement d'expansion de population (Slatkin & Hudson 1991) ou lors d'un balayage

sélectif (Barton 2000). Une taille réduite de l'arbre de coalescence est attendue à la suite d'un balayage sélectif (Barton 2000). Enfin, le déséquilibre des arbres peut également être observé dans certains scénarios démographiques très spécifiques (Blum et al. 2006) ou pour certaines formes de sélection purificatrice (Maia et al. 2004). Tous ces phénomènes alternatifs ont cependant un impact plus limité sur le déséquilibre des arbres que la transmission du succès reproducteur. Ce déséquilibre apparaît donc comme l'une des caractéristiques les plus spécifiques de la transmission du succès reproducteur (Blum et al. 2006; Sibert et al. 2002).

Détecter le phénomène

La transmission du succès reproducteur peut être détectée sur plusieurs échelles de temps. L'ethnologie ou la sociologie permettent de décrire si certaines pratiques sociales ou culturelles des populations actuelles peuvent entraîner ce phénomène, mais rien ne nous assure qu'il y a une continuité de ces pratiques au cours du temps. La démographie historique permet de reconstruire des généalogies d'individus sur une ou quelques dizaines de générations au sein des populations. Ceci permet de calculer la corrélation du succès reproducteur entre parents et enfants. Ceci a été étudié par exemple dans la population anglaise (Pearson et al. 1899), dans la population islandaise (Helgason et al. 2003), dans la population irlandaise (Madrigal et al. 2003), dans la population du Saguenay Lac-Saint-Jean au Québec (Austerlitz & Heyer 1998), et dans la population huttérite (Pluzhnikov et al. 2007). Les généalogies sont reconstruites à partir des actes de naissances, de décès et de mariage des populations. La reconstruction d'une généalogie dépend donc de la présence de tels actes dans les populations et est limitée par le temps qu'elle recouvre et le risque de biais dû aux informations manquantes.

La troisième source est l'analyse des données de polymorphisme génétique au sein des populations, ces données étant souvent plus faciles à obtenir, en particulier dans les populations sans tradition écrite. En pratique, la transmission du succès reproducteur est évoquée comme une cause probable de distributions atypiques de la diversité génétique quand il existe des indices biologiques ou culturels de corrélation intergénérationnelle du succès reproducteur. Ce phénomène a par exemple été invoqué pour expliquer une faible diversité génétique dans certaines populations, pour des données mitochondriales chez les baleines matrilinéaires (Whitehead 1998) ou sur le chromosome Y dans les populations mongoles (Dulik et al. 2011; Zerjal et al. 2003). La transmission du succès reproducteur a aussi été

invoquée pour expliquer la fréquence élevée de certaines maladies génétiques sévères dans certaines populations (Heyer 1995; Heyer et al. 2005).

Comme nous l'avons vu dans le paragraphe « *Prédire les effets du phénomène »,* ces observations au niveau du polymorphisme génétique ne peuvent cependant pas suffire à elles seules à détecter la transmission du succès reproducteur parce que d'autres forces démographiques ou sélectives peuvent avoir des impacts similaires. Sibert (2002) avait déjà proposé de combiner plusieurs tests de déviation à la neutralité afin de trouver une méthode spécifique pour détecter le phénomène.

De manière alternative, Blum et *al.* (2006) ont montré que l'utilisation du déséquilibre de l'arbre à partir de reconstructions phylogénétiques constitue une voie particulièrement intéressante pour détecter spécifiquement la transmission du succès reproducteur. À partir du déséquilibre des arbres obtenus par reconstruction phylogénétique sur le mitochondrial, Blum et *al.* (2006) ont ainsi mis en évidence une transmission matrilinéaire du succès reproducteur chez les populations de chasseurs-cueilleurs.

Interactions avec d'autres phénomènes sociaux-culturels ou génétiques

D'autres phénomènes sociaux ou génétiques peuvent agir en même temps que la transmission du succès reproducteur et interférer avec ses impacts sur les arbres de coalescence et sur la diversité génétique. Le premier de ces phénomènes est une augmentation de la variance du succès reproducteur entre familles due à des facteurs socio-économiques ou comportementaux qui pourraient jouer sur le niveau de mortalité (Sastry 1997). Dans le cas de la population du Québec ancien, la mortalité des individus d'une famille donnée et donc leur descendance finale dépend par exemple des épidémies, de la survie de la mère et intervalles entre naissances (Pavard et al. 2005). En deuxième lieu, le succès reproducteur peut varier parmi les individus d'une même famille en fonction de différents facteurs (sociaux ou génétiques), notamment le rang de naissance, l'âge de la mère à la naissance et la survie du précédent enfant (Cohen 1975; Nault et al. 1990; Pedersen 2000; Ronsmans 1995). Pour finir, on observe toute une part d'hétérogénéité propre à chaque individu qui n'arrive pas à être reliée à des variables en particulier (biologiques, socio-économiques, culturelles, génétiques, etc.) (Vaupel et al. 1979).

Un autre phénomène important est le mode de choix du conjoint qui a été décrit comme dépendant autant de facteurs biologiques que de facteurs matériels et culturels (Geary et al. 2004). En raison de facteurs sociaux (alliances familiales, endogamie, choix d'un

conjoint proche socio-culturellement), la taille de fratrie d'un individu peut être corrélée à la taille de fratrie de son conjoint. Une corrélation entre les tailles de fratrie des conjoints a ainsi été mise en évidence dans diverses populations : dans le village d'Arthez d'Asson en France (Bocquet-Appel & Jakobi 1993), dans la population anglaise (Murphy 2006) et dans la population Uto au Japon (Imaizumi et al. 1970).

Objectifs de cette partie

Le travail effectué dans les chapitres qui suivent s'ancre dans les travaux présentés ci-dessus, notamment les travaux théoriques de Sibert et *al.* (Sibert 2002; Sibert et al. 2002) et Blum et *al.* (2006). Il est à trois objectifs : (i) modéliser l'impact de la transmission du succès reproducteur sur la diversité génétique et les arbres de coalescence, (ii) utiliser cette approche théorique pour développer une méthodologie pour détecter le phénomène et (iii) appliquer cette méthodologie sur des données réelles et simulées pour en déterminer la puissance et la robustesse.

En ce qui concerne l'aspect de modélisation, les travaux effectués précédemment ont décrit l'impact de la transmission du succès reproducteur sur la diversité génétique et sur la forme des arbres de coalescence, mais uniquement dans un modèle haploïde. De ce fait ils ne prennent pas en compte une part de la complexité des individus et des structures sociales, notamment l'existence de sexes séparés et les mécanismes de formation des couples. La diploïdie et les recombinaisons n'ont pas été non plus prises en compte. Par conséquent, les conclusions de ces modèles ne peuvent s'appliquer que lors d'événements spécifiques comme lors d'une transmission uniparentale du succès reproducteur sur des génomes haploïdes non recombinants (chromosomes Y ou ADN mitochondrial).

Dans une première partie (I.B), nous présentons nos travaux théoriques sur l'impact de la transmission du succès reproducteur sur des variables démographiques (variance entre individus et corrélation intergénérationnelle du succès reproducteur) et sur le déséquilibre des arbres de coalescence dans un modèle diploïde sexué, pour des gènes se situant sur des chromosomes à hérédités différentes, à savoir des chromosomes mitochondriaux, X, Y et autosomaux, dans différents cas de transmission du succès reproducteur (matrilinéaire, patrilinéaire ou biparental). Nous étudions aussi l'interaction de la transmission du succès reproducteur avec les autres paramètres décrits ci-dessus (variance accrue du succès reproducteur, homogamie pour la taille de fratrie). Dans une deuxième partie (I.C.i), nous développons une méthodologie qui vise à reconstruire par des méthodes de phylogénie les

arbres de coalescence d'un échantillon de gènes pris dans la population afin de détecter la transmission du succès reproducteur. Nous appliquons cette méthodologie sur des données génétiques simulées afin de valider la méthode et nous testons son applicabilité sur des données issues du projet HapMap. Ensuite, nous comparerons les règles de transmission du succès reproducteur entre des populations « patrilinéaire » et « cognatique » à partir de données génétiques séquencées sur le chromosome Y des deux groupes de populations (partie I.C.ii). Dans un dernier temps, nous ne nous intéressons plus à des données génétiques mais à des données généalogiques. Nous analysons dans ce cadre l'impact de la transmission du succès reproducteur sur des généalogies diploïdes sexuées simulées que nous comparerons à des données réelles provenant de la région du Cilento en Italie (partie I.D).

I.B Impact of fertility transmission and other socio-demographic factors on reproductive success and coalescent trees

Jean-Tristan Brandenburg Frédéric Austerlitz, Bruno Toupance

Article en revue dans l'état à Genetics Research

I.B.i Abstract

Fertility transmission (FT) is a phenomenon with cultural and/or genetic bases, in which positive relation exists between the number of offspring of an individual and that of his/her parents. Theoretical studies using a haploid individual-based model have shown that FT increases the variance and intergenerational correlation in reproductive success, as well as the imbalance in the coalescent tree of sampled genes. It has been documented in several populations through demographic studies or the reconstruction of the genealogical trees of mitochondrial DNA sequences. However, as mtDNA is a single locus potentially subject to other forces (e.g. natural selection), it is of interest to extend the theory of FT to nuclear loci. We show that because of the mixing process linked with random mating, FT likely has less impact on the variance and intergenerational correlation of reproductive success, and the shape of the coalescent trees, for these loci than for uniparentally inherited loci. Nevertheless, other phenomena such as high heterogeneity in reproductive success and homogamy for family size increase the impact of FT on the shape of the coalescent tree. Thus, FT should be easier to detect when occurring in conjunction with these other factors. We also show the utility of analyzing different kinds of loci (X-Linked, Y-linked, mitochondrial and autosomal) to assess whether CTF is matrilineal, patrilineal or biparental. Finally, from a theoretical perspective, we show that unlike the classical Kingman coalescent process, the shape of the coalescent tree depends upon population size.

I.B.ii Introduction

A positive correlation between the fertility of children and their parents has been observed in many populations of humans and several other species (Heyer et al. 2005). This

phenomenon, denoted fertility transmission (FT) is highly variable among populations, time periods or genders (Murphy 1999). Because fertility is one of the major life-history traits, the relative contribution of genetic, social and environmental factors on fertility has been a subject of study and debate for over a century among geneticists, demographers and economists (see e.g. Kohler et al. 1999). Whatever its determinants and its evolution, FT is also known to influence the evolution of the neutral part of the genome, as it notably reduces effective population size N_e (Nei & Murata 1966). It was also shown to have a strong influence on the occurrence of genetic diseases in some human populations (Austerlitz & Heyer 1998).

FT can arise from genetic or cultural factors. Regarding cultural transmission of fertility (CTF), the Saguenay–Lac Saint Jean (SLSJ) population in Quebec is one of the best documented examples in humans, notably because both population genetic data and accurate deep genealogical data are available. In this population, intergenerational correlations in effective family size (EFS) are much higher than correlations in census family size (CFS), which was interpreted as a sign of cultural rather than genetic basis for FT (Austerlitz & Heyer 1998). Indeed, unlike CFS which incorporates all born children, EFS considers only individuals who reproduced at least once in their native population and is thought to be strongly influenced by cultural factors, such as transmission of wealth or migration behavior. Despite being non-genetically based, Austerlitz et *al.* (1998) showed that CTF had dramatic effects on neutral genetic diversity in SLSJ and largely explained the observed high incidence of several, usually rare, recessive inherited disorders in this population. Conversely, this process also resulted in the complete absence of many inherited disorders that are widespread in other human populations. A similar study linking genetic data and demographic analysis on deep genealogical records in Iceland also showed strong effects of CTF on genetic diversity (Helgason et al. 2003). This CTF was proposed as the major explanation for the substantial fluctuation in the haplotypic frequencies of mitochondrial DNA observed in this population since its foundation.

In other human populations, the involvement of CTF in shaping genetic diversity generally relies on more indirect evidence, such as the simultaneous observation of atypical genetic diversity distributions (i.e. unexpected from classical population genetics theory) and of transmission of cultural traits linked to fertility. For instance, the transmission of female social ranks observed in Maoris was proposed to explain the peculiar frequency distribution of mtDNA haplotypes. Indeed, high-ranking mothers may give birth to high-ranking

daughters, who also have higher survival rates than standard women. This created an intergenerational positive correlation in reproductive success of matrilineal lines (Murray-McIntosh et al. 1998). Similarly, cultural transmission of higher male social status was suggested to explain the increase in gene frequency of some closely related Y-chromosome haplotypes in modern Asia. The establishment of the Mongol empire in Asia would indeed have contributed directly to the spread of a specific lineage by the establishment of a long-lasting male dynasty initiated by Genghis Khan and his relatives, due to social prestige (Dulik et al. 2011; Zerjal et al. 2003). Thus CTF has clearly been observed in several human populations, even though it is not present in all populations (Lansing et al. 2008).

CTF is not restricted to human populations and several examples have been documented in other species. For instance, species of matrilineal whales show lower levels of mitochondrial diversity than species of non-matrilineal whales (Whitehead 1998). In the former, selectively advantageous cultural traits, such as migration strategy, foraging techniques or babysitting, may be matrilineally transmitted, which would yield CTF and thus could explain their low observed genetic diversity in mtDNA sequences. In cheetahs, a 25-year demographic survey in the Serengeti National Park (Tanzania) showed that effective population size is dramatically lower than census population size and resulted in a drastic loss of matrilineal lineages. In this population, lifetime reproductive success was not only highly variable among females, but also preferentially transmitted from mothers to daughters and was assumed to rely on cultural transmission of behavioral traits, such as vigilance against potential predators, a major cause of juvenile mortality in this species (Kelly 2001).

Contrary to the above studies emphasizing the role of social and cultural determinants, other studies have instead advocated a significant genetic basis for FT. For instance, in contrast to SLSJ, the intergenerational correlations in sibship size observed in Huterrites were interpreted as indicating a significant genetic component to reproductive fitness because environmental heterogeneity is reduced in this “egalitarian” population and because none of the social explanations of FT invoked in other societies were found in this population (Pluzhnikov et al. 2007). These findings were further confirmed in a quantitative genetics approach designed to reduce confounding effects due to shared social environments (Kosova et al.). Similarly, a Danish twin pairs study found moderate genetic heritabilities for fertility (Rodgers et al. 2001). Actually, any pattern of selection on a quantitative trait occurring during several generations will lead to FT. This was shown for instance in wheat, where FT was generated through natural selection favoring the tallest plants in an experimental

population (Goldringer et al. 2001), yielding a strong decrease in effective population size. Note that in all cases, this genetic FT is not linked to selection at a single locus but on a quantitative trait coded by many loci located throughout the genome. Finally, FT may be even be linked to both genetic and social factors in some cases (Frère et al. 2010).

Viewed from the neutral part of the genome, FT is basically a process which can be summarized into a simple intergenerational correlation of family size (Heyer et al. 2005). Thus its impacts on neutral genetic diversity can be theoretically studied using a parametric model simulating the transmission of reproductive success to various degrees of strength. Up to now, only haploid models have been considered (Blum et al. 2006; Sibert et al. 2002). They showed that under FT, coalescent trees for neutral genes are expected to differ strongly from standard coalescent trees (Kingman 1982a): i) the time to the most recent common ancestor (TMRCA) is reduced, resulting in lower genetic diversity; ii) the ratio between external branch lengths and internal branch lengths is increased, yielding more starlike trees; iii) coalescent trees tend to be highly imbalanced, i.e. for a given binary node, the distribution of tips is not balanced between the two sub-trees. While a starlike tree can also be observed in cases of population expansion (Slatkin & Hudson 1991), tree imbalance for neutral sequences is much more specific of FT (Blum et al. 2006; Sibert et al. 2002). Thus, estimating the imbalance of coalescence trees reconstructed from samples of neutral DNA sequences constitutes an unambiguous method to infer whether FT occurred or not (Blum et al. 2006). This possibility to detect past periods of FT from genetic data is noteworthy, as genealogical data of sufficient depth are rarely available, whereas genetic data can be obtained more easily by sampling current populations. For instance, phylogenetic trees reconstructed from human mitochondrial DNA samples appear on average more imbalanced in hunter-gatherer populations (HGPs) than in food-producer populations (FPPs), thus indicating more matrilineal FT in the former populations (Blum et al. 2006).

However, previous theoretical studies of the influence of FT on genetic diversity are limited in several aspects. First, they only modeled haploid genomes without recombination and thus can only be applied to uniparentally inherited markers such as Y-linked or mitochondrial genes. A well-known limitation of these markers is that they provide information on a unique gene history, while the ability to detect genome-wide processes should increase with the number of independent markers (Nielsen 2001). Another limitation of non-recombining uniparentally transmitted loci is the difficulty to discriminate FT (either determined by social and/or genetic factors) from selection on a single-locus. Indeed, Maia et

al. (2004) demonstrated theoretically that purifying selection mimics many aspects of FT, including tree imbalance at the locus under selection. Furthermore, many studies have shown that numerous aspects of mitochondrial evolution are indeed far from being neutral (Balloux et al. 2009; Bazin et al. 2006; Elson et al. 2004; Ruiz-Pesini et al. 2004; Stewart et al. 2008). For instance, the large proportion of mutations confined on the external branches of mitochondrial trees may reflect an ongoing purifying selective process that would progressively remove mildly deleterious mutations (Ruiz-Pesini et al. 2004). Interestingly, Kivisild et *al.* (Kivisild et al. 2006) suggested a purifying selection process that would differ between populations eating grain and those eating other food. This may be part of the explanation of the more imbalanced genealogical trees inferred in HGPs (Blum et al. 2006). Note however that Maia *et al.* (2004) showed that only limited imbalance can be obtained through the action of selection, and that the maximum level of tree imbalance was obtained with quite high values of selection coefficients and high mutation rates. Therefore, it appears rather unlikely that natural selection could be the main explanation of the difference in imbalance between genealogical trees of HGPs and FPPs. However, analyzing many independent autosomal neutral markers is obviously a more powerful strategy to infer FT than relying on a unique gene history possibly affected by selection.

Moreover, previous theoretical studies on FT did not distinguish between genders. However, empirical data show that fertility can be transmitted either by the mother alone (matrilineal FT), or by the father alone (patrilineal FT) or by the two parents simultaneously (biparental FT). For instance, matrilineal correlation in CFS was much stronger than patrilineal correlation in the British (Murphy 2006; Pearson et al. 1899) and Icelandic populations (Helgason et al. 2003), while biparental FT was observed in the French Canadian (Austerlitz & Heyer 1998), Hutterite (Pluzhnikov et al. 2007) and Danish populations (Murphy & Knudsen 2002). A joint analysis of biparentally and uniparentally inherited markers (autosomal, X-linked, mitochondrial, Y-linked markers) may help to differentiate sex-specific contributions on FT, in the same manner as gender differences were found for other demographic forces such as migration (Segurel et al. 2008).

Moreover, other demographic phenomena may occur simultaneously with FT. One of these phenomena is an increased heterogeneity in family size among couples as compared to the Wright-Fisher expectation. This was shown for instance in the SLSJ population in Quebec (Austerlitz & Heyer 1998), in which only the conjunction of FT and of this heterogeneity could explain the observed high frequencies of severe genetic disorders. This heterogeneity

may stem from socioeconomic conditions and behavioral characteristics of parents (Ronsmans 1995; Sastry 1997) or from other sources of heterogeneity as for instance the individuals' "frailty" at birth that determines the variance in survival (Vaupel et al. 1979) and thus that in reproductive success.

Finally, when couples are considered, one may question whether spouses tend to marry more often with partners from families with sizes similar to their own family. Numerous studies showed that mate choice is generally not random and may depend on social factors such as material resources, cultural factors (e.g. religion) as well as biological factors (e.g. indicators of good genes in the context of sexual selection) (Geary et al. 2004). As the same social factors may determine not only mate choice but also the number of siblings, homogamy in sibship size, i.e. a positive correlation between the spouses' number of sibs, is an expected feature of many human populations. This phenomenon has been observed in several populations including the Arthez d'Asson village in France (Bocquet-Appel & Jakobi 1993), the English (Murphy 2006) and the Uto in Japan (Imaizumi et al. 1970).

Thus, FT is likely to interact with many others factors in natural populations. It is unclear to which extent these interactions will affect the demographic features of the population (variance among individuals and correlation between parents and offspring in reproductive success). They may also affect the shape of the coalescent tree, and hence the ability to infer FT in natural populations from genetic data. In this paper, we study the joint impact of all these factors on demographic features and on the shape of gene trees inside populations. For this purpose, we simulated a population of diploid individuals with separate sexes, carrying genes belonging to the four genomic compartments: autosomal genes, X-linked genes, Y-linked genes and mitochondrial genes, i.e. genes with respectively biparental, maternally-biased, paternal or maternal inheritance.

In addition to population size, our simulation procedure allowed us to control the levels of FT, the heterogeneity in family size and the homogamy in sibship size. We could thus investigate the joint impact of all these phenomena on the variance in reproductive success among individuals in a given generation and on the correlation of this reproductive success between parents and offspring. Then, we investigated how these phenomena jointly affect the level of imbalance of coalescent trees, as measured by an index that we modified from Fusco and Cronk (1995) to allow for non-binary nodes. This allowed us also to investigate the differences between our diploid model and the haploid model previously

studied by Sibert et *al.* (2002) and Blum et *al.* (2006). We investigated in particular the conditions that resulted in a maximal level of imbalance. Moreover, we also evaluated the impact of FT on the different genomic compartments (autosomal, X-linked, Y-linked and mitochondrial), in particular in the cases where FT is not biparental, but matrilineal or patrilineal. Finally we analyzed how population size may modulate the impact of FT on tree imbalance and reproductive success.

I.B.iii Materials and Methods

(i) Simulations

Populations were simulated using a forward individual-based approach where the pairing rules of individuals and the Mendelian rules of gene transmission from parents to children were explicitly implemented. This allowed us to incorporate a flexible model of fertility transmission into the classical Wright-Fisher framework. All simulations were performed assuming non-overlapping generations and a constant-size population of N individuals. In order to study the properties of the resulting coalescent trees, we stored the genealogical links that were progressively built between genes copies in each simulation. In other words, we did not directly simulate coalescent trees with a backward algorithm, but we considered the gene trees that resulted from our forward-in-time simulation procedure.

First, we performed a set of simulations in a haploid model as in Sibert et *al.* (2002) and Blum et *al.* (2006). In this model, each child had a single parent which was drawn randomly among the individuals of the previous generation according to a probability distribution modeling FT. As in Sibert *et al.* (2002), the probability p_i for a given individual i to be chosen as parent was:

$$p_i = \frac{\gamma_i(a) \times s_i^{\alpha}}{\sum_{j=1}^{M} \gamma_j(a) \times s_j^{\alpha}} \quad \text{(Eq. I.B-a)}$$

where M was the number of individuals ($M=N$), s_i the sibship size of individual i (i.e. the progeny size of its own parent), α the intensity of FT ($\alpha = 0$ meant no FT), and $\gamma_i(a)$ was a random deviate drawn independently for each individual i from a gamma distribution with shape parameter a and mean 1. The a parameter is a means to control the variance in

reproductive success among individuals, as this variance increases with decreases in *a*. As in Sibert et *al.* (2002), we considered only the cases ($a = \infty$) and ($a = 1$), which correspond respectively to a Poisson-like and a geometric-like distribution of offspring number, the latter having a higher variance.

Then we modified this model to consider a population of diploid individuals with separate sexes assuming a constant 1:1 sex-ratio and simulated the transmission of autosomal, mitochondrial, X-linked and Y-linked genes. We assumed that each individual produced offspring with one partner only (strict monogamy). In each generation, coupled were formed either randomly, or under the model of homogamy for sibship size described by McKusick et *al.* (1990). In the latter case, we first ranked the individuals according to their sibship size. An individual coming from a family of rank *k* (i.e. with *k* offspring) randomly chose a mate among the individuals coming from families of rank between (k - s/2) and (k + s/2) (s thus being the "homogamy window size"). Therefore the choice of mates was limited to individuals with similar sibship sizes. When such mates were not available, the actual mate was chosen among families with the nearest rank. As in McKusick *et al.* (1990), the ranking was circularized in order to ensure that everyone could reproduce: individuals coming from the highest-ranking families could choose a mate among individuals coming from the lowest-ranking families and vice versa. This parametric pairing rule allowed us to quantitatively model homogamy for sibship size: the level of correlation between the sibship sizes of the husband and the wife increased as the homogamy window size (s) decreased (Figure I.B-1).

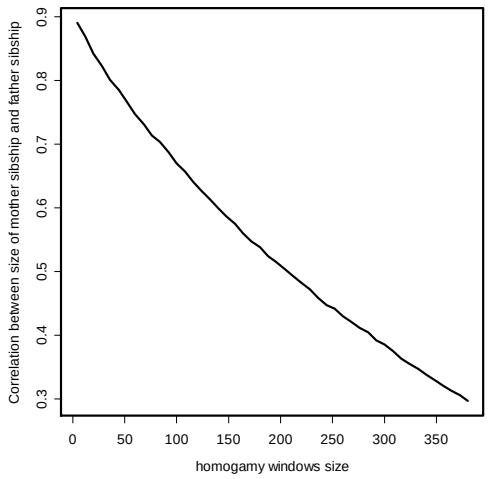


Figure I.B-1. Impact of homogamy window size *s* on the correlation *r* between the mother's sibship and the father's sibship size without fertility transmission and without heterogeneity in reproductive success.

Once the couples were formed, we then drew among them the parents for each individual in the next generation. The probability p_i for a given couple *i* to be chosen as

parents was computed with equation (I.B-a) except that M was the number of couples (M=N/2) and s_i was either the mean sibship size of the two parents, the sibship size of the mother or the sibship size of the father, in order to simulate respectively biparental, matrilineal or patrilineal FT. As in the haploid model, the parameter α controlled the level of FT and the parameter a controlled the level of heterogeneity in family size.

In order to be consistent, we performed the comparisons between haploid and diploid models for the same number of gene copies (N_c). In the haploid model, we have $N_c = N$, since each individual carries one allele for each locus. However, in the diploid model, the number of gene copies for a given locus in a population of N individuals varies according to the compartment: $N_c = 2N$ for autosomal loci whereas, because we assumed a 1:1 sex-ratio, $N_c = N/2$ for Y-linked and mitochondrial loci and $N_c = 3N/2$ for X-linked loci. For each simulation, we recorded the variance in reproductive success, the intergenerational correlation in this reproductive success, and the level of imbalance of the coalescent tree, as detailed below. All investigated parameters are summarized in Table I.B-1.

Table I.B-1. Model parameters

Parameter	*Values*
Heterogeneity (*a*)	1, ∞
Fertility transmission (α)	0-2
Transmission mode	Haploid Diploid (biparental, matrilineal, patrilineal),
Population size (*N*)	500-8000

(ii) Variance and correlation in reproductive success

Comparing the reproductive success in haploid and diploid models is not trivial, as in the first case the number of offspring of an individual is the same as the number of copies in which a gene of this individual will be found in the population in the next generation, while in a diploid model a gene of a given individual is transmitted on average to only half of its offspring. In order to make valid comparisons between haploid and diploid models, we considered in all cases the "gene reproductive success" (GRS), defined for each gene in a given generation as the number of copies of this gene occurring in the next generation. For all simulations, we computed the variance in GRS among genes at the final generation and the correlation between the GRS of a gene in the final generation and the GRS of its parent. In diploid models, we computed these values for each genomic compartment (mtDNA, X-linked, Y-linked and autosomal). As variances and correlations stabilize rapidly, each simulation was run until the 100^{th} generation and we computed the mean values of variance and correlation in GRS over 1,000 replicates for each set of parameters.

(iii)Tree imbalance analysis

FT reduces the height of coalescent trees (Sibert et al. 2002) and thus increases the frequency of polytomies (i.e. when a node has more than two direct sub-nodes). Therefore, in order to measure tree imbalance in this context, we developed a modified version of the imbalance index *I* of Fusco and Cronck (1995) used in its unbiased form (*I'*) in Blum et *al.* (2006), as this index does not allow to analyze non-binary nodes and is biased for nodes with an even number of tips (Purvis et al. 2002). For each node in a tree, this modified index (I_{nb}) was computed as:

$$I_{nb} = \frac{B - m_{k,n}}{M - m_{k,n}}$$, (Eq. I.B-b)

where *k* is the number of direct sub-nodes of the studied node (e.g. 2 for a binary node, 3 for a tertiary node, etc.); *n* is its number of tips; *B* is the maximum number of tips across all its sub-nodes; $M = n - k + 1$ is the maximal possible value for *B;* $m_{k,n}$ was a correction factor allowing the expected value of I_{nb} to be 0.5 in a standard population (without FT). It is computed as $m_{k,n} = 2\,B_{k,n,coal} - M$, where $B_{k,n,coal}$ is the expected value of *B* in a standard population (without FT). $B_{k,n,coal}$ was obtained in practice by performing repeated simulations of a random Kingman coalescent for *n* tips that was stopped when only *k* nodes remained in the genealogy. In each simulation, we recorded the *B* value as the number of tips of the node with the highest number of tips among these *k* nodes. $B_{k,n,coal}$ was obtained by averaging the *B* values over 1,000 independent simulations. To handle polytomies, we considered only nodes with a minimum of $n = k+1$ tips (Figure I.B-2). As for Fusco and Cronk's (1995) *I* index, the expected value of I_{nb} was 0.5 for each node in the coalescent for a standard population, whereas it was greater than 0.5 when the nodes were imbalanced, as for instance under FT. The imbalance measure of the whole tree was simply the mean of I_{nb} across all its nodes on which it could be computed, removing only the most terminal nodes (i.e. the nodes occurring less than four generations before present), as a preliminary study had shown us that using these nodes was biasing downward the I_{nb} value of the tree (result not shown). Each retained node was given an equal weight in the final measure of total tree imbalance, as in Blum et al (2006).

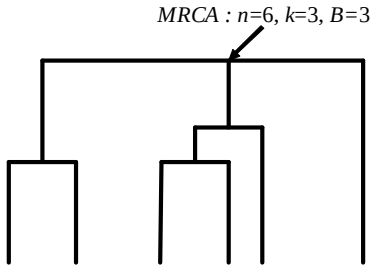


Figure I.B-2: Example of tree where the most recent common ancestor (MRCA) has three direct children (k = 3), six tips (n = 6) and where the maximum number of tips per sub-node is three (B = 3).

I.B.iv Results

(i) Haploid model

Considering first the haploid model without heterogeneity in family size ($a = \infty$), the variance in gene reproductive success (GRS) and the level of correlation in GRS between parents and offspring increased with the level of fertility transmission (α), whatever the assumed population size (N). Starting from a value of 1.0 as expected in the standard Wright-Fisher model (Figure I.B-3a), variance increased first slowly with α for values up to 1.2, but then increased much more steadily to reach values around 3.0 for $\alpha = 2$. The strength of correlation showed the opposite pattern (Figure I.B-3b): starting from the expected null value for α = 0, it increased first rapidly until α = 1.2, but much more slowly afterwards. Starting from the expected value of 0.5 for $\alpha = 0$, the level of imbalance I_{nb} increased with α according to a sigmoid shape with an inflexion point for α = 1.0 (Figure I.B-3c).

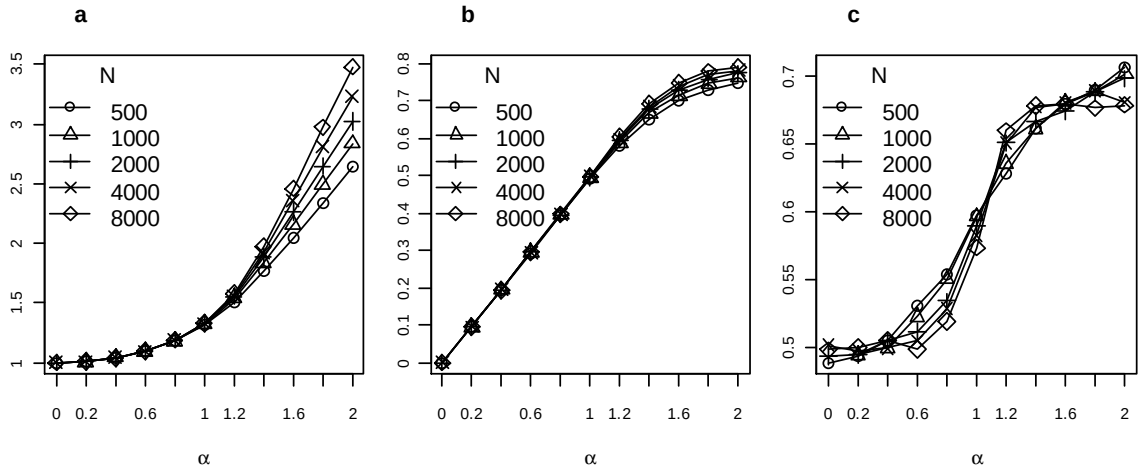


Figure I.B-3: Impact of the intensity of fertility transmission (α) on the variance in gene reproductive success (GRS) (**a**), the intergenerational correlation in GRS (**b**) and the tree imbalance index I_{nb} (**c**) for five population sizes (N = 500, 1000, 2000, 4000, 8000), and a low level of heterogeneity (a=∞), in a haploid model.

Regarding the impact of population size (N), for α below 0.8, the levels of variance and correlation were similar for all N values (Figures I.B-3a, I.B-3b). Nevertheless, for higher values of α, variance increased substantially with N, while the correlation slightly decreased with N. The pattern was more complex for I_{nb}: the higher the population size N, the steeper the slope at the inflexion point (Figure I.B-3c). The initial increase of I_{nb} was thus delayed for high N while the opposite pattern was observed after the inflexion point. The final pattern was also complex as I_{nb} reached a plateau for the highest N while it kept increasing for the lowest N. These parameters were also affected by the level of heterogeneity in family size (Figure I.B-4). Variance in GRS was larger when heterogeneity in family size was high ($a = 1$) than when it was low ($a = \infty$), especially for high α values (Figure I.B-4a): the ratio between the variances obtained in these two cases increased from 2.0 for $\alpha = 0$ to 3.3 for $\alpha = 2$. Conversely, the correlation in GRS between generations was lower for $a = 1$ than for $a = \infty$ (Figure I.B-4b). Finally I_{nb} was higher for $a = 1$ than for $a = \infty$ as a result of a left-shift of the curve in case of high heterogeneity in family size (Figure I.B-4c). The difference was particularly noticeable for small values of α: for instance for $\alpha = 0.8$, I_{nb} was only of ~0.53 when $a = \infty$, while it was of ~0.64 when $a = 1$.

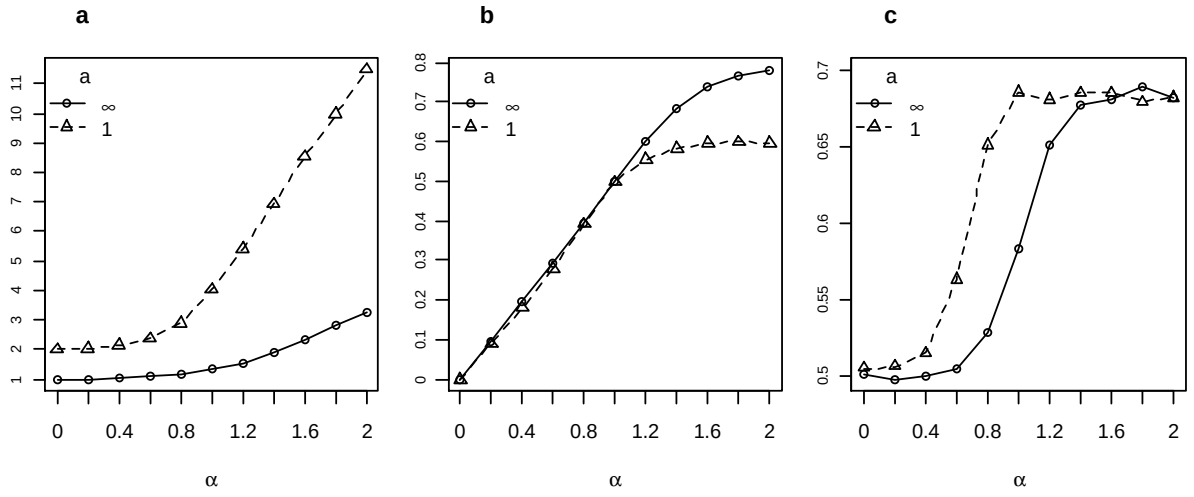


Figure I.B-4: Impact of the intensity of fertility transmission (α) on the variance in gene reproductive success (GRS) (**a**), the intergenerational correlation in GRS (**b**) and the tree imbalance index I_{nb} (**c**), for two levels of heterogeneity in reproductive success ($a = 1$ and $a = \infty$), in a haploid model, assuming a population size $N = 4000$.

(ii) Diploid model

We then investigated whether the levels of variance and correlation in GRS were different between the haploid and the diploid model for autosomal genes (Figure I.B-5). The levels of variance and correlation in GRS were much lower for the diploid model with

biparental transmission than for the haploid one (Figures I.B-5a, I.B-5b). Regarding I_{nb}, while it increased rapidly in the haploid model, until it reached a plateau, it increased slowly but continuously in the diploid model (Figure I.B-5c).

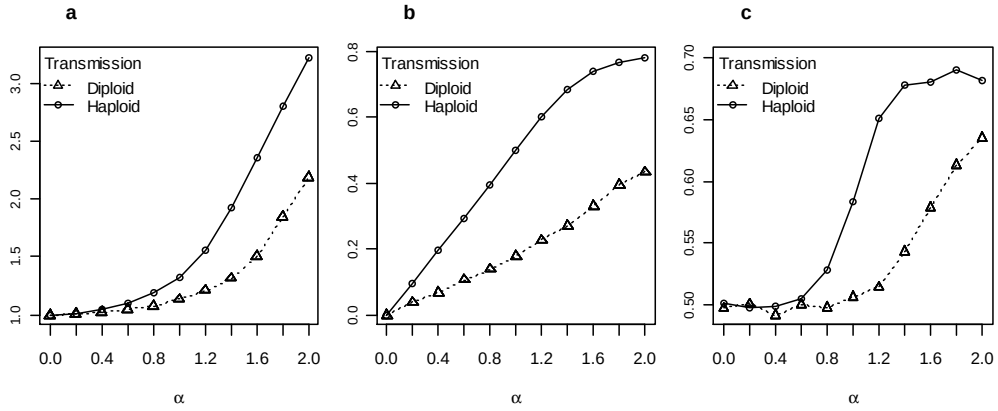


Figure I.B-5: Impact of the intensity of fertility transmission (α) on the variance in gene reproductive success (GRS) (**a**), the intergenerational correlation in GRS (**b**) and the tree imbalance index I_{nb} (**c**) for haploid and diploid models, considering for the latter autosomal loci with biparental transmission and no homogamy in family size (s = 0), assuming a number of gene copies N_c = 4000 (N = 4000 for the haploids and N = 2000 for the diploids).

In this diploid model, we found no differences among the different genetic compartments (autosomes, X-linked, Y-linked and mtDNA) in their level of variance and correlation, as long as FT was biparental (Figures I.B-6a, I.B-6b). Nevertheless, we observed that I_{nb} values were slightly higher for the haploid compartments (mitochondrial and Y-linked) than for the autosomal and X-linked compartments (Figure I.B-6c).

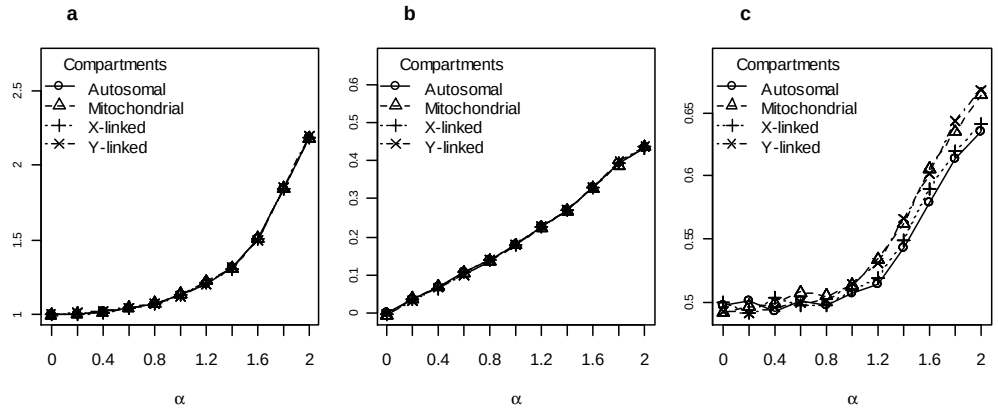


Figure I.B-6 : Impact of the intensity of fertility transmission (α) on the variance in gene reproductive success (GRS) (**a**), the intergenerational correlation in GRS (**b**) and the tree imbalance index I_{nb} (**c**) for different genomic compartments (Autosomal, Mitochondrial, X-linked, Y-linked) in a diploid model without homogamy in family size (s = 0) using a biparental transmission of fertility and without heterogeneity in reproductive success ($a = \infty$), assuming a population size N = 2000.

When considering matrilineal transmission (Figure I.B-7), the level of variance in GRS increased strongly with α in a similar manner for all four compartments, even for the

strictly paternally transmitted Y chromosomes (Figure I.B- 7a). This contrasted with the patterns observed for the correlation and I_{nb} (Figures I.B-7b, I.B-7c), which where very strong for mitochondrial genes, but reached lower values for X-Linked genes, even lower values for autosomal genes, and null for Y-Linked genes.

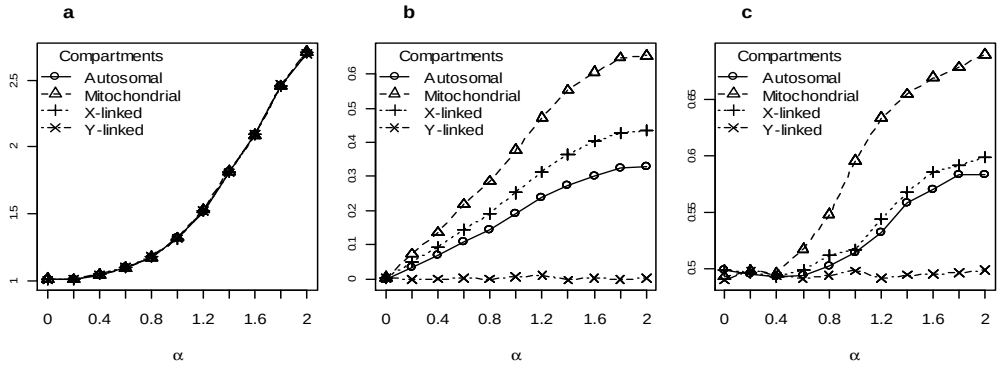


Figure I.B-7 : Impact of the intensity of fertility transmission (α) on the variance in gene reproductive success (GRS) (**a**), the intergenerational correlation in GRS (**b**) and the tree imbalance index I_{nb} (**c**) for different genomic compartments (Autosomal, Mitochondrial, X-linked, Y-linked) in a diploid model without homogamy in family size ($s = 0$) using a matrilineal transmission of fertility and without heterogeneity in reproductive success ($a = \infty$), assuming a population size $N = 2000$.

A symmetric pattern was observed for patrilineal transmission: variance was the same for all compartments (Figure I.B-8a), while correlation and I_{nb} values were the highest for Y-linked genes and then were lower for autosomal genes, even lower for X-linked genes and null for mitochondrial genes (Figures I.B-8b, I.B-8c). Finally, as in the haploid model, variances in GRS and imbalance I_{nb} were higher in the diploid model with high heterogeneity in family size ($a = 1$) than in the standard diploid model ($a = \infty$) while intergenerational correlation in GRS were lower in the former than in the latter (Figures I.B-S1 to I.B-S3). Nevertheless, regarding imbalance, the global patterns of variation of I_{nb} with α for the different genomic compartments were identical for the diploid and the haploid model.

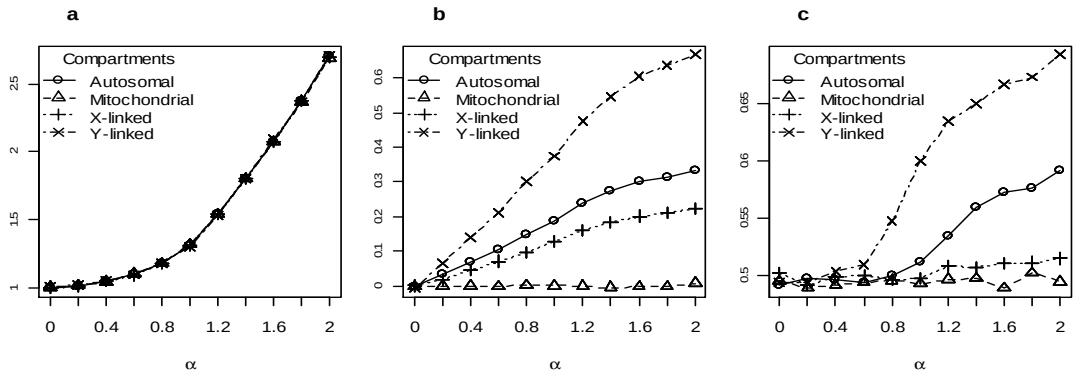


Figure I.B-8 : Impact of the intensity of fertility transmission (α) on the variance in gene reproductive success (GRS) (**a**), the intergenerational correlation in GRS (**b**) and the Imbalance index I_{nb} (**c**) for different genomic compartments (Autosomal, Mitochondrial, X-linked, Y-linked) in a diploid model with homogamy in family (s = 0) size using a patrilineal transmission of fertility and without heterogeneity in reproductive success ($a = \infty$), assuming a population size N = 2000.

(iii) Homogamy in family size

We observed that both the correlation and the variance in GRS increased as the level of homogamy increased (i.e. as *s* decreased, Figures I.B-9a, I.B-9b). Regarding I_{nb}, we observed no difference for low α value ($\alpha < 0.6$) between the situations with or without homogamy (Figure I.B-9c). For intermediate α values ($0.6 < \alpha < 1.4$), the addition of homogamy yielded an increase in I_{nb}, which increased as *s* decreased. However, for large α values ($\alpha > 1.4$), I_{nb} was larger without homogamy and decreased as the level of homogamy increased.

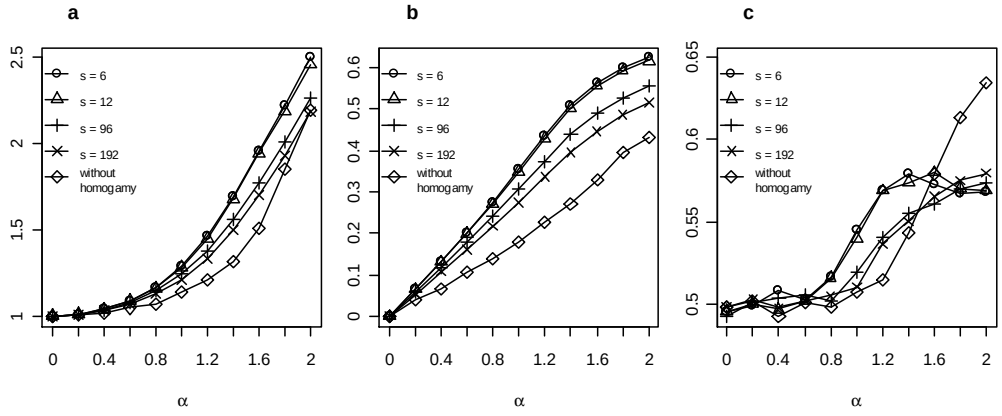


Figure I.B-9 : Impact of the intensity of fertility transmission (α) and homogamy windows size (s) on the variance in gene reproductive success (GRS) (**a**), the intergenerational correlation in GRS (**b**) and the tree imbalance index I_{nb} (**c**) for autosomal genes in a diploid model without homogamy in family size ($s = 0$) using a patrilineal transmission of fertility and without heterogeneity in reproductive success ($a = \infty$), assuming a population size $N = 2000$.

I.B.v Discussion

(i) Interaction between parameters in the haploid model.

Using a individual-based model of FT, our simulations have provided a theoretical framework to detect and understand FT and its impact on the demographic patterns of the population (namely the variance and correlation between parents and offspring in reproductive success) and on the level of imbalance of genealogical trees, as measured by our new index (I_{nb}) that can account for polytomies in the tree. When demographic data are available, FT can be directly measured through the correlation between parents and offspring in reproductive success. In most cases, however, demographic data are not available, so FT cannot be inferred directly. Therefore, Blum et *al.* (2006) developed a method to detect FT

from genealogical tree imbalance. We showed here with our simulation study that the imbalance index (I_{nb}) increased indeed in a population submitted to FT.

However, this index is also affected by other parameters, in particular population size (N) and heterogeneity in family size (a). Regarding the latter parameter, I_{nb} increased when heterogeneity increased (Figure I.B-4). This is a consequence of the increased variance in GRS in that case. By definition, a genealogical tree can be imbalanced only if some of its nodes have more tips than others, and this is only possible if some individuals have a larger reproductive success than others. Thus, an increased level of heterogeneity in progeny size will increase the potentiality of observing tree imbalance, but this imbalance will only occur in case of FT, as the nodes with more direct sub-nodes will always occur in the same part of the tree. In other words, variance is the prerequisite for the creation of imbalance in the genealogical tree, but this imbalance can be maintained in subsequent generations only when GRS is correlated between generations.

It is interesting however to note that increasing the level of heterogeneity resulted in reductions in the level of correlation for a given level of FT (α), while I_{nb} increased at the same time. This decrease in correlation results naturally from the increase of variance, which appears in the denominator of the formula used to compute the correlation coefficient. However, this decrease in correlation does not preclude an increase of I_{nb}. This shows that the relation between I_{nb} and correlation is not completely straightforward: because of the increased variance in reproductive success, a higher value of I_{nb} can be reached when variance is increased, even if the correlation is lower.

The other factor that affected the relation between I_{nb} and FT is population size (N). An increase in population size led to an increase in variance and correlation for a given value of α (Figure I.B-3). Hence, in small populations, variances and covariances between generations were limited by the total size of the population. As a result, I_{nb} was higher for large population sizes than for small ones, when FT was high enough but not too extreme ($1 < \alpha < 1.6$). The opposite pattern was however observed for lower levels of FT ($\alpha < 1$), indicating a rather complex interaction between the parameters. From a theoretical perspective, it is interesting to observe that under FT, the topology of the tree depends therefore on population size. This contrasts strongly with the classical Kingman coalescent process, in which population size does not affect the shape of the coalescent tree, but is only a scaling factor for

its branch lengths (Hudson 1990; Kingman 1982a). Thus, under FT, trees cannot be simply normalized by population size as in the classical Kingman model.

Regarding the possibility to detect FT in natural populations from the inferred coalescent trees, without heterogeneity in family size, the I_{nb} value did almost not deviate from 0.5 for α values below 0.8 (Figures I.B-3c, I.B-4c). Thus, this value of 0.8 appears as the minimal value for which FT could be detected. This would correspond to the correlation of 0.394 (Figure I.B-3b), which is higher than the highest correlation ever observed in human populations, namely the Hutterite population (Pluzhnikov et al. 2007). Here, our conclusions differ from those of Blum et *al.* (2006), who stated that even without heterogeneity in family size, FT could be detected in human populations. However their inferences were based on the simulations of Sibert et *al.* (2002), who considered only populations of 50 individuals. In such small populations, when α is below 1.0, correlations are expected to be lower and I_{nb} values higher than in larger populations. Thus, the conclusion that FT can be detected in populations without heterogeneity in family size does not hold in larger populations ($N \geq 500$). However, when there is heterogeneity in family size, I_{nb} deviates substantially from 0.5 for α values as low as 0.5, corresponding to a correlation value of 0.22 which is realistic for human populations. Thus, it is likely that the populations in which FT was detected through tree imbalance estimations, like hunter-gatherer populations (Blum et al. 2006), were also submitted to other phenomena such as heterogeneity in family size, which is a common phenomenon in human populations (Austerlitz & Heyer 1998; Ronsmans 1995; Sastry 1997; Vaupel et al. 1979). In this context, it is interesting to notice that the high frequency of rare diseases alleles in Saguenay Lac-Saint-Jean was also explained by an interaction between FT and high heterogeneity in family size (Austerlitz & Heyer 1998).

(ii) Diploid model

Furthermore, simulating diploid individuals with autosomal, X-linked, Y-linked and mitochondrial genes allowed us to predict the level of variance and correlation in GRS, as well as gene tree imbalance, for the different kind of genes under biparental, patrilineal or matrilineal FT. We showed that the haploid model is a good approximation of the impact of FT on mitochondrial genes for matrilineal scenarios (Figure I.B-7) and Y-linked genes for patrilineal scenarios (Figure I.B-8), but not for autosomal and X-linked genes for which the pairing process of individuals has to be taken into account. In this context, we demonstrated that the variance, correlation and imbalance of the different compartments depended upon the

features of FT. Firstly, under matrilineal transmission, tree imbalance and correlation in GRS were higher for mitochondrial genes than for autosomal and X-linked genes. They were absent for Y-linked genes as expected. However, it is interesting to note that even when FT is strictly matrilineal, as is probably the case for the Maoris (Murray-McIntosh et al. 1998) and the cetaceous species (Frère et al. 2010; Whitehead 1998), the variance in GRS is the same for all compartments, even for the Y chromosome that is not directly affected by the matrilineal transmission. This remarkable result is a consequence of the fact that, as each male was paired with a unique female in our model (strict monogamy), the variance in reproductive success is necessarily the same for males and females. Thus, effective population size will be decreased for the Y chromosome because of FT occurring only between mothers and daughters. As pointed out by Sibert et al (2002), individuals lose their exchangeability when FT occurs, because unlike in the classical Wright-Fisher model, they do not have the same propensity to reproduce. This will equally apply to paternally inherited genes in a pure matrilineal FT. Phenomena like serial or simultaneous polygamy should to some extent decouple the reproductive variance of males and females. Yet, even in that case, the variance in reproductive success of males should be affected by the variance in the reproductive success of the females they marry.

Conversely, for a patrilineal FT transmission such as in the Mongol population in Eurasia (Zerjal et al. 2003), we expect higher correlation and imbalance for Y-linked genes than for X-linked and autosomal genes, and neither correlation nor imbalance for mitochondrial genes. However the effective population size of these mitochondrial genes should still be decreased as a result of the increased variance in GRS. Biparental FT, as for example described in the French Canadian population (Austerlitz & Heyer 1998) will have a similar impact on all the compartments. Therefore, investigating different genetic compartments could allow researchers to assess whether FT is gender-biased or not in their study population, provided the conditions are fulfilled for the detection of FT. Contrasting the autosomes and X chromosomes should be particularly relevant in this context, as they harbor many independent neutral loci, while the non-recombining Y and mitochondrial chromosomes may be submitted to selective processes that may affect their tree imbalance.

(iii) Future studies

The analysis of autosomal neutral markers could provide a means to distinguish between FT and selection on a single locus, as for instance in the case of HGPs, were FT was

inferred only from the mitochondrial HV1 sequence (Blum et al. 2006). FT should affect all neutral loci in the same way, while single-locus selection should only affect the target locus. Inferring tree imbalance for several neutral nuclear DNA sequences could thus seperate these two phenomena. One practical problem will be that nuclear DNA sequences are prone to intragenic recombination, which leads to some distortions when reconstructing the genealogical tree using phylogenetic methods (Schierup & Hein 2000), hence necessitating a appropriate design for the analysis of nuclear DNA sequences. One solution could be to restrict the analyses to "haploblocks", i.e. portion of sequences that did not undergo any recombination events. Detecting these haploblocks can be performed through different methods, such as the level of linkage disequilibrium (measured for instance by D or D') or the four gamete test (Hudson & Kaplan 1985). If the number of mutations is sufficiently large on a given haploblock to build a phylogenetic tree, we may expect to be able to analyze the imbalance of these trees and thus detect FT. Otherwise more elaborate methods that account for the recombination process, such as Approximate Bayesian Computation (Beaumont et al. 2002) should be developed. Further analyses are needed to assess the feasibility of these methods.

I.B.vi Supplementary Figures

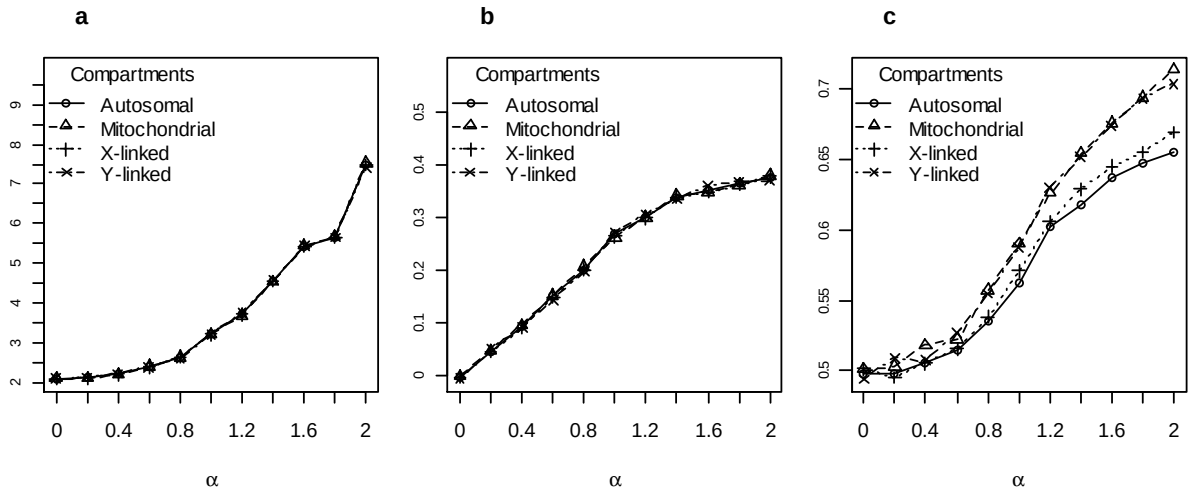


Figure I.B-S1: Impact of the intensity of fertility transmission (α) on the variance in gene reproductive success (GRS) (**a**), the intergenerational correlation in GRS (**b**) and the tree imbalance index I_{nb} (**c**) for different genomic compartments (Autosomal, Mitochondrial, X-linked, Y-linked) in a diploid model without homogamy in family size (s = 0) using a biparental transmission of fertility and with heterogeneity in reproductive success (a = 1), assuming a population size N = 2000.

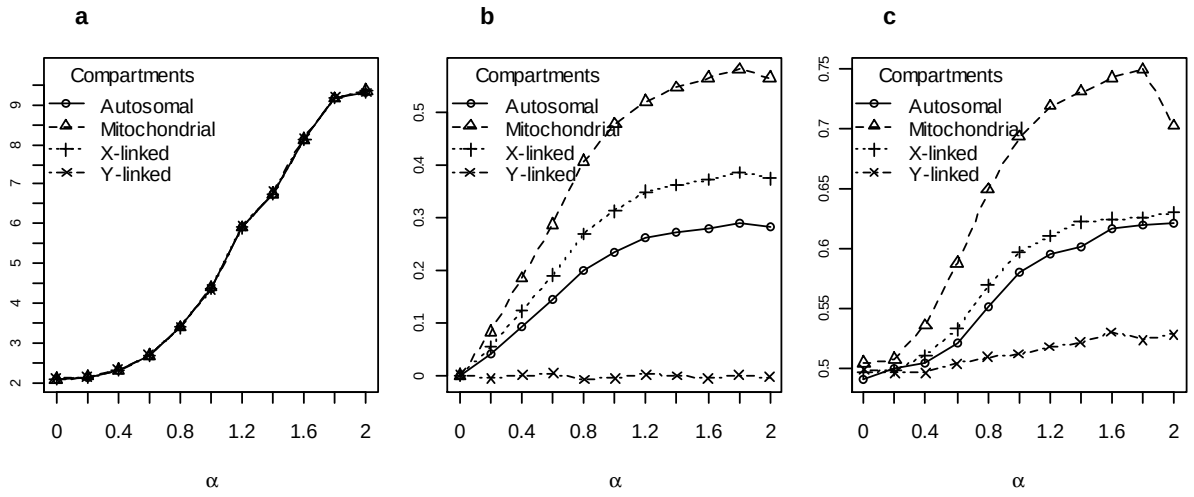


Figure I.B-S2: Impact of the intensity of fertility transmission (α) on the variance in gene reproductive success (GRS) (**a**), the intergenerational correlation in GRS (**b**) and the tree imbalance index I_{nb} (**c**) for different genomic compartments (Autosomal, Mitochondrial, X-linked, Y-linked) in a diploid model without homogamy in family size (s = 0) using a matrilineal transmission of fertility and with heterogeneity in reproductive success (a = 1), assuming a population size N = 2000.

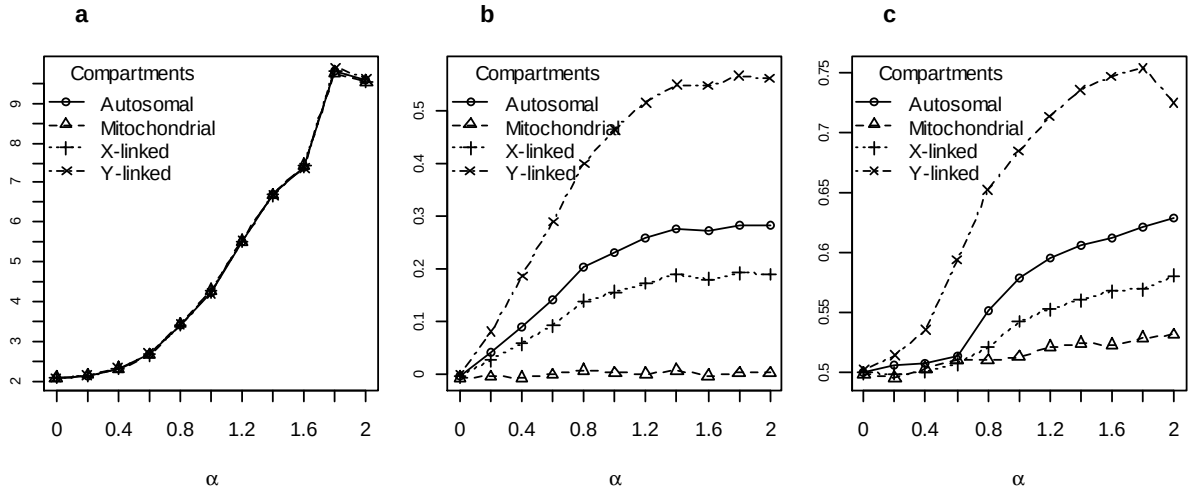


Figure I.B-S3: Impact of the intensity of fertility transmission (α) on the variance in gene reproductive success (GRS) (**a**), the intergenerational correlation in GRS (**b**) and the tree imbalance index I_{nb} (**c**) for different genomic compartments (Autosomal, Mitochondrial, X-linked, Y-linked) in a diploid model without homogamy in family size (s = 0) using a patrilineal transmission of fertility and with heterogeneity in reproductive success (a = 1), assuming a population size N = 2000.

I.C Détection de la transmission du succès reproducteur à partir du polymorphisme génétique

I.C.i Détection à partir de données de chromosomes autosomaux soumises à la recombinaison

I.C.i.a Introduction

La transmission du succès reproducteur est un phénomène qui affecte autant la démographie des populations que l'histoire sous-jacente des gènes et donc, par voie de conséquence, la diversité génétique au sein des populations. Ceci se traduit notamment dans la forme des arbres de coalescence d'un échantillon de gènes pris dans la population. Dans les populations soumises à une forte transmission de la fertilité, Sibert et *al.* (2002) ont en effet montré que ces arbres se distinguent d'un arbre standard de coalescence (Kingman 1982a) par les caractéristiques suivantes : une forme en étoile, une taille plus petite et des arbres sont plus déséquilibrés qu'attendus (cf. Figure I.A-2)

Blum et *al.* (2006) ont utilisé cette dernière propriété des arbres de gènes pour détecter le phénomène de transmission de la fertilité dans les populations humaines. En effet, le déséquilibre des arbres de coalescence est le caractère qui distingue sans doute le mieux la transmission du succès reproducteur et ce, même si les arbres de gènes peuvent aussi être plus déséquilibrés que le coalescent standard dans certains cas de sélection purificatrice (Maia et al. 2004) ou dans le cas d'une structuration spatiale des populations (Blum et al. 2006).

Par simulation, Blum et *al.* (2006) ont montré aussi que les méthodes classiques de phylogénie moléculaire permettaient de reconstruire des arbres dont le déséquilibre était proche de celui des vrais arbres de coalescence et qu'on pouvait donc inférer l'existence de transmission du succès reproducteur dans une population à partir d'un jeu de données de séquences. Néanmoins, ils ont également montré que certaines méthodes de reconstruction surestimaient le déséquilibre réel, en particulier la méthode UPGMA (Sokal & Michener 1958) et dans une moindre mesure la méthode de maximum de vraisemblance.

En utilisant ce procédé, ils ont montré que les arbres phylogénétiques reconstruits à partir d'un échantillon de séquences de la portion hypervariable de la mitochondrie humaine sont généralement plus déséquilibrés dans les populations de

chasseurs-cueilleurs que dans les autres populations. Ce résultat a été interprété comme étant la trace d'une transmission culturelle du succès reproducteur via les femmes. Cependant, l'interprétation univoque d'une transmission culturelle du succès reproducteur implique qu'aucun autre processus ne modifie le déséquilibre de l'arbre comme cela pourrait être le cas sous sélection purificatrice (Maia et al. 2004), processus vraisemblablement non négligeable pour la mitochondrie (Kivisild et al. 2006).

Aujourd'hui, le séquençage de génomes complets, comme le projet « 1000 Genomes » (Via et al. 2010) et la multiplication de données de génotypage pour plusieurs millions de « Single-Nucleotide Polymorphism » (SNPs), c'est-à-dire pour le polymorphisme d'un nucléotide particulier du génome, dans plusieurs populations comme par exemple le projet « HapMap » (2003; 2007), nous donnent accès à une information quasi-complète de la diversité sur les chromosomes autosomaux. Nous avons par ailleurs montré que le déséquilibre des arbres de coalescence pour les différents types de gènes (mitochondriaux, situés sur le chromosome X, sur le chromosome Y ou sur les autosomes) dépend du type de transmission (patrilinéaire, matrilinéaire ou bi-parental), le niveau de déséquilibre attendu sur leurs histoires de gènes étant variable selon le compartiment (cf. partie I.B). Quel que soit le type de transmission, on s'attend néanmoins à ce que les arbres de coalescence des gènes autosomaux soient toujours plus déséquilibrés en présence de transmission du succès reproducteur qu'en son absence. Austerlitz et *al.* (2000) avaient déjà montré que le compartiment autosomal était affecté par la transmission du succès reproducteur puisque les associations entre allèles pour des locus autosomaux étaient augmentées dans ce cas.

Les autosomes et le chromosome X sont soumis au phénomène de recombinaison durant la méiose. Chaque recombinaison entraîne une dissociation partielle de l'histoire des différents locus dans la population. Reconstruire un arbre phylogénétique sans tenir compte de la recombinaison induit toute une série de biais. La topologie inférée peut être une topologie « moyenne », ne correspondant à aucune des topologies réelles des différents locus (Posada & Crandall 2002) et cette topologie inférée est généralement biaisée vers une forme en étoile et sa taille inférée est souvent trop faible (Schierup & Hein 2000). De plus, la reconstruction des séquences ancestrales est mauvaise (Arenas & Posada) et certains tests de neutralité sélective (notamment ceux qui se basent sur les fréquences haplotypiques) peuvent aussi être sérieusement biaisés (Ramírez-Soriano et al. 2008). Enfin, le taux de faux-positifs dans la détection de sélection positive site-spécifique par les méthodes de maximum de vraisemblance peut devenir extrêmement élevé (Anisimova et al. 2003). Pour reconstruire les

histoires de gènes autosomaux, il est donc nécessaire de détecter, au sein des génomes, des portions où aucune recombinaison ne s'est produite au sein de la population. Nous appellerons ces portions « haploblocks » ou « blocks » dans la suite du texte. De nombreuses méthodes ont été décrites pour détecter la position des événements passés de recombinaison et donc pour délimiter des haploblocks comme le test des quatre gamètes (Hudson & Kaplan 1985), le D' (Lewontin 1964), le r^2 (Hill 1976).

Une fois les blocks déterminés, il est nécessaire de savoir si le nombre de SNPs est suffisant sur les blocks pour reconstruire l'histoire des gènes de la population. Or Gabriel et *al.* (2002) ou la description de HapMap (2005) ont montré que sur les zones définies sans recombinaison, le nombre de polymorphismes présents pouvait varier entre 1 et plusieurs centaines, tant au vu de la taille des haploblocks observés dans la population que du nombre moyen de SNPs.

Dans ce chapitre, nous allons développer une méthodologie permettant de détecter la transmission du succès reproducteur sur des chromosomes autosomaux, c'est-à-dire des chromosomes diploïdes soumis à recombinaison. Dans un premier temps, nous allons étendre les travaux effectués sur la détection de la transmission du succès reproducteur par reconstruction phylogénétique (Blum et al. 2006) et analyser les limites de cette détection dans le cas d'un modèle de mutation à nombre infini de sites. Nous étudierons ensuite l'impact des recombinaisons sur la détection du phénomène. Pour finir, nous testerons la méthodologie mise en place sur les données de polymorphisme du projet HapMap (The International HapMap Consortium 2007).

I.C.i.b Matériels et Méthodes

Simulation de séquences et des arbres de coalescence

Simulation du coalescent neutre

Nous simulons les arbres de coalescence constitués de 100 copies de gène dans le cadre d'un modèle neutre sans transmission du succès reproducteur, avec ou sans recombinaison en utilisant le logiciel *ms* (Hudson 2002). Dans le cadre de la recombinaison, nous simulons des arbres avec des taux de recombinaison par paire de bases allant d'une forte valeur (10^{-8}, correspondant à 200 recombinaisons par séquence en moyenne) à une faible recombinaison (5×10^{-10}, correspondant à 0.4 recombinaison par séquence en moyenne), pour

une longueur de séquence de 100 000 paires de bases et une population de 2 000 individus diploïdes.

Simulation de coalescents déséquilibrés par la transmission du succès reproducteur

Nous simulons des arbres de coalescence constitués de 100 échantillons copies de gène dans des populations de 2000 individus diploïdes soumises à de la transmission du succès reproducteur, pour des séquences autosomales sans recombinaison, en utilisant le modèle décrit partie I.B.iii. Ceci est effectué pour différents niveaux de transmission du succès reproducteur (α =0, 0.8, 1.0, 1.2, et 1.4). 500 arbres sont simulés par valeurs de α.

Simulation des séquences

A partir de chaque arbre de coalescence simulé, nous générons un échantillon de séquences, en simulant les mutations le long des branches des arbres. Nous considérons un modèle de mutation à nombre infini de sites où chaque SNP occupe une position différente sans possibilité d'homoplasie. La probabilité d'apparition d'un SNP dans une branche est proportionnelle à la longueur de la branche, c'est-à-dire du temps qu'il y a entre deux ancêtres.

Reconstruction des arbres

Nous comparons trois méthodes phylogénétiques de reconstruction des arbres généalogiques. Les deux premières reposent sur des méthodes d'analyse de groupe (*cluster analysis*), méthodes qui se basent elles-mêmes sur l'analyse de distances : *Neigbour Joining* (NJ, Saitou & Nei 1987) et *Unweighted Pair Group Method with Arithmetic mean* (UPGMA, Sokal & Michener 1958). Dans les deux cas, nous utilisons la matrice du nombre de différences entre toutes les paires de séquences. La fonction *nj* du package *ape* du logiciel R a été utilisée pour la méthode NJ (Paradis et al. 2004) et la fonction *average* du package *phangorn* du logiciel R pour la méthode UPGMA (Schliep 2010).

Nous comparons ces deux méthodes avec la méthode de reconstruction par maximum de vraisemblance implémentée par le logiciel PhyML (Guindon et al. 2010) en supposant un modèle d'évolution moléculaire HKY85 (Hasegawa et al. 1985) et en optimisant la longueur des branches, la topologie des arbres et le taux de substitution.

Détection des blocks

A partir des séquences des individus, nous détectons les recombinaisons à partir d'une méthode des quatre gamètes (Hudson & Kaplan 1985), nous extrayons les blocks et reconstruisons l'arbre (cf. Figure I.C.i-1).

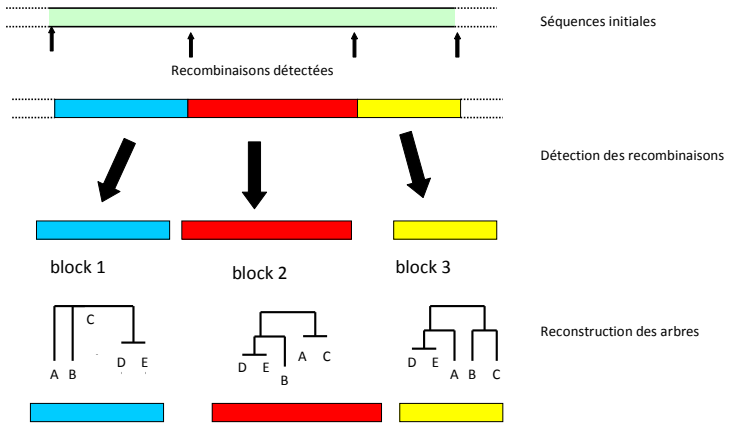


Figure I.C.i-1, Exemple de reconstructions d'arbres sur des zones d'un génome recombinant avec la détection de 4 événements de recombinaison, l'extraction des 3 blocks détectés et la reconstruction d'un arbre sur chacun des blocks (pour un échantillon de 5 séquences).

Calcul de l'indice de déséquilibre des arbres et calcul de p-value

Nous calculons le déséquilibre des arbres avec l'indice I_{nb}, présenté dans la partie I.1, que nous avons développé pour évaluer correctement le déséquilibre des nœuds non binaires. Nous ne considérerons que les arbres qui ont au moins trois nœuds sur lesquels a pu être calculé le déséquilibre, comme l'avaient précédemment fait Blum et *al.* (2006). Pour définir si la valeur du déséquilibre d'un arbre est significativement supérieure à 0.5 (la valeur attendue pour un coalescent standard de Kingman), nous utilisons deux procédures. La première est celle développée par Agapow et Purvis (2002), que nous appellerons dans la suite du texte « méthode par permutation ». Dans cette méthode, pour chaque arbre observé, une distribution nulle est construite à partir de « permutations ». Ces « permutations » consistent à remplacer avec une probabilité de ½ la valeur du déséquilibre I_{nb} de chaque nœud par son symétrique par rapport à 0.5, c'est-à-dire 1-I_{nb}. La valeur globale du déséquilibre est ensuite calculée en faisant la moyenne sur l'ensemble des nœuds de l'arbre des valeurs « randomisées ». La distribution nulle est construite en effectuant 5000 « permutations » indépendantes de ce type et la *p-value* d'un déséquilibre observé est la proportion de permutations conduisant à des arbres « randomisés » de déséquilibre supérieur ou égal à la valeur I_{nb} observée initialement.

La seconde procédure consiste à construire la distribution nulle de I_{nb} à partir de la distribution du déséquilibre d'arbres reconstruits en utilisant la même méthode de reconstruction que celle utilisée pour le jeu de données observé (c'est-à-dire UPGMA, NJ ou PhyML) mais en partant d'un jeu de données simulées sous un coalescent standard de Kingman, en utilisant le même nombre de séquences et en faisant en sorte que le niveau de polymorphisme soit identique à celui des données observées. Par exemple, pour déterminer la *p-value* du déséquilibre d'un arbre reconstruit avec la méthode de NJ pour un jeu de 100 séquences présentant 50 SNPs, nous simulons des arbres de coalescence de Kingman pour 100 copies de gènes, sur lesquels nous simulons exactement 50 SNPs. La suite est strictement identique à ce qui est effectué pour le jeu de données observé : calcul de la matrice de distances, reconstruction de l'arbre par la méthode NJ et calcul du déséquilibre de l'arbre reconstruit. La procédure est répétée 500 fois de manière indépendante pour construire la distribution nulle et nous définissons la *p-value* comme la proportion de ces répétitions pour lesquelles le déséquilibre est supérieur ou égal au niveau de déséquilibre I_{nb} de l'arbre observé. Dans la suite du texte cette procédure sera appelée la méthode « Kingman reconstruit » (cf. Figure I.C.i-2).

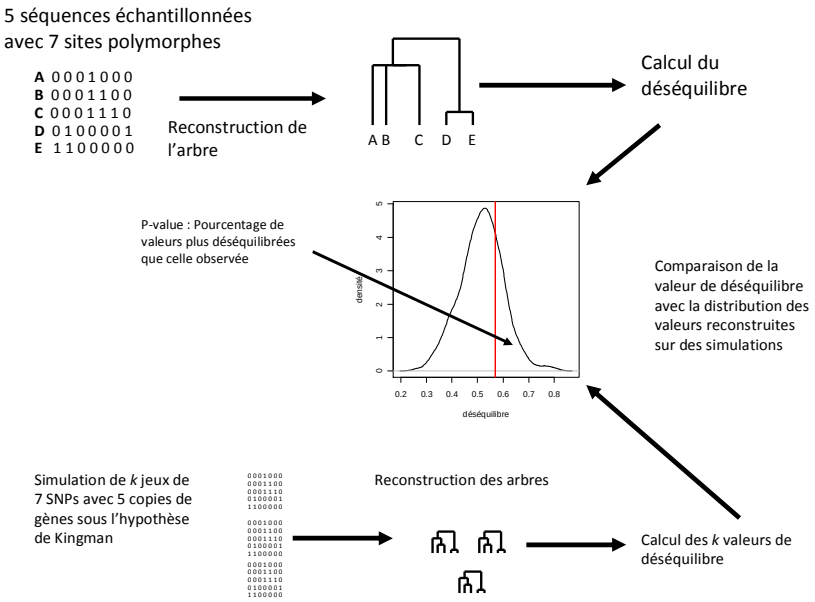


Figure I.C.i-2, Exemple de calcul d'une *p-value* selon la méthode « Kingman reconstruit », pour un échantillon de 5 copies de gène avec 7 SNPs

Pour les deux procédures, nous considérerons qu'un arbre est déséquilibré lorsque la *p-value* est inférieure au seuil de 0.1, c'est-à-dire que nous travaillerons avec un seuil de significativité de 10%. Nous définissons la puissance comme le pourcentage d'arbres qui sont plus déséquilibrés que l'attendu ; donc pour un seuil de 0.1, il s'agit du pourcentage d'arbres qui ont une *p-value* inférieure à 0.1. Pour chaque jeu de paramètres (nombre de SNPs, taux de recombinaison, valeur de α), nous répétons 500 fois nos simulations.

Analyse des données de Simulation

(1) Influence de la méthode et du nombre de SNPs

Nous testons le déséquilibre sur trois méthodes de reconstructions : NJ, PhyML et UPGMA. Pour chaque méthode de reconstruction, nous testons sa sensibilité au nombre de SNPs, en simulant sur les arbres un nombre différent de SNPs : 20, 30, 40, 50, 100, 200, 500. Pour une méthode de reconstruction *i* et un nombre de SNPs *n*, nous simulons ces *n* SNPs sur 500 arbres de coalescence neutre. Pour chaque jeu de n SNPs, nous reconstruisons l'arbre à partir de la méthode *i* de phylogénie.

Pour le jeu de 500 arbres de 100 copies de gène reconstruits avec la méthode *i* et *n* SNPs, nous calculons le pourcentage d'arbres ayant au moins 3 nœuds résolus et donc conservés pour le calcul moyen du déséquilibre et le pourcentage d'arbres plus déséquilibrés qu'attendu (c'est-à-dire dont la *p-value* est inférieure à 0.1).

(2) Influence de la transmission du succès reproducteur

Nous testons l'impact du déséquilibre des arbres reconstruits en utilisant deux méthodes (*i*) de reconstruction, NJ et PhyML ; 4 valeurs (*n*) de SNPs, 50, 100, 200 et 500 ; et 4 intensités (α) de transmission du succès reproducteur : 0, 0.8, 1.0, 1.2, et 1.4. Pour chaque méthode de reconstruction (*i*), chaque nombre de SNPs (*n*) et chaque intensité de transmission du succès reproducteur (α), nous simulons 500 arbres avec le modèle décrit en partie I.B.iii (taille de population de 2000 individus et 100 copies de gène). Nous simulons *n* SNPs sur notre jeu d'arbres et reconstruisons avec la méthode i les arbres (arbres dits « reconstruits »)

Pour chaque arbre de coalescence initial, nous calculons son déséquilibre moyen et sa *p-value* avec la méthode par permutation. Parmi les 500 arbres de coalescences initiaux, nous extrayons ceux dont le nombre de nœuds est suffisant (plus de trois nœuds où le déséquilibre peut être calculé) et calculons sur ceux-ci la moyenne des déséquilibres et la puissance du test de déséquilibre des arbres (pourcentage d'arbres significativement déséquilibrés avec la méthode par permutation pour une *p-value* < 0.1).

De même, sur le jeu d'arbres reconstruits, nous calculons chaque moyenne de déséquilibre et chaque *p-value*. Nous calculons la moyenne des déséquilibres sur les 500 arbres et la puissance, c'est-à-dire le pourcentage d'arbres plus déséquilibrés qu'attendu (cette puissance est calculée pour une *p-value* < 0.1) avec la méthode par permutation. Puis, nous calculons l'autre *p-value* dite de « Kingman reconstruit » (cf. « Calcul de l'indice de déséquilibre des arbres et calcul de *p-value* » et Figure I.C.i-2). Sur les arbres reconstruits où nous avons calculé un déséquilibre moyen, une *p-value* par permutation et une *p-value* de « Kingman reconstruit » et dont le nombre de nœuds est suffisant, nous calculons la moyenne des déséquilibres et le pourcentage d'arbres plus déséquilibrés qu'attendu avec la méthode par permutation et la méthode de « Kingman reconstruit » (ces puissances sont calculées pour une *p-value* < 0.1).

Pour nos quatre valeurs de SNPs, cinq valeurs d'α et nos deux méthodes de reconstruction, nous comparons, dans une première analyse, la moyenne de déséquilibre entre les arbres initiaux et les arbres reconstruits. Puis, dans une seconde analyse, nous comparons la puissance avec une méthode par permutation des arbres initiaux avec la puissance par permutation des arbres reconstruits. Dans une troisième analyse, nous comparons la puissance avec une méthode par permutation des arbres initiaux avec la puissance de « Kingman reconstruit » des arbres reconstruits.

Pour finir, nous analysons les erreurs de conclusion des tests. Pour cela, nous comparons la *p-value* de l'arbre initial et la *p-value* de l'arbre reconstruit. Nous considérons, premièrement, le pourcentage d'arbres reconstruits qui apparaissent comme significativement déséquilibrés (*p-value*<0.1) alors que leurs arbres initiaux apparaissent comme équilibrés (*p-value*>0.1), que nous appellerons erreur de type I. Puis, nous considérerons le pourcentage d'arbres reconstruits équilibrés et dont leurs arbres initiaux apparaissent comme significativement déséquilibrés, appelé erreur de type II.

(3) Influence de la recombinaison

Nous simulons 500 jeux d'arbres pour divers taux de recombinaisons *r* (10^{-10}, 5×10^{-10}, 10^{-9}, 5×10^{-9}, 10^{-8}, 5×10^{-8}) en considérant une séquence de longueur 100 000 paires de bases et 5000 individus.

Analyse avec détection des blocks

Pour chaque simulation, nous extrayons les arbres correspondants à chacun des blocks non-recombinants et simulons un total de 1000 SNPs. Sur ces 1000 SNPs, nous détectons les recombinaisons avec un test des « quatre gamètes ». Nous sélectionnons les blocks sans recombinaison et reconstruisons sur chaque bloc l'arbre avec une méthode de NJ et calculons le déséquilibre de nos arbres. Pour un taux de recombinaison donné, sur tous les arbres reconstruits (indépendamment du jeu d'arbres), nous faisons une moyenne des arbres en fonction du nombre de SNPs avec lequel ils ont été reconstruits. Nous regardons la moyenne du déséquilibre en fonction du nombre de SNPs utilisés pour le reconstruire et du taux de recombinaison.

Analyse sans détection des blocks

Sur chaque jeu d'arbres, nous simulons *n* SNPs (50, 100, 200 et 500). Nous reconstruisons un arbre sur la séquence sans détecter les recombinaisons avec une méthode de NJ sans détecter les blocks. Nous comparons le déséquilibre en fonction du taux de recombinaison et du nombre de SNPs.

Analyse des Données HapMap

Nous utilisons les données phasées (res. #22) de la version 2 du projet HapMap (2007) qui totalisent 3.1 millions SNPs génotypés dans trois populations : une population d'origine européenne « CEU » (Utah, Etats-Unis - 60 individus), une population africaine « YRI »

(Yoruba, Nigéria - 60 individus) et une population asiatique « JPT+CHB » (mélange de Japonais de Tokyo, Japon et de Hans de Beijing, Chine - 90 individus).

Dans un premier temps, nous effectuons un premier filtre des données afin d'exclure les chromosomes présentant une trop forte ressemblance au sein de chaque population. Nous éliminons les chromosomes qui sont très proches génétiquement car ils pourraient entraîner un biais dans le déséquilibre de l'arbre : des chromosomes génétiquement trop proches signifie une « parenté » trop importante ce qui nous éloignerait d'un échantillonnage aléatoire de gènes, et donc de la théorie de Kingman sur laquelle repose le calcul du déséquilibre. Comme lorsque nous reconstruisons les histoires de gènes, nous ne nous intéressons pas aux individus mais à leurs chromosomes (c'est-à-dire que nous considérons de manière indépendante deux chromosomes provenant ou non du même individu), nous calculons la valeur de parenté sur les chromosomes. Pour chacun des 22 chromosomes autosomaux humains, nous considérons toutes les paires possibles de chromosomes homologues présents dans la population afin d'en estimer la parenté génétique R suivant la formule (Rousset 2002) :

$$R = \frac{Q_c - Q_m}{1 - Q_m} \quad \text{(Eq. I.C.i-a)}$$

où Q_c est la proportion de variants identiques entre les deux chromosomes considérés et Q_m la proportion moyenne de variants identiques entre deux chromosomes de la population. Nous ne conservons aléatoirement qu'un chromosome par paire de chromosomes dont le coefficient de parenté est supérieur à 1/32, correspondant à des cousins issus de germains si nous ramenons la mesure à des individus, afin d'obtenir un jeu de chromosomes non-apparentés de manière notable.

Parmi ces chromosomes, nous définissons un haploblock comme une portion de chromosome où aucune recombinaison n'est détectée dans la population. Pour détecter si les sites sont en déséquilibre de liaison, nous utilisons un test des « quatre gamètes » (Hudson & Kaplan 1985) car la taille des blocks est en moyenne plus faible que celle obtenue avec un test du D' (Lewontin 1964). L'analyse des tailles de blocks de HapMap a en effet montré des tailles plus petites avec une détection par un test des « quatre gamètes » qu'avec l'utilisation d'un test du D' (The International HapMap Consortium 2005). Nous avons donc choisi d'être assez stringents à cette étape en utilisant le test des « quatre gamètes », afin de limiter le plus possible l'occurrence de recombinaisons non détectées au sein des séquences.

Nous comptons le nombre de SNPs présents sur chaque bloc afin de déterminer la distribution du nombre de ces SNPs par bloc. Sur chaque bloc possédant plus de 40 SNPs[1], nous reconstruisons les arbres par une méthode de maximum de vraisemblance (PhyML) et une méthode de distance (NJ). Les arbres sont reconstruits en utilisant les allèles ancestraux déterminés par la méthode décrite par Voight et *al.* (2006) et enracinés à partir de ces allèles. Le déséquilibre de ces arbres est estimé à l'aide de l'indice de déséquilibre I_{nb}, et la signification statistique de cet indice est évaluée soit selon la méthode par « permutation » soit selon notre nouvelle méthode « Kingman reconstruit ».

Finalement, nous cherchons à savoir si les blocks détectés sont représentatifs du reste du génome en comparant la diversité des blocks détectés à celle du génome complet. Enfin, nous cherchons à savoir si les arbres reconstruits à partir des zones géniques présentent des niveaux de déséquilibre différents de ceux des arbres reconstruits à partir des zones non géniques. Pour cela, nous utilisons les annotations et les fonctions des gènes obtenues dans la banque de données « UCSC data base » (Fujita et al. 2011).

I.C.i.c Résultats

Simulation

Influence des méthodes de reconstruction sur l'évaluation du déséquilibre dans un modèle neutre sans recombinaison

Nous comparons d'abord le déséquilibre des arbres reconstruits selon les trois méthodes de reconstruction (UPGMA, NJ, et PhyML) sur un jeu d'arbres simulés avec le logiciel *ms* dans un modèle sans transmission du succès reproducteur et sans recombinaison. Dans ce cas, par définition, les arbres de coalescence simulés présentent en moyenne un indice de déséquilibre de 0.5 et la question est ici d'évaluer dans quelle mesure les méthodes de reconstruction parviennent à préserver cette caractéristique de la topologie des arbres de coalescence standard de Kingman. Nous montrons que les trois méthodes de reconstruction possèdent un biais positif, avec une valeur moyenne du déséquilibre I_{nb} supérieur au 0.5 attendu (Tableau I.C.i-1). Le biais est faible pour les méthodes NJ et PhyML, il varie entre 0.03 à 0.07, la méthode NJ étant moins biaisée en moyenne que la méthode PhyML. En revanche, le biais est beaucoup plus élevé pour la méthode UPGMA, avec une valeur qui va

[1] Cette valeur est un consensus entre le temps de calcul nécessaire pour reconstruire les arbres et calculer leur *p-value* et la maximisation du nombre de blocks pour faire l'analyse.

de 0.075 à 0.2. Par ailleurs, ce biais des différentes méthodes dépend du nombre de sites polymorphes utilisés pour reconstruire les arbres : plus la diversité haplotypique est importante, moins le biais est important, toutes méthodes confondues. En d'autres termes, la reconstruction biaise d'autant plus la topologie des arbres vers des topologies déséquilibrées que la quantité d'information disponible est faible.

Comme pour le déséquilibre, la puissance (*p* : proportion de tests significatifs au seuil de 10%), calculée à partir de la méthode par « permutation », est également plus élevée que la valeur attendue (10%) puisqu'elle est de l'ordre de 15% à 20% pour PhyML et NJ et de 75%, pour la méthode de l'UPGMA. Néanmoins, contrairement au comportement de l'indice de déséquilibre, la puissance reste stable quel que soit le nombre de sites polymorphes (n) présents sur la séquence (Tableau I.C.i-1).

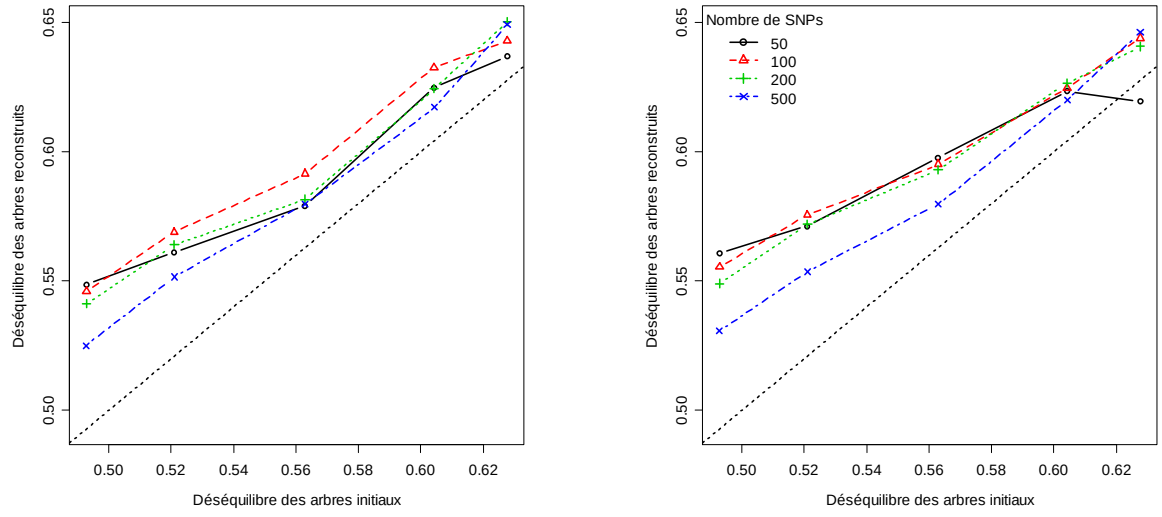
Le nombre d'arbres reconstruits qui ont plus de trois nœuds sur lesquels le déséquilibre peut être calculé dépend autant de la méthode de reconstruction que du nombre de sites polymorphes. La méthode UPGMA permet de calculer le déséquilibre sur la totalité des arbres (n = 500) quelque soit le nombre de SNPs. Au contraire, pour les méthodes de reconstruction NJ et PhyML, lorsque le nombre de SNPs est inférieur à 100, le déséquilibre ne peut pas être calculé sur tous les arbres. Par exemple, il n'a pu être calculé que sur 17% des arbres en NJ et sur 14% en PhyML pour 20 SNPs. Néanmoins, dès 40 SNPs, ces pourcentages grimpent à plus de 80% en NJ et à plus de 70% en PhyML.

Tableau I.C.i-1, Valeurs moyennes de déséquilibre I_{nb}, puissance du test de déséquilibre évaluée selon la méthode par « permutation » (p), nombre d'arbres sur lesquels on a pu calculer le déséquilibre (nb), en fonction de la méthode de reconstruction (UPGMA, NJ ou PhyML) et du nombre de SNPs (n) sur des arbres de coalescence de Kingman en absence de recombinaison.

	20	*30*	*40*	*50*	*100*	*200*	*500*
UPGMA	0.6984 (p=0.714, nb=500)	0.6854 (p=0.76, nb=500)	0.6731 (p=0.754, nb =500)	0.6699 (p=0.77, nb=500)	0.6373 (p=0.73, nb=500)	0.6104 (p=0.638, nb=500)	0.575 (p=0.464, nb=500)
NJ	0.5676 (p=0.1744, nb =86)	0.5432 (p=0.1513, nb =238)	0.5472 (p=0.194, nb=402)	0.5487 (p=0.1737, nb=449)	0.546 (p=0.1804, nb =499)	0.5412 (p=0.21, nb=500)	0.525 (p=0.182, nb=500)
PhyML	0.5672 (p=0.1528, nb=72)	0.5636 (p=0.195, nb=241)	0.5582 (p=0.1803, nb=355)	0.5607 (p=0.2078, nb=438)	0.5554 (p=0.1824, nb=499)	0.549 (p=0.234, nb=500)	0.5306 (p=0.194, nb=500)

Impact de l'hérédité du succès reproducteur sur le déséquilibre des arbres reconstruits

Dans cette partie, nous utilisons des arbres de gènes extraits de populations simulées par notre méthode « forward » pour différents niveaux α de transmission du succès reproducteur (cf partie I.B). Pour des valeurs non nulles de α (c'est-à-dire quand il y a effectivement de la transmission du succès reproducteur), ces arbres sont déséquilibrés (leur indice de déséquilibre I_{nb} est supérieur à 0.5) et le test de déséquilibre permet de le détecter. Ces arbres généalogiques initiaux ont été utilisés pour simuler un polymorphisme de type SNP en suivant exactement la même méthode que celle employée ci-dessus pour les arbres de coalescence standard de Kingman. A partir des données haplotypiques obtenues, nous avons reconstruits des arbres par les méthodes de reconstruction NJ et PhyML. Nous avons enfin comparé le déséquilibre des arbres généalogiques initiaux à celui des arbres reconstruits. Il s'agit donc ici d'évaluer dans quelle mesure la méthode de reconstruction permet ou non de préserver un déséquilibre dans des coalescents obtenus sous transmission du succès reproducteur.



a *b*

Figure I.C.i.3, Déséquilibre des arbres reconstruits en fonction du déséquilibre des coalescents initiaux, pour quatre niveaux de polymorphisme (nombre de SNPs), et pour deux méthodes de reconstruction : (*a*) reconstruction par *Neighbour Joining* (NJ) et (*b*) reconstruction par maximum de vraisemblance (PhyML). La droite représentant y = x est en pointillées noires.

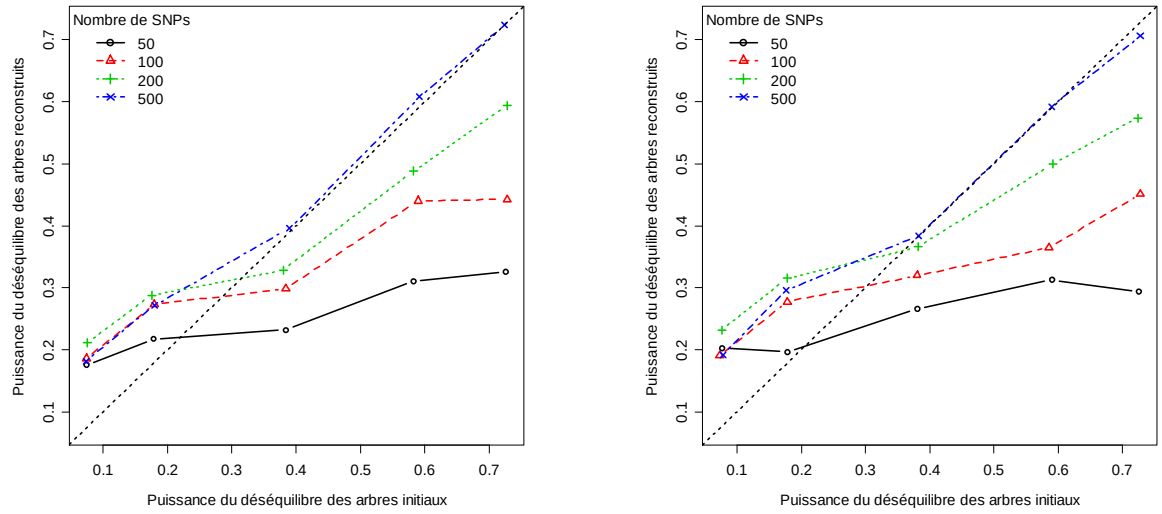
Le déséquilibre des arbres reconstruits augmente avec celui des arbres initiaux (Figure I.C.i.3). Néanmoins, quelle que soit la valeur initiale du déséquilibre, celle des arbres reconstruits est toujours plus élevée (biais variant entre 0.02 et 0.06). Cette différence diminue avec l'augmentation du déséquilibre initial. Les valeurs moyennes du déséquilibre des arbres

reconstruits par NJ (Figure I.C.i.3a) sont peu différentes de celles des arbres reconstruits par PhyML (Figure I.C.i.3b).

Pour chaque valeur α, correspondant à une valeur moyenne de déséquilibre des arbres initiaux, nous calculons la corrélation entre les valeurs de déséquilibre de chaque paire d'arbres (arbre initial, arbre reconstruit). La liaison entre ces deux valeurs ne semble pas dépendre de la valeur du déséquilibre initial mais plutôt du niveau de polymorphisme (nombre de SNPs) des blocks à partir desquels les arbres sont reconstruits (Figure I.C.i-S1). On observe en effet une corrélation entre 0.2 et 0.3 quel que soit le déséquilibre initial pour 50 SNPs alors que cette corrélation est aux alentours de 0.6 lorsque 500 SNPs sont simulés.

Puissance de détection du déséquilibre selon la méthode par « permutation »

Si on utilise la méthode par « permutation » décrite par Agapow et Purvis (2002) et utilisée par Blum et *al.* (2006) pour construire un test de détection du déséquilibre, nous montrons que les puissances de détection obtenues sur les arbres reconstruits ne dépendent pratiquement pas de la méthode de reconstruction employée (NJ ou PhyML) (Figure I.C.i.4). Dans les deux cas, les puissances de détection du déséquilibre après reconstruction augmentent avec les puissances calculées sur les arbres généalogiques véritables. Néanmoins, la puissance calculée après reconstruction est soit inférieure soit supérieure à la puissance initiale, l'ampleur et le sens de la différence entre les deux mesures dépendant à la fois de la puissance initiale mais également du niveau de polymorphisme des séquences (Figure I.C.i.4).



a *b*

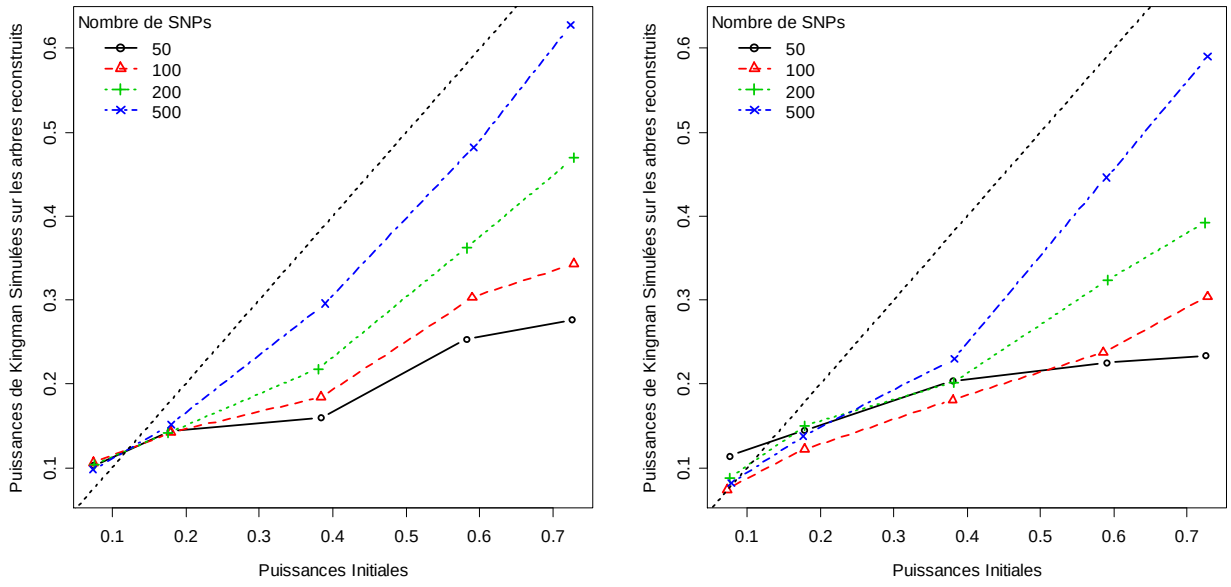
Figure I.C.i.4 Puissance de détection du déséquilibre des arbres reconstruits selon la méthode par « permutation » en fonction de la puissance de détection du déséquilibre des coalescents initiaux, pour quatre niveaux de polymorphisme (nombre de SNPs), et pour deux méthodes de reconstruction : (*a*) reconstruction par *Neighbour Joining* (NJ) et (*b*) reconstruction par maximum de vraisemblance (PhyML). Chaque point correspond à un couple de moyennes de puissance calculées sur 500 simulations indépendantes obtenues pour un même niveau α de transmission du succès reproducteur et en utilisant un seuil de significativité de 10%. La droite représentant y = x est en pointillées noires.

Lorsque le nombre de SNPs est faible (par exemple avec 50 SNPs), même si la puissance initiale de détection du déséquilibre est forte, la puissance après reconstruction n'atteint pas 0.3. Plus le niveau de polymorphisme des blocks est important, plus la valeur de la puissance après reconstruction s'approche de la puissance de détection du déséquilibre sur le véritable coalescent. La puissance de détection après reconstruction surestime la puissance de détection du déséquilibre du coalescent véritable lorsque celle-ci est faible (< 0.2) alors que l'inverse est observé lorsque celle-ci est élevée (> 0.4). Dans ce dernier cas, la puissance de détection après reconstruction sous-estime en effet la puissance de détection du déséquilibre du coalescent véritable. Ces différences entre les deux mesures (avant et après reconstruction) sont très fortement liées au nombre de nœuds qui sont résolus ainsi qu'au niveau de déséquilibre du coalescent véritable. Pour les arbres initialement déséquilibrés, la puissance de détection du déséquilibre après reconstruction est beaucoup plus dépendante du nombre de SNPs que de la valeur du déséquilibre initial car le nombre de nœuds résolus augmente avec le nombre de SNPs (la *p-value* dépend du nombre de nœuds résolus).

Puissance de détection du déséquilibre selon la méthode de « Kingman reconstruit »

Plutôt qu'utiliser le calcul d'une *p-value* centrée sur 0.5 comme cela est supposé dans la méthode par « permutation », nous comparons maintenant les valeurs de déséquilibre des arbres avec les valeurs de déséquilibre obtenues sur un jeu d'arbres reconstruits à partir de séquences simulées sur des coalescents standards de Kingman en utilisant la même taille d'échantillon et le même nombre de SNPs.

Quelle que soit la méthode de reconstruction des arbres (NJ ou PhyML), les puissances du test de détection du déséquilibre obtenues par cette méthode « Kingman reconstruit » (Figure I.C.i.5) sont légèrement inférieures à celles obtenues avec la *p-value* traditionnelle (méthode par « permutation ») (Figure I.C.i.4). Lorsque les arbres ne sont pas initialement déséquilibrés (puissance initiale < 0.1, Figure I.C.i.5), le pourcentage d'arbres déclarés déséquilibrés après reconstruction reste aussi inférieur au seuil de 10%.

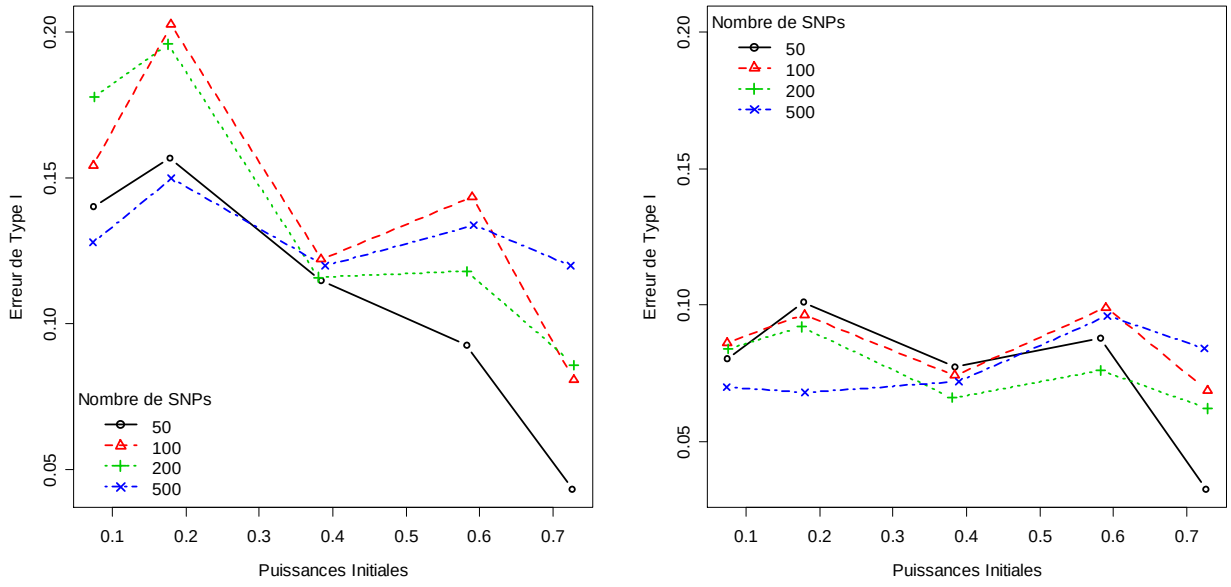


a *b*

Figure I.C.i.5, Puissance de détection du déséquilibre des arbres reconstruits selon la méthode « Kingman reconstruit » en fonction de la puissance de détection du déséquilibre des coalescents initiaux, pour quatre niveaux de polymorphisme (nombre de SNPs) et pour deux méthodes de reconstruction : (*a*) reconstruction par *Neighbour Joining* (NJ) et (*b*) reconstruction par maximum de vraisemblance (PhyML). Chaque point correspond à un couple de moyennes de puissance calculées sur 500 simulations indépendantes obtenues pour un même niveau α de transmission du succès reproducteur et en utilisant un seuil de significativité de 10%.

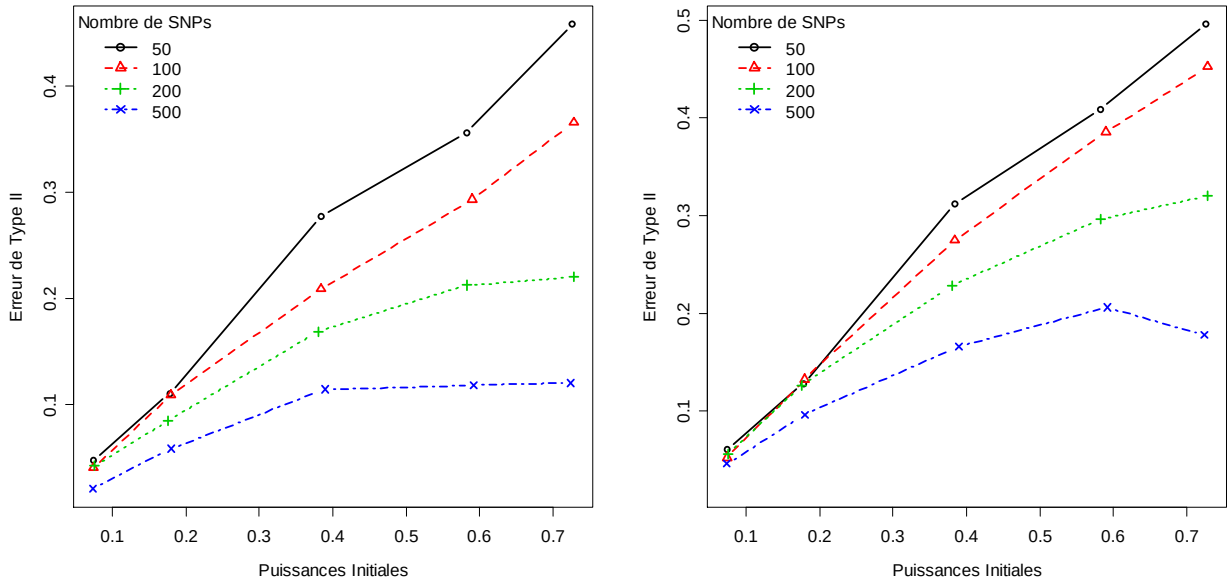
Comparaison des erreurs de détection entre les deux méthodes

De plus, l'erreur de type I (c'est-à-dire la proportion d'arbres reconstruits significativement déséquilibrés alors que les arbres initiaux correspondants sont significativement équilibrés) est plus faible lorsque nous utilisons une *p-value* basée sur la simulation de coalescents standards (méthode « Kingman reconstruit ») qu'avec la *p-value* traditionnelle calculée par une méthode de « permutation ». Par exemple, pour une reconstruction en NJ, l'erreur de type I est inférieure à 10% pour la méthode « Kingman reconstruit » alors qu'elle est entre 15 et 20% pour la méthode par « permutation » (Figure I.C.i.6). Cela est équivalent pour la reconstruction en PhyML (Résultats non montrés). Au contraire, l'erreur de type II (c'est-à-dire la proportion d'arbres initialement déclarés déséquilibrés qui sont finalement déclarés équilibrés l'issue de la reconstruction) est plus forte avec la *p-value* « Kingman reconstruit » qu'avec la *p-value* traditionnelle (Figure I.C.i.7). Au contraire de l'erreur de type I, elle est d'un part plus dépendante du nombre de SNPs puisqu'elle diminue lorsque le nombre de SNPs augmente et d'autre part, elle augmente avec la puissance de détection du déséquilibre des arbres initiaux (cf. Figure I.C.i.7).



a *b*

Figure I.C.i.6, Proportion d'arbres déclarés déséquilibrés après reconstruction par NJ alors que le coalescent initial est déclaré équilibré (Erreur de type I) en fonction de la puissance de détection du déséquilibre du coalescent initial, pour quatre niveaux de polymorphisme (nombre de SNPs) et pour deux méthodes de calcul des *p-values* : méthode par « permutation » (a) et méthode « Kingman reconstruit » (b).



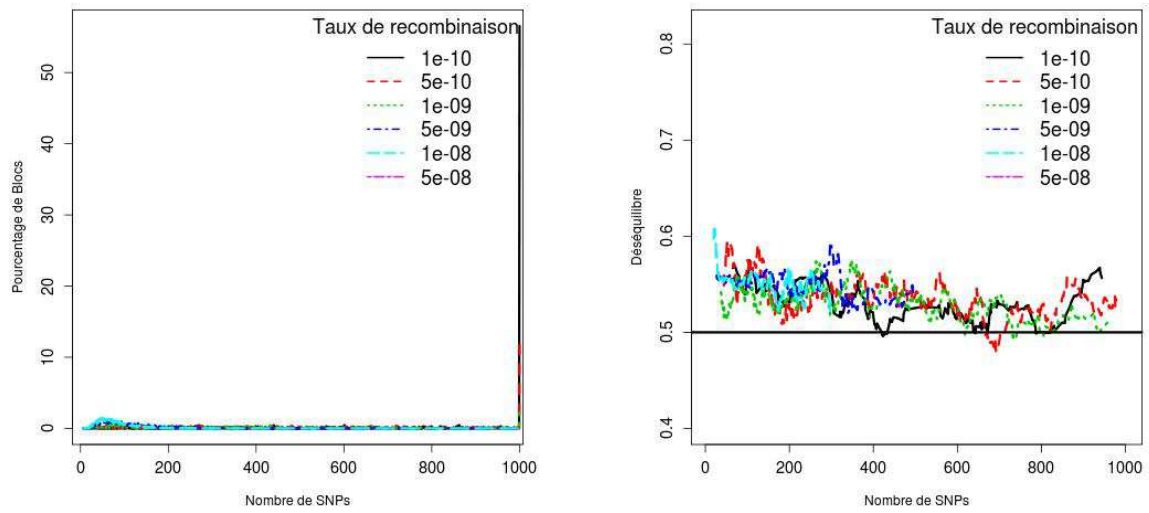
a *b*

Figure I.C.i.7, Proportion d'arbres déclarés équilibrés après reconstruction par NJ alors que le coalescent initial est déclaré déséquilibré (Erreur de type II) en fonction de la puissance de détection du déséquilibre du coalescent initial, pour quatre niveaux de polymorphisme (nombre de SNPs) et pour deux méthodes de calcul des *p-values* : méthode par « permutation » (a) et méthode « Kingman reconstruit » (b).

Reconstruction d'arbres sur des séquences simulées avec recombinaison et détection de blocks

Nous simulons 1000 SNPs sur une séquence de longueur de 100000 paires de base et des taux de recombinaison variables sans transmission du succès reproducteur. Sur chaque bloc extrait, nous reconstruisons l'arbre avec une méthode de NJ et calculons le déséquilibre de l'arbre pour vérifier si la valeur de déséquilibre calculée sur des arbres reconstruits à partir d'haploblocks va être influencée par la recombinaison.

La figure I.C.i.8.a montre la distribution du nombre d'haploblocks détecté sur lesquels ont pu être reconstruits les arbres avec plus de 3 nœuds pour calculer le déséquilibre. Comme attendu, plus le taux de recombinaison diminue plus la taille des blocks est grande puisque que pour un fort taux de recombinaison (10^{-10}) dans plus 50 % des cas il n'y aucune recombinaison (Figure I.C.i.8.a).



a *b*

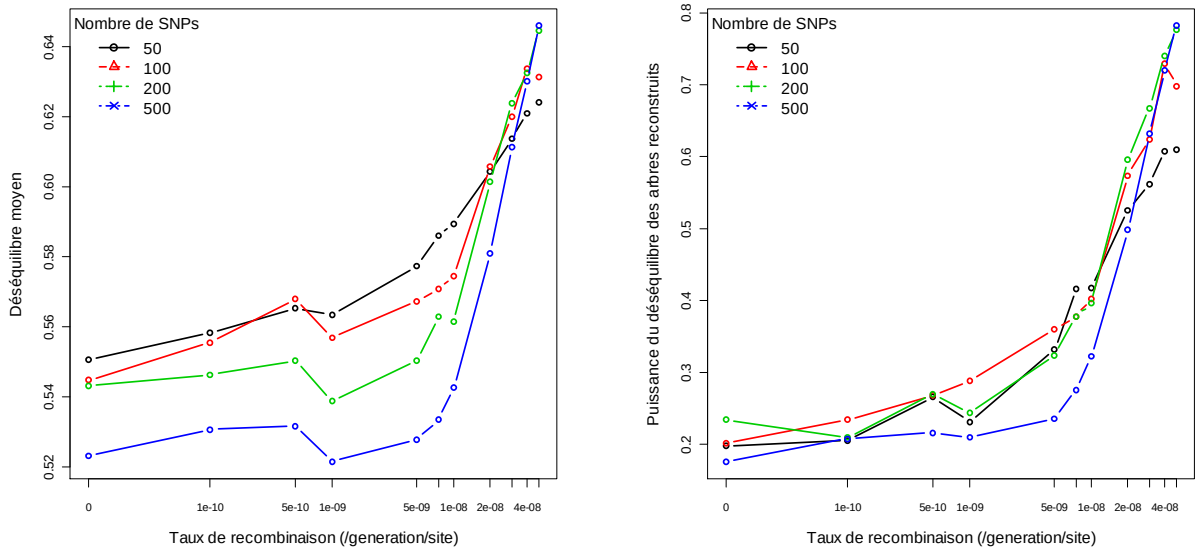
Figure I.C.i.8, Distribution du nombre de blocks sur lesquels ont pu être reconstruits les arbres avec plus de 3 nœuds pour calculer le déséquilibre (a) et déséquilibre moyen en fonction du nombre de SNPs trouvés (moyenne mobile sur 20 valeurs de SNPs). Pour chaque taux de recombinaison sont simulés 500 jeux de données de 1000 SNPs d'où sont extraits les blocks par un test des « quatre gamètes » et sur chaque bloc a été reconstruit l'arbre.

La figure I.C.i.8.b présente la moyenne mobile sur vingt valeurs du déséquilibre pour un nombre de SNPs donné. Le déséquilibre comme précédemment décroît avec le nombre de SNPs à partir desquels les arbres ont été reconstruits mais il ne dépend pas du taux de recombinaison (Figure I.C.i.8.b). Nous observons plus de stochasticité dans la valeur de déséquilibre (Figure I.C.i.8.b) que dans le tableau I.C.i-1 ; ceci est dû au plus faible nombre d'arbres utilisés pour calculer les moyennes pour chaque nombre de SNPs puisque le nombre de blocks utilisé pour la reconstruction va être inférieur aux 500 valeurs simulées dans le tableau et varier d'un nombre de SNPs à l'autre.

Impact des recombinaisons sur le déséquilibre

Nous simulons des séquences, pour différents taux de recombinaison et quatre valeurs du nombre de SNPs. Sur ces séquences, nous reconstruisons les arbres par une méthode de NJ sans prendre en compte les recombinaisons éventuelles. Pour les simulations sans recombinaison, le déséquilibre diminue avec le niveau de polymorphisme des séquences (Figure I.C.i.3a). Le déséquilibre des arbres augmente avec le taux de recombinaison, mais si le taux de recombinaison par site est inférieur à 10^{-9}, le déséquilibre moyen et la puissance sont peu sensibles aux différents taux de recombinaison (Figure I.C.i.9.a). Alors que si le taux de recombinaison est supérieur à cette valeur, le déséquilibre va augmenter avec le taux de recombinaison. Au contraire des résultats obtenus ci-dessus, où la puissance sur les arbres reconstruits diminuait avec le nombre de SNPs quand il y avait de la présence de succès

reproducteur, la valeur de la puissance ne dépend pas de ce nombre de SNPs. Elle augmente par contre fortement avec le taux de recombinaison, jusqu'à des valeurs proches de 80%. Ceci indique que de nombreux arbres reconstruits apparaissent comme déséquilibrés alors qu'il n'y a pas de transmission du succès reproducteur dans la population.



a *b*

Figure I.C.i.9, Déséquilibre moyen (a) et puissance de détection avec une méthode de permutation (b) en fonction du taux de recombinaison pour des reconstructions d'arbres sur une échelle logarithmique par la méthode NJ sur des séquences de longueur 100 000 paires de bases simulées sur des coalescents standards avec recombinaison. La puissance de détection est calculée par la méthode de « permutation ». Chaque point correspond à une moyenne du déséquilibre ou à une puissance de détection calculée sur 500 arbres indépendants reconstruits.

HapMap

Parenté entre copies de chromosome

Le nombre initial de copies de chromosome était de 180 dans la population asiatique et 120 dans les populations européenne et africaine. A l'issue du filtrage effectué pour retirer les copies de chromosome trop apparentées, 32% des copies ont été conservées chez les asiatiques (JPT+CHB) et les européens (CEU) alors que chez les africains (YRI) ce pourcentage est de 42%. Le nombre de copies de chromosome retirées varie selon le chromosome. Par exemple, sur le chromosome 21, il ne reste que 25% des copies par population, alors que sur chromosome 2, 52% des copies sont conservées dans les trois populations (Tableau I.C.i-S1).

Distribution du nombre de sites polymorphes par bloc

Comme nous avons vu précédemment, le nombre d'arbres qui ont suffisamment de nœuds résolus (>3) sur lesquels le déséquilibre peut être calculé, augmentait avec le nombre

de SNPs (cf. partie *Simulation*). Il est donc important d'évaluer dans quelle mesure le jeu de données HapMap est susceptible de contenir des blocks suffisamment longs pour espérer reconstruire correctement des arbres et évaluer correctement leur déséquilibre. Le nombre moyen de sites polymorphes par bloc est variable selon les trois populations du projet HapMap 2 (Tableau I.C.i-2). Si le nombre de blocks détectés est plus important dans la population africaine (YRI) que dans les populations européenne (CEU) et asiatique (JPT+CHB), le nombre moyen de SNPs par bloc y est par contre plus faible.

Tableau I.C.i-2, Nombre de blocks, nombre de SNPs et nombre moyen de SNPs par bloc polymorphe présents dans les trois populations d'Asie (JPT+CHB), d'Europe (CEU) et d'Afrique (YRI)

	CEU	*JPT+CHB*	*YRI*
Nombre de blocks détectés	352 823	388 721	597 921
Nombre de SNPs	2 479 649	2 359 017	2 806 262
Nombre moyen de SNPs par bloc polymorphe	7.03	6.07	4.69

La majorité des blocks contiennent moins de cinq SNPs, avec un faible nombre de blocks qui ont plus de 40 SNPs (Tableau I.C.i-3) Il existe une variabilité en fonction des populations, le nombre de blocks de plus de 40 SNPs étant plus important dans la population européenne, environ deux fois plus qu'en Asie et six fois plus qu'en Afrique.

Tableau I.C.i-3, Répartition des blocks détectés en pourcentage (et en nombre entre parenthèses) en fonction de cinq classes de nombre de SNPs par bloc dans les trois populations d'Asie (JPT+CHB), d'Europe (CEU) et d'Afrique (YRI)

Nombre de SNPs par bloc	*CEU*	*JPT+CHB*	*YRI*
< 5	50.67 (178 786)	55.38 (215 273)	63.45 (379 368)
5-9	26.18 (92 378)	25.90 (100 680)	25.33 (151 441)
10-19	16.54 (58 364)	14.22 (55 266)	9.60 (57 421)
20-39	5.66 (19 983)	4.04 (15 720)	1.54 (9 202)
≥ 40	0.94 (3 312)	0.46 (1 781)	0.08 (489)
Total	100 (352 823)	100 (388 721)	0.08 (597 921)

Calcul du déséquilibre des arbres reconstruits

Dans chacune des trois populations de HapMap (CEU, JPT+CHB et YRI), pour chaque bloc détecté présentant plus de 40 SNPs, nous reconstruisons l'arbre généalogique du bloc en utilisant la méthode NJ ou la méthode PhyML.

Les déséquilibres moyens observés sur les arbres des trois populations de HapMap sont au-dessus de la valeur de 0.5 attendue pour un arbre de coalescence standard de Kingman que cela soit avec une reconstruction NJ ou une reconstruction par PhyML (Tableau I.C.i-4 et Tableau I.C.i-5). Les valeurs de déséquilibre obtenues par NJ sont inférieures à celles obtenues par PhyML pour les populations européenne et africaine (test bilatéral de Student sur séries appariées, à un seuil de 5%) alors qu'elles ne sont pas significativement différentes dans la population asiatique.

Par ailleurs, si on compare les trois populations entre elles, elles présentent des valeurs moyennes de déséquilibre significativement différentes lorsque la reconstruction a été faite par la méthode NJ (Tableau I.C.i-4) alors que par la méthode PhyML, seuls les déséquilibres moyens des populations asiatique et européenne sont significativement différents (Tableau I.C.i-5) (tests t de Student, à 5%).

Significativité du déséquilibre des arbres

Calcul des *p-value* par permutation

Si l'on calcule les *p-values* selon la méthode de « permutation », en fonction de la population considérée, entre 32% et 44 % des blocks correspondent à des arbres significativement déséquilibrés pour un seuil de 10%, quelle que soit la méthode de reconstruction (Tableau I.C.i-4,5). Pour la méthode de reconstruction en NJ et la *p-value* calculée par permutation, le pourcentage est le plus élevé dans la population asiatique (43.21%) suivie de la population africaine (36.19 %), puis de la population européenne (36.19%) (Tableau I.C.i-4). Pour une reconstruction avec le logiciel PhyML, nous ne retrouvons pas le même classement entre populations, puisque le pourcentage d'arbres déséquilibrés dans la population africaine (43.53%) est supérieur à celui trouvé dans la population asiatique (42.02%) (Tableau I.C.i-5).

Calcul des *p-value* par « Kingman Reconstruit »

Nous calculons la puissance grâce à la distribution nulle des déséquilibres obtenue en reconstruisant des arbres avec le même nombre de SNPs et la même taille d'échantillon. La proportion d'arbres déclarés déséquilibrés au seuil de 10% est plus faible avec cette méthode que la précédente (Tableau I.C.i-4).

Tableau I.C.i-4, Nombre moyen de séquences, nombre d'arbres reconstruits, nombre moyen de nœuds par arbre reconstruit, déséquilibre moyen des arbres reconstruits et proportion d'arbres déclarés déséquilibrés au seuil de 10% par la méthode par « permutation » ou par la méthode « Kingman reconstruit », en utilisant la méthode de reconstruction NJ sur les blocks identifiés à partir des données HapMap de trois populations (JPT+CHB, CEU et YRI).

	Nombre moyen de séquences	*Nombre d'arbres*	*Nombre de nœuds*	*Déséquilibre moyen*	*Proportion d'arbres déclarés déséquilibrés au seuil de 10%*	
					Méthode par « permutation »	*Méthode « Kingman reconstruit »*
CEU	42.13	1650	10.55	0.6334	34.55%	27.09%
JPT+CHB	60.3	759	9.93	0.6777	43.21%	36.76%
YRI	56.03	315	15.04	0.6257	36.19%	21.59%

Tableau I.C.i-5, Nombre moyen de séquences, nombre d'arbres reconstruits, nombre moyen de nœuds par arbre reconstruit, déséquilibre moyen des arbres reconstruits et proportion d'arbres déclarés déséquilibrés au seuil de 10% par la méthode par « permutation », en utilisant la méthode de reconstruction PhyML sur les blocks identifiés à partir des données HapMap de trois populations (JPT+CHB, CEU et YRI).

	Nombre moyen de séquences	*Nombre d'arbres*	*Nombre de nœuds*	*Déséquilibre moyen*	*Proportions d'arbres déclarés déséquilibrés au seuil de 10%*
					Méthode par « permutation »
CEU	42.06	1046	10.28	0.6591	38.34%
JPT+CHB	60.19	426	9.64	0.6818	42.02%
YRI	55.65	232	14.2	0.6710	43.53%

Comparaison des déséquilibres calculés entre NJ et PhyML

Nous sélectionnons dans les trois populations les arbres dont la valeur de déséquilibre a pu être calculée en NJ et en PhyML donc lorsque le nombre de nœuds où le déséquilibre peut être calculé est supérieur à trois pour les deux arbres. Cela nous fait un total de 1684 arbres. Du fait de ces sélections, nous conservons 99% des arbres reconstruits par la méthode de PhyML mais seulement 66% des arbres reconstruits initialement par la méthode de NJ ; ce qui est logique car beaucoup plus d'arbres avaient pu être reconstruits par NJ Parmi ces arbres reconstruits par NJ, nous constatons que les arbres conservés sont en moyenne plus déséquilibrés que la totalité des arbres reconstruits par cette méthode.

Nous trouvons donc un ensemble de 1684 arbres en commun dans les trois populations. Pour un même bloc, les deux méthodes de reconstruction aboutissent à des topologies relativement similaires en termes de déséquilibre puisque le coefficient de corrélation entre les valeurs de déséquilibre des deux méthodes est de 0.88 sur les 1684 arbres.

Tableau I.C.i-6 Nombre d'arbres, déséquilibres moyens et proportions d'arbres déséquilibrés à un seuil de 10 % pour les déséquilibres en commun reconstruits par une méthode de NJ et une méthode PhyML

		Déséquilibres Moyens		*Proportions d'arbres déclarés déséquilibrés au seuil de 10%*	
	Nombre d'arbres	*NJ*	*PhyML*	*NJ*	*PhyML*
CEU	1030	0.6444	0.6597	36.31%	38.45%
JPT+CHB	422	0.6746	0.6825	41.71%	42.18%
YRI	232	0.6393	0.671	40.95%	43.53%

Si, dans ce cas, on compare les valeurs de déséquilibre entre les populations, nous observons que les populations européenne (CEU) et africaine (YRI) ne sont pas différentes significativement (test bilatéral t de Student à 5%) dans les deux modes de reconstruction. Les populations européenne et asiatique (JPT+CHB) sont significativement différentes (test bilatéral t de Student à 5%) dans les deux modes de reconstruction. La seule différence entre les deux méthodes est observée quand on compare les populations africaine et asiatique ; la différence est significative en NJ et non significative en PhyML.

Pour les proportions d'arbres déséquilibrés, les différences entre les deux méthodes sont relativement faibles. On compare les différences de « décision » entre les deux méthodes (c'est-à-dire la proportion d'arbres qui sont dits équilibrés dans une méthode et qui ne le sont pas dans l'autre méthode et *vice versa*). La proportion de ces différences de décision est de 12 % pour les populations européenne et africaine et de 18 % pour la population asiatique.

Analyse de la différence d'hétérozygotie entre le génome et les blocks sélectionnés

Nous analysons la différence d'hétérozygotie entre les blocks sélectionnés pour savoir si nos résultats sont représentatifs du reste du génome.

Les hétérozygoties moyennes des blocks sélectionnés pour la reconstruction phylogénétique (c'est-à-dire les blocks possédant plus de 40 SNPs) sont significativement inférieures à celle du génome entier pour toutes les populations (tests bilatéraux t de Student, seuil de 5%) (Tableau I.C.i-7).

Tableau I.C.i-7, Moyenne de l'hétérozygotie par base sur l'ensemble du génome et sur les blocks sélectionnés pour la reconstruction phylogénétique (blocks possédant plus de 40 SNPs).

	Génome entier	*Blocks sélectionnés*
CEU	0.3192	0.2968
JPT+CHB	0.3165	0.2792
YRI	0.2991	0.2769

Enfin, parmi les blocks, nous séparons ceux qui sont strictement inter-géniques (SNP « non-génique » dans le Tableau I.C.i-7) de ceux qui sont étiquetés comme étant localisés dans des introns, des exons, des régions 5' UTR ou des régions 3' UTR (SNP « génique » dans le Tableau I.C.i-8). Dans la population européenne (CEU), le pourcentage d'arbres déclarés déséquilibrés pour les portions non géniques est inférieur à celui trouvé pour les portions géniques (test de χ^2, seuil de 5%). Il n'y a pas de différences significatives pour les deux autres populations.

Tableau I.C.i-8, Pourcentage (nombre entre parenthèses) d'arbres déclarés déséquilibrés au seuil de 10% selon la localisation génomique (« non-génique » ou « génique ») des blocks ayant servis à les reconstruire (reconstruction NJ).

	Non génique	*Génique*
CEU	24.80 (754)	29.02 (896)
JPT+CHB	38.15 (346)	35.59 (413)
YRI	17.46 (126)	24.34 (189)

I.C.i.d Discussion

Nous montrons que l'utilisation des séquences autosomales est une possibilité pour détecter des phénomènes de transmission du succès reproducteur en utilisant le déséquilibre des arbres que crée ce phénomène. Le déséquilibre calculé sur des arbres reconstruits à partir de séquences neutres dépend de la méthode de reconstruction utilisée, du nombre de nœuds présents dans l'arbre (liés au nombre de sites polymorphes) et aussi de la bonne détection des recombinaisons.

Sous un modèle de mutation à nombre infini de sites, nous montrons que le déséquilibre des arbres reconstruits dépend de la méthode de reconstruction. Les méthodes de *Neighbour-Joining* (NJ) et de maximum de vraisemblance (PhyML) permettent de reconstruire des arbres dont le déséquilibre est beaucoup plus proche de celui de l'arbre de coalescence réel que ne le permet la méthode UPGMA. Nous retrouvons donc des résultats similaires à ceux obtenus par Blum et *al.* (2006) qui montraient que les méthodes de

reconstruction par maximum de vraisemblance sur des données simulées sous le modèle d'évolution moléculaire HKY85 (Hasegawa et al. 1985) étaient meilleures que la méthode UPGMA. Ces résultats sont en contradiction avec ceux de Huelsenbeck et *al.* (1996) qui ont montré que la méthode UPGMA retrouvait mieux la valeur initiale du déséquilibre que les méthode du maximum de vraisemblance et de *Neighbor-Joining*. Pour cela, ces auteurs ont simulé des séquences avec un modèle d'évolution moléculaire JC69 (Jukes et al. 1969) et de forts taux de substitution et ils ont calculé le déséquilibre des arbres en utilisant d'autres méthodes de calcul que l'indice de déséquilibre I_{nb} que nous avons utilisé.

En général, l'excès de déséquilibre dans des arbres reconstruits peut être dû au fait que les branches résolues, c'est-à-dire la présence d'au moins une mutation, sont préférentiellement les branches externes qui se connectent haut dans l'arbre. Comme elles sont plus longues que les autres branches, la probabilité qu'une mutation se produise au cours du temps est plus grande. Le déséquilibre moyen d'un coalescent est une moyenne de nœuds équilibrés et non équilibrés ; si les nœuds portant des branches externes sont mieux résolus que les autres, on observera donc une augmentation de la valeur de déséquilibre.

L'augmentation du nombre de SNPs au sein de nos haploblocks diminue le déséquilibre observé, les arbres se rapprochent des arbres initiaux car le nombre de branches et le nombre de nœuds résolus augmentent. Lorsque les blocks sur lesquels sont reconstruits les arbres ne sont pas bien définis (la reconstruction se fait directement sur des séquences où il y a présence de recombinaisons), nous montrons que le déséquilibre des arbres augmente avec le nombre de recombinaisons dans la séquence. Ce résultat se situe dans la lignée d'études antérieures (Anisimova et al. 2003; Arenas & Posada 2010; Posada & Crandall 2002; Ramírez-Soriano et al. 2008) qui ont montré que la présence de recombinaisons dans une séquence modifiait fortement la forme de l'arbre qui était reconstruit. En effet, les séquences recombinées correspondent à plusieurs arbres de coalescence sous-jacents qui ne peuvent pas se résumer en un seul arbre. Nous montrons par ailleurs que la reconstruction de ce type d'arbre moyen ne permet pas d'avoir accès au déséquilibre moyen de tous les arbres sous-jacents. Au contraire, si on définit bien les blocks, la recombinaison n'influence pas le déséquilibre de notre arbre qui est alors d'autant mieux estimé que le nombre de SNPs est élevé.

La transmission du succès reproducteur augmente le déséquilibre au sein des arbres. Lorsque les arbres de gènes sont extraits d'une population soumise à ce phénomène, ils sont plus déséquilibrés que le coalescent standard. Ce déséquilibre dépend de l'intensité du

phénomène, de la taille de population ou de l'hétérogénéité du succès reproducteur (cf. partie I.B). Nous observons que le déséquilibre est conservé dans les arbres reconstruits puisque la moyenne du déséquilibre de ces arbres augmente avec la valeur moyenne initiale du déséquilibre même si elle reste surestimée légèrement par rapport à la valeur initiale.

La valeur du déséquilibre ne nous permet pas de définir si statistiquement l'arbre est plus déséquilibré qu'attendu. Dans le cas d'un coalescent de Kingman, la valeur attendue pour le déséquilibre était de l'ordre de 0.5 et la permutation, avec une probabilité de ½ de la valeur du déséquilibre de chaque nœud, permettait de définir une *p-value* et de déclarer si l'arbre étudié était déséquilibré ou non, pourvu qu'un seuil de significativité ait été défini. Nous montrons que cette méthode de calcul des *p-values* par « permutation » est non conservative puisque le pourcentage d'arbres déclarés déséquilibrés s'élève à 15% voire 20%, alors qu'il devrait être proche de 10%. Cette *p-value* traditionnelle qui montrait sa performance pour les arbres complets montre ses limites lorsque les arbres reconstruits ne sont pas entièrement résolus. Lorsque les arbres initiaux sont effectivement déséquilibrés, les *p-values* de ces arbres reconstruits sont plus dépendantes du niveau de polymorphisme des blocks (nombre de SNPs) que du déséquilibre du coalescent initial. Plus le nombre de SNPs est important, plus la puissance de détection du déséquilibre après reconstruction se rapproche de la valeur initiale. Comme la variance du déséquilibre augmente avec la diminution du nombre de SNPs, même si les moyennes du déséquilibre sont proches, les différences entre les puissances vont être plus marquées.

Pour corriger ces biais, nous proposons d'utiliser une *p-value* basée sur la simulation de coalescents standards de Kingman servant d'arbres sous-jacents pour la simulation d'un polymorphisme de même nature que celui qui est observé. Le polymorphisme est utilisé pour reconstruire un arbre par la même méthode de reconstruction utilisée sur les données observées. Cette approche permet de rendre la *p-value* indépendante de la méthode de reconstruction et du type d'information présente dans les séquences. Ce mode de calcul de la *p-value* repose sur l'idée que la surestimation du déséquilibre des arbres reconstruits par la méthode classique de « permutation » est principalement le résultat d'une combinaison entre la méthode de reconstruction de l'arbre et le nombre de nœuds résolus dans l'arbre. Le nombre de nœuds observés dans l'arbre reconstruit est lui-même une combinaison de la taille de l'échantillon, du nombre de sites polymorphes et de la forme initiale du coalescent. Nous montrons que dans un jeu d'arbres reconstruits à partir de séquences obtenues selon des coalescents non déséquilibrés, la puissance de détection du déséquilibre selon notre méthode

alternative est inférieure à celle qui est calculée de manière traditionnelle, ce qui signifie que nous limitons par cette approche le risque de conclure à tort à la présence de déséquilibre. Lorsqu'on utilise un jeu d'arbres initialement déséquilibrés (obtenus sous transmission du succès reproducteur), les puissances de détection après reconstruction augmentent avec la puissance de détection initiale. Comme pour la méthode par « permutation », les puissances calculées selon la méthode « Kingman reconstruit » sont dépendantes du nombre de SNPS présents sur les séquences. Ce nouveau calcul de la *p-value* limite le risque de conclure après reconstruction à la présence de déséquilibre alors que le coalescent initial est équilibré. Cependant, cette méthode est conservative et a tendance à nous faire conclure que l'arbre est équilibré alors qu'il est initialement déséquilibré, puisque les puissances calculées sur les arbres reconstruits sont inférieures aux puissances des arbres initiaux.

En ce qui concerne l'analyse des données HapMap, en appliquant le test de parenté (cf. Eq. I.C.i-a) et en considérant le niveau chromosomique, plus de 50% des chromosomes sont supprimés de notre échantillon. Le constat de la présence d'individus fortement apparentés au sein d'HapMap est connu : 8 paires d'individus le sont (voir par exemple le tableau The International HapMap Consortium 2005). Notre approche qui considère non plus les individus mais chaque chromosome montre que de nombreux chromosomes sont fortement similaires entre eux.

Nous montrons que le nombre de blocks pour lesquels le nombre de sites polymorphes est supérieur à 40 représentent moins de 1% du nombre de blocks dans chaque population. Les valeurs de déséquilibre moyen observées sur les arbres reconstruits sur ces blocks sont plus importantes dans la population asiatique que dans les deux autres populations quelle que soit la méthode. Si on utilise la méthode par « permutation » pour calculer la *p-value* du déséquilibre des arbres, entre 35% à 45% des arbres sont déclarés déséquilibrés au seuil de 10%. Grâce à la méthode basée sur la distribution nulle des déséquilibres dans les arbres de Kingman - méthode qui prend en compte la taille de l'échantillon, le nombre de SNPs et la méthode de reconstruction - le pourcentage d'arbres déclarés déséquilibrés diminue. Cela montre que, dans le cas d'une reconstruction avec une méthode de *Neighbor-Joining*[2], le pourcentage d'arbres faussement significatifs par la méthode de permutation varie de 20% pour les Africains à 35% pour les Asiatiques

[2] Pour des raisons de temps de calcul, nous n'avons pu faire la même approche avec PhyML.

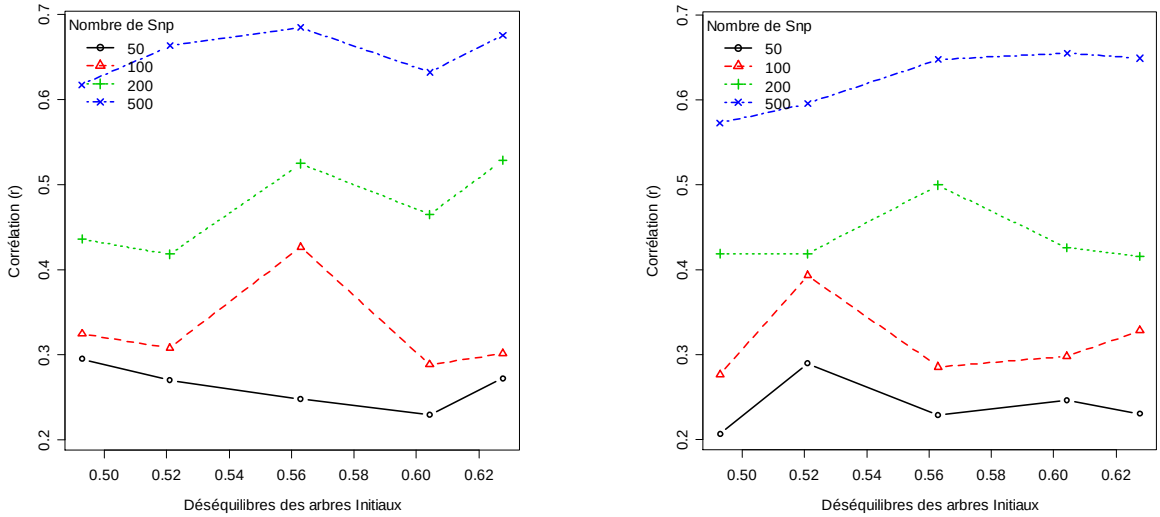
Aucun a priori existait sur une possible structure sociale au sein des populations pouvant expliquer un déséquilibre dans les arbres. Pour autant, la proportion d'arbres déséquilibrés dans les trois populations est supérieure au seuil attendu de 10 %. Pour vérifier la validité de notre méthode « Kingman reconstruit », nous avons utilisé la même approche en utilisant une reconstruction par un algorithme de maximum de vraisemblance plus rapide (fastme, Richard & Olivier 2002) que PhyML. Nous trouvons des proportions d'arbres plus déséquilibrés qu'attendus légèrement plus importantes pour les puissances de « Kingman reconstruit » que dans le cas d'une reconstruction NJ ; ce pourcentage d'arbres déséquilibrés dans les populations asiatiques est plus important que pour les deux autres (Tableau I.C.i-S3). Ceci confirme les résultats obtenus sur les puissances de « Kingman reconstruit » obtenus en NJ : nous obtenons des puissances plus élevées pour la population asiatique que pour la population européenne, la population africaine présentant les puissances les plus faibles.

Plusieurs hypothèses peuvent expliquer cet excès de déséquilibre dans les trois populations. Premièrement, les données HapMap ont un déficit dans les SNPs en faible fréquence (Clark et al. 2005), ce qui pourrait entraîner un défaut de détection de certaines recombinaisons et par voie de conséquence, l'augmentation du déséquilibre dans les arbres reconstruits. En second lieu, notre méthode alternative de calcul de la *p-value* prend en compte à la fois la théorie du coalescent standard de Kingman, la taille de l'échantillon, le niveau de polymorphisme mais également la méthode de reconstruction, et permet ainsi de diminuer le pourcentage d'arbres déclarés à tort déséquilibrés. Pour autant, les séquences simulées sur des coalescents de Kingman ont une hétérozygotie moyenne plus faible que celle observée sur HapMap, ce qui peut être expliqué par le déficit en allèles rares dans HapMap. Par conséquent, à même nombre de SNPs, ce déficit entraîne une différenciation moindre entre les individus deux à deux (SNPs manquants) et donc une reconstruction d'arbre moins performante. Ce problème pourra sans doute être réglé en développant une méthode de simulation de la distribution nulle prenant en compte le biais en défaveur des allèles rares.

Par ailleurs, vu que nous n'observons pas de différence entre les arbres sur des portions géniques et non géniques, nous pouvons dire que le fort déséquilibre dans la population asiatique ne semble pas lié à la présence de sélection sur des locus précis. Ce déséquilibre pourrait être lié à des structurations sociales plus complexes au sein des populations asiatiques que dans les autres populations. Il convient aussi de noter que la population asiatique de HapMap est en réalité constituée de Japonais (JPT) et de Chinois (CHB). Il conviendra donc de répéter ces analyses en séparant ces deux populations. En ce

qui concerne la population européenne (CEU), l'enrichissement significatif en arbres déséquilibrés dans les régions géniques pourrait être lié à la présence de sélection. Il conviendra dans ce cadre d'analyser plus finement la nature des régions géniques ayant des arbres déséquilibrés. La population africaine apparaît quant à elle comme étant la plus proche de l'équilibre.

I.C.i.e Informations supplémentaires



a *b*

Figure I.C.i-S1, Corrélation *r* entre le déséquilibre de l'arbre initial et le déséquilibre de l'arbre reconstruit en fonction du déséquilibre moyen des arbres initiaux, pour quatre niveaux de polymorphisme (nombre de SNPs) et pour deux méthodes de reconstruction : (*a*) reconstruction par *Neighbour-Joining* (NJ) et (*b*) reconstruction par maximum de vraisemblance (PhyML).

Tableau I.C.i-S1, Nombre de copies de chromosome utilisées pour reconstruire les arbres pour les trois populations de HapMap et pour chacun des 22 chromosomes autosomaux

Chromosome	*CEU*	*JPT+CHB*	*YRI*
1	55	78	76
2	56	73	76
3	44	64	65
4	50	70	62
5	44	68	67
6	43	65	55
7	44	64	61
8	40	60	63
9	41	52	53
10	46	62	58
11	38	52	55
12	39	65	54
13	36	53	48
14	38	48	39
15	33	47	41
16	38	52	45
17	24	51	44
18	31	48	47
19	30	43	35
20	36	46	37
21	31	34	35
22	27	41	35

Tableau I.C.i-S2, Longueur moyenne par bloc (kb).

CEU	*JPT-CHB*	*YRI*
6.4	5.6	3.4

Tableau I.C.i-S3, nombre d'échantillons, d'arbres, proportions d'arbre significatif au seuil 10% par méthode de permutation ou avec des simulations d'arbres de Kingman calculése sur des arbres reconstruits par fastme sur les blocks extraits dans trois populations.

	Nombres d'échantillons	*Nombres d'arbres*	*Nombres de nœuds*	*Déséquilibres moyens*	*Proportion d'arbre significatif au seuil de 10% (méthode par permutation)*	*Proportion d'arbre significatif au seuil de 10% (Méthode de « Kingman reconstruit »*
CEU	42.17	1740	10.62	0.6471	38.45%	30.75%
JPT+CHB	60.22	797	10.03	0.687	46.8%	41.41%
YRI	55.97	321	15.53	0.6497	43.61%	25.23%

I.C.ii Détection sur le Chromosome Y : Exemple de la transmission patrilinéaire en Asie Centrale

Collaboration avec Evelyne Heyer (MNHN) et Michela Leonardi (Palaeogenetic group, Institute of Anthropology, Johannes Gutenberg-Universität Mainz).

I.C.ii.a Introduction

Depuis le 4e millénaire avant notre ère, des populations de type pastoral à langue d'origine turco-mongole et des populations agricultrices à langue d'origine indo-iranienne coexistent en Asie centrale (Cavalli-Sforza et al. 1994). Ces populations, si elles coexistent, présentent une organisation sociale différente. Les sociétés d'agriculteurs sont organisées en familles endogames et de filiation cognatique (mode de descendance passant indifféremment par les hommes et les femmes) alors que les sociétés d'éleveurs sont constituées en groupes préférentiellement exogames (choix du conjoint préférentiellement en dehors du groupe) et patrilocaux (migration préférentielle des femmes vers le lieu de résidence des maris). Ces groupes sont organisés en tribus, clans et lignées, suivant une filiation patrilinéaire (mode de descendance passant par les hommes) (Chaix et al. 2007; Jacquesson 2002).

Au niveau génétique, on observe une réduction de la diversité sur le chromosome Y (Chaix et al. 2007) chez les populations pastorales par rapport aux populations agricultrices ainsi qu'une réduction de l'effectif efficace des hommes par rapport à celui des femmes chez les populations pastorales. Ces observations ne peuvent pas être expliquées en totalité par une différence de migration sexe-spécifique et pourraient en réalité provenir principalement du caractère patrilinéaire de l'organisation sociale des populations agricultrices (Heyer et al. 2009; Segurel et al. 2008), dans la mesure où ce type d'organisation sociale pourrait être à l'origine d'une transmission du succès reproducteur.

La transmission du succès reproducteur entraîne plusieurs impacts sur la forme des arbres de coalescence : (i) des arbres de taille plus petite et une réduction de l'effectif efficace ; (ii) des arbres en forme d'étoile et une augmentation de la fréquence des allèles rares ; (iii) des arbres plus déséquilibrés (Blum et al. 2006; Sibert et al. 2002). L'utilisation du déséquilibre des histoires de gènes reconstruits à partir de données du chromosome Y semble être le meilleur moyen de détecter la transmission du succès reproducteur par la voie masculine. Plusieurs approches ont été tentées pour reconstruire les arbres phylogénétiques sur les séquences d'individus échantillonnés dans des populations. Les méthodes de

vraisemblance et de *Neigbour-joining* ont montré leur performance par rapport aux méthodes d'UPGMA (Blum et al. 2006 et partie I.C.i). De plus, les reconstructions d'arbres pour détecter la transmission du succès reproducteur ont été effectuées sur des séquences, en supposant un modèle mutationnel permettant l'homoplasie (Blum et al. 2006) ou avec un modèle de mutation à nombre infini de sites (partie I.C.i). Nous allons chercher à déterminer si la transmission du succès reproducteur peut expliquer les différences observées au niveau du chromosome Y entre populations patrilinéaires (éleveurs) et cognatiques (agriculteurs) en utilisant des données microsatellites et SNPs au mode de mutation différent.

I.C.ii.b Matériels et Méthodes

Au total, 8 populations patrilinéaires et 11 populations cognatiques ont été échantillonnées à travers l'Asie centrale. Le tableau I.C.ii.S1 résume le nombre d'individus, le nombre de SNPs et de microsatellites analysés. Les données et populations utilisées sont décrites en détail dans Ségurel et *al.* (2008) (leur localisation géographique est donnée dans la Figure I.C.ii-1 de ce chapitre, reproduite ci-dessous).

Figure I.C.ii-1. Carte géographique de l'aire d'échantillonnage de l'étude. Les populations cognatiques sont en bleu, (Tajiks) et jaune (Uzbek) ; les populations patrilinéaires avec un mode de vie semi-nomade sont en violet (Kazak), rouge (Kirghise), vert (Karakalpaks) et rose (Turkmen) ; Carte extraite de (Heyer et al. 2009).

Pour reconstruire les arbres, nous combinons les méthodes de reconstruction phylogénétique. Pour chaque population nous regroupons les individus en fonction de leur haplotype SNPs. Ensuite, nous construisons l'arbre des haplotypes avec une méthode de maximum de vraisemblance, en utilisant le logiciel PhyML v. 3.0 (Guindon et al. 2010). Puis, pour chaque groupe d'individus ayant le même haplotype SNPs, nous reconstruisons sur les données microsatellites le sous-arbre avec une méthode de NJ (Saitou & Nei 1987) à l'aide du package APE (Paradis et al. 2004). Nous fusionnons l'arbre des haplotypes avec tous les sous-arbres obtenus à partir des données microsatellites pour obtenir un arbre complet (cf. Figure I.C.ii-2). Nous calculons avec la méthode décrite en partie I.B.iii le déséquilibre (I_{nb}, cf. équation I.B-a) et calculons une *p-value* évaluée selon la méthode par « permutation » d'Agapow et Purvis (Agapow & Purvis 2002).

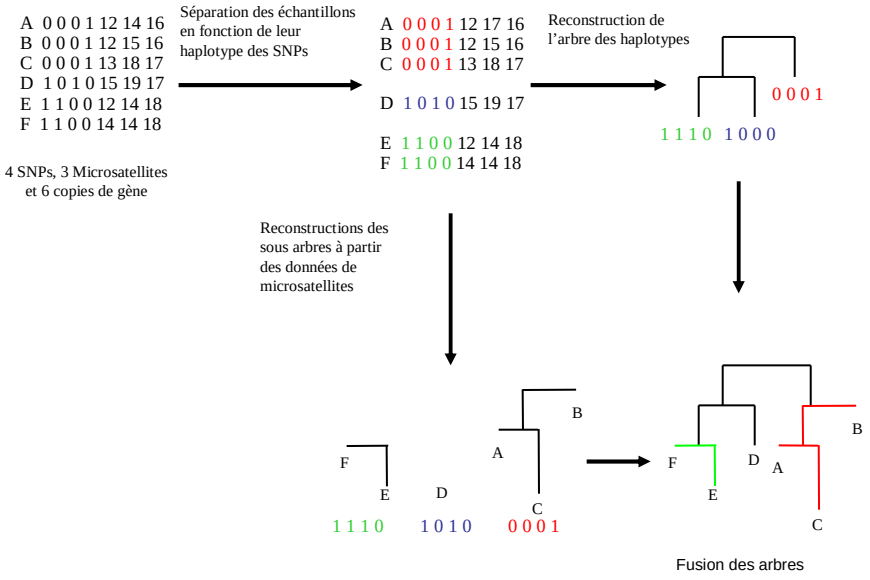


Figure I.C.ii-2, Méthode de reconstruction combinant des données de SNPs et de microsatellites pour 4 SNPs , 3 Microsatellites et 6 copies de gène.

Aussi, pour chaque population, nous calculons l'hétérozygotie sur chacun des sites polymorphes puis nous calculons une moyenne sur les données de type SNP et de type microsatellite. Nous comparons les moyennes d'hétérozygotie par site et par population avec les valeurs de déséquilibre.

I.C.ii.c Résultats

Dans toutes les populations, les valeurs de déséquilibre observées sont en moyenne plus élevées que la valeur attendue de 0.5 pour un coalescent de Kingman. Les populations qui possèdent une filiation patrilinéaire ont un déséquilibre plus important que les populations de filiation cognatique (Test unilatéral de rangs de Wilcoxon entre les valeurs de déséquilibre, *p-value* = 0.037, cf. Tableau I.C.ii-1, 2, 3).

Tableau I.C.ii-1 Médiane, moyenne et écart-type des déséquilibres des arbres reconstruits pour les populations à filiation cognatique et les populations à filiation patrilinéaire.

	Populations cognatiques	*Populations patrilinéaires*
Médiane	0.6606	0.8201
Moyenne	0.682	0.811
Écart-type	0.1441	0.112

Lorsqu'on compare les déséquilibres significatifs des populations à filiation cognatique, 27% des arbres (soit 3 populations sur 11) sont plus déséquilibrés à un seuil de 5% alors que 87.5% des populations à filiation patrilinéaire sont déséquilibrés (soit 7 populations sur 8). Cette proportion de populations à filiation patrilinéaire en déséquilibre est significativement plus élevée que celle des populations à filiation cognatique (Test exact de Fisher, *p-value* = 0.0012, cf. Tableau I.C.ii-1, 2, 3).

Tableau I.C.ii-2 Valeur du déséquilibre, niveau de significativité (*p*) du déséquilibre, nombre de nœuds (*n*) pour 11 populations cognatiques d'Asie centrale. Les *p-values* (p) en gras sont les valeurs significatives à un seuil de 5%.

Population (Aire géographique)	*Acronyme*	*Déséquilibre* I_{nb}	*p*	*n*
Uzbek (Zarmank)	LUZ	0.5773	0.248	11
Tajik (Penjinkent Chink mnt.)	TDS	0.8522	**0**	9
Tajik (Penjinkent mountains)	TDU	0.6956	0.0562	15
Tajik (Samarkand Agalic)	TJA	0.8511	**0.0066**	8
Tajik (Gharm Nimich)	TJE	0.5135	0.4544	10
Tajik Ferghana (Kaptarana)	TJK	0.5471	0.3346	13
Tajik (Gharm Nimich)	TJN	0.5476	0.3814	10
Tajik (Ferghana Richtan)	TJR	0.5943	0.241	12
Tajik (Gharm Nouchor)	TJT	0.9351	**0**	7
Tajik (Samarkand)	TJU	0.6606	0.0994	13
Tajik Gharm Nouchor (Yagnobs from Douchambe)	TJY	0.7272	0.1008	8

Tableau I.C.ii-3 Valeur du déséquilibre, niveau de significativité (*p*) du déséquilibre, nombre de nœuds (*n*) pour 8 populations patrilinéaires d'Asie centrale. Les *p-values* (p) en gras sont les valeurs significatives à un seuil de 5%.

Population (Aire géographique)	*Acronyme*	*Déséquilibre I_{nb}*	*P*	*n*
Kazakh (Karakalpakia)	KAZ/KZ	0.6319	0.1422	12
Karakalpak (Qongrat from Karakalpakia)	KAR/KKK	0.6802	**0.015**	16
Kirghize (Andijan, sud)	KRA	0.8738	**0**	13
Kirghize (Narin, nord, 1)	KRG	0.8048	**0**	3
Kirghize (Narin, nord, 2)	KRM	0.9837	**0**	7
Kazak Gasli	LKZ	0.8027	**0**	7
Karakalpaks (On Tört Uruw from Karakalpakia)	OTU	0.7247	**0.0348**	10
Turkmen (Karakalpakia)	TUR/TK	0.9452	**0**	9

Analyse de la diversité

Les populations à filiation patrilinéaire ont une moyenne d'hétérozygotie par site par population plus faible que les populations à filiation cognatique que cela soit pour les données microsatellites (Test unilatéral de Wilcoxon, *p-value* = 0.010) ou les données de SNPs (Test unilatéral de Wilcoxon, p = 0.025) (Tab I.C.ii-4).

La corrélation entre les valeurs de déséquilibre dans les arbres avec les moyennes d'hétérozygotie par locus est significativement négative pour les valeurs calculées sur les microsatellites (Test bilatéral de corrélation de rangs de Spearman, r = -0.53, p = 0.020), mais non significative pour les valeurs calculées sur les SNPs (Test bilatéral de corrélation de rangs de Spearman, r = -0.22, *p-value* = 0.361 ; cf. Tableau I.C.ii.S2). Par ailleurs, lorsque l'analyse est restreinte aux huit populations à filiation patrilinéaire, la valeur de la corrélation *r* entre l'hétérozygotie des microsatellites et le déséquilibre est plus forte que dans le cas où toutes les populations sont considérées, mais elle devient non significative (r = -0.57, *p-value* = 0.15). En ce qui concerne les populations à filiation cognatique, la corrélation entre la moyenne des microsatellites et le déséquilibre calculée sur l'arbre de phylogénie du Y est de -0.32 (non significatif). Aussi, la corrélation entre l'hétérozygotie moyen des SNPs et le déséquilibre des arbres pour les populations cognatiques est négative, non significative et faible (r=-0.07) ; il en est de même pour les populations patrilinéaires.

Tableau I.C.ii-4 Moyenne du nombre de SNPs et des microsatellites par population, des hétérozygoties par site polymorphe par population pour les SNPs et les microsatellites pour les deux types de filiation (cognatique et patrilinéaire)

	Nombre moyen de SNPs polymorphes par population	*Hétérozygotie moyenne par site et par population pour les SNPs*	*Nombre moyen de microsatellites polymorphes par population*	*Hétérozygotie moyenne par site et par population pour les microsatellites*
Cognatique	10.6364	0.2768	11.0909	0.5793
Patrilinéaire	12.875	0.2174	11.875	0.4668

I.C.ii.d Discussion

Les données génétiques semblent confirmer que les populations à filiation patrilinéaire sont soumises à un phénomène de transmission du succès reproducteur patrilinéaire que l'on ne retrouverait pas dans les populations à filiation cognatique. En effet, le déséquilibre des arbres reconstruits sur le chromosome Y est plus important dans les populations à filiation patrilinéaire que dans les populations à filiation cognatique. Les différences d'effectif efficace entre hommes et femmes dans les populations patrilinéaires ou entre les hommes des populations cognatiques et patrilinéaires (Chaix et al. 2007) pourraient donc être dues à la structuration sociale dont une des conséquences serait la transmission du succès reproducteur. L'origine de cette transmission pourrait provenir d'une différence de succès reproducteur en fonction du rang social chez les hommes ; par exemple un homme de rang social élevé pourrait pratiquer plus facilement la polygynie. L'hypothèse d'une migration différentielle selon le rang social ou d'un accès différentiel aux terres à l'origine de ce phénomène est probable seulement si les migrations se font préférentiellement entre groupes ethniques et non au sein des groupes ethniques (Karakalpak, Kazakh, Kirgyz, Turkmen) puisqu'il a été montré une plus forte différenciation sur les chromosomes Y au sein des groupes ethniques qu'entre ces mêmes groupes dans les populations pastorales (Heyer et al. 2009).

Le déséquilibre des arbres comme marqueur de la transmission du succès reproducteur a déjà été utilisé par Blum et *al.* (2006) sur des données génétiques de l'ADN mitochondrial pour mettre en évidence une transmission matrilinéaire chez les chasseurs-cueilleurs que l'on ne retrouve pas chez les populations d'agriculteurs. Aussi, comme nous le soulignons dans la partie I.B la présence de déséquilibre dans les arbres est liée à la présence d'hétérogénéité du succès reproducteur. Dans ce cas, il est possible que d'autres phénomènes sociaux ou

génétiques entraînent aussi dans les populations chasseurs-cueilleurs, une disparité du succès reproducteur entre les individus.

II faut cependant noter que Blum et al (2006) et nous-mêmes (partie I.C.i) ont montré les biais liés à la reconstruction phylogénétique des arbres de coalescence, puisque la valeur du déséquilibre obtenue à partir des arbres reconstruits était supérieure au déséquilibre véritable des coalescents de Kingman utilisés pour la simulation des jeux de données de séquences. En effet, 15% à 20% des arbres reconstruits phylogénétiquement étaient significativement déséquilibrés, au lieu des 10% attendus (seuil de la *p-value* des études). Néanmoins, dans notre étude des populations de l'Asie centrale, nous montrons que la différence entre les valeurs de déséquilibre pour les deux types de populations (cognatiques et patrilinéaires) est significative et que sept populations patrilinéaires sur huit présentent un déséquilibre, ce qui est beaucoup plus que les 20% attendus du fait du biais sous une hypothèse neutre.

Aussi, la diversité des microsatellites et des SNPs est réduite dans les populations à filiation patrilinéaire par rapport à celle des populations à filiation cognatique, confirmant les précédentes études (Chaix et al. 2007; Segurel et al. 2008). De plus, nous observons que la valeur du déséquilibre augmente quand la diversité des microsatellites diminue et qu'elle a tendance à être plus forte quand nous ne considérons que les populations patrilinéaires. Cette corrélation est attendue puisque les travaux théoriques ont montré que la transmission du succès reproducteur entraînait à la fois un déséquilibre plus fort des arbres et une diversité réduite (Blum et al. 2006 et partie I.B; Sibert et al. 2002), ce qui conforte bien le signal de transmission du succès reproducteur dans ces populations. Ajoutons qu'il se pourrait qu'il y ait une différence de transmission du succès reproducteur ou des différences d'hétérogénéité du succès reproducteur entre les populations patrilinéaires ; ceci expliquerait le gradient négatif entre la diversité et le déséquilibre au sein des populations à filiation patrilinéaire.

La prochaine étape de ce travail sera de comparer ces données obtenues pour le chromosome Y avec les données de l'ADN mitochondrial des femmes échantillonnées dans les mêmes populations. Dans le cas d'une patrilinéarité stricte, comme nous le montrons partie I.B, nous ne nous attendons à aucun déséquilibre sur les arbres de ces gènes mitochondriaux, mais à une réduction de la diversité par rapport aux populations cognatiques, car nous avons montré que la variance du succès reproducteur des femmes est aussi augmentée dans le cas d'une transmission matrilinéaire. A l'inverse, si dans certaines

populations un fort déséquilibre était aussi observé pour l'ADN mitochondrial, un scénario avec une transmission biparentale du succès reproducteur serait à envisager.

I.C.ii.e Informations Supplémentaires

Tableau I.C.ii.S1, Nom des populations, localisation entre parenthèses, Acronyme, nombre d'individus échantillonnés par population, type de transmission (C pour Cognatique et P pour patrilinéaire), Nombre de SNPs séquencés (polymorphe ou non) et nombre de microsatellites séquencés par population,

Population	*Acronyme*	*Nombre d'individus*	*Type de transmission*	*Nombre de SNPs*	*Nombre de microsatellites*
Uzbek (Zarmank)	LUZ	31	C	13	11
Tajik (Penjinkent Chink mnt.)	TDS	24	C	13	10
Tajik (Penjinkent mountains)	TDU	31	C	18	11
Tajik (Samarkand Agalic)	TJA	32	C	49	12
Tajik (Gharm Nimich)	TJE	27	C	18	11
Tajik Ferghana (Kaptarana)	TJK	30	C	22	12
Tajik (Gharm Nimich)	TJN	30	C	18	11
Tajik (Ferghana Richtan)	TJR	29	C	22	12
Tajik (Gharm Nouchor)	TJT	24	C	13	10
Tajik (Samarkand)	TJU	29	C	49	12
Tajik Gharm Nouchor (Yagnobs from Douchambe)	TJY	25	C	13	10
Kazakh (Karakalpakia)	KAZ/KZ	50	P	33	12
Karakalpak (Qongrat from Karakalpakia)	KAR/KK	54	P	33	12
Kirghize (Andijan, sud)	KRA	46	P	22	12
Kirghize (Narin, nord)	KRG	20	P	22	12
Kirghize (Narin, nord)	KRM	25	P	22	12
Kazak Gasli	LKZ	20	P	18	11
Karakalpaks (On Tört Uruw from Karakalpakia)	OTU	54	P	33	12
Turkmen (Karakalpakia)	TUR/TK	51	P	33	12

Tableau I.C.ii.S2, Nombre de SNPs et de microsatellites, hétérozygotie moyenne par site polymorphe pour les SNPs et les microsatellites pour les populations échantillonnées en Asie centrale.

Population	*Acronyme*	*Nombre de SNPs polymorphes par population*	*Hétérozygotie moyenne par site de type SNP*	*Nombre de microsatellites polymorphes par population*	*Hétérozygotie moyenne par site de type microsatellite*
Uzbek (Zarmank)	LUZ	10	0.236	11	0.6162
Tajik (Penjinkent Chink mnt.)	TDS	8	0.2678	10	0.5428
Tajik (Penjinkent mountains)	TDU	7	0.3015	11	0.5733
Tajik (Samarkand Agalic)	TJA	19	0.2303	12	0.6038
Tajik (Gharm Nimich)	TJE	9	0.3274	11	0.6004
Tajik Ferghana (Kaptarana)	TJK	11	0.1925	12	0.602
Tajik (Gharm Nimich)	TJN	8	0.3547	11	0.5626
Tajik (Ferghana Richtan)	TJR	12	0.2798	12	0.6351
Tajik (Gharm Nouchor)	TJT	9	0.2788	10	0.598
Tajik (Samarkand)	TJU	19	0.223	12	0.586
Tajik Gharm Nouchor (Yagnobs from Douchambe)	TJY	5	0.3533	10	0.4521
Kazakh (Karakalpakia)	KAZ/KZ	15	0.1673	12	0.4216
Karakalpak (Qongrat from Karakalpakia)	KAR/KK	19	0.227	12	0.6025
Kirghize (Andijan, sud)	KRA	11	0.279	12	0.4712
Kirghize (Narin, nord)	KRG	11	0.3614	12	0.5089
Kirghize (Narin, nord)	KRM	5	0.1331	12	0.2813
Kazak Gasli	LKZ	9	0.1706	11	0.4552
Karakalpaks (On Tört Uruw from Karakalpakia)	OTU	19	0.2098	12	0.5937
Turkmen (Karakalpakia)	TUR/TK	14	0.1912	12	0.3998

I.D Impacts de la transmission du succès reproducteur sur les généalogies d'individus. Méthodes de détection.

I.D.i Introduction

Nous avons décrit dans les chapitres précédents l'impact de la transmission du succès reproducteur sur l'histoire des gènes au sein d'une population ainsi que les approches pour détecter ce phénomène en utilisant directement l'information génétique. L'utilisation de la forme du coalescent comme signature spécifique pour détecter la transmission du succès reproducteur a été décrite par Blum et *al.* (2006). D'une manière plus directe et peut-être plus intuitive, la transmission du succès reproducteur peut aussi être détectée en utilisant l'information des généalogies d'individus construites à partir des actes de naissance, de mariage et de décès des individus (Austerlitz & Heyer 1998; Bocquet-Appel & Jakobi 1993; Helgason et al. 2003; Pearson et al. 1899; Pluzhnikov et al. 2007). Néanmoins, contrairement aux données génétiques, les généalogies individuelles sont d'accès difficile voire impossible pour certaines populations, elles sont souvent fragmentaires et ont une profondeur temporelle relativement restreinte. Par ailleurs, il faut noter que l'on ne dispose le plus souvent que de données ascendantes, reconstituées à partir d'individus contemporains. La détection du phénomène de transmission du succès reproducteur au sein de généalogies ascendantes nous a donc obligé à définir un ensemble de nouvelles mesures ou à adapter les mesures existantes (ainsi que les statistiques associées) à ce type de données.

L'étude de la transmission du succès reproducteur sur les généalogies d'individus a commencé avec Pearson et *al.* (1899) qui ont montré qu'au sein de la noblesse anglaise il existait des corrélations positives du succès reproducteur entre les générations et que la corrélation entre les fécondités d'une mère et de sa fille était plus importante que celles observées entre un père et son fils. La corrélation entre les tailles de fratrie et le nombre d'enfants des individus donne une relation directe entre le succès reproducteur d'un individu et celui de ses parents. Cette relation permet de mesurer s'il existe une différence de contribution génétique des individus à la génération suivante et si elle est liée à la taille de leur fratrie. Notons cependant que parmi les descendants d'un individu, tous ne sont pas pertinents du point de vue génétique. De nombreux individus peuvent en effet ne pas participer au pool génétique de la population à la génération suivante soit parce qu'ils ont

migré avant de se reproduire, soit parce qu'ils sont morts à un âge jeune. Pour prendre en compte ce phénomène, Heyer et *al.* (1999) ont proposé de ne compter que les individus « utiles » (c'est-à-dire les individus qui ont au moins un enfant dans la population). Il est donc intéressant de mesurer la corrélation entre parents et enfants du nombre d'enfants utiles, ce qui implique d'observer trois générations successives, pour pouvoir dénombrer les individus s'étant reproduits dans la population. Austerlitz et *al.* (1998) ont montré que dans la population de Saguenay-Lac Saint Jean il existait une corrélation positive entre la taille de fratrie utile et le nombre d'enfants utiles des individus alors que la corrélation entre les tailles totales de fratrie et le nombre total d'enfants était plus faible. La différence entre les deux types de corrélation a été un argument supplémentaire pour affirmer que la transmission du succès reproducteur était culturelle dans ce cas. En revanche Pluzhnikov et *al.* (2007) ont conclu que dans la population Huttérite cette transmission du succès reproducteur était plutôt d'origine génétique, puisque les corrélations mesurées étaient similaires que l'on considère l'ensemble des enfants ou seulement les enfants utiles, et qu'aucun facteur socio-économique ne pouvait expliquer la différence de reproduction entre les individus dans cette population.

La transmission du succès reproducteur a des impacts sur d'autres mesures comme la contribution génétique des individus. Cette mesure permet d'estimer la proportion de gènes autosomaux qu'un individu laisse aux descendants de la population au bout d'un nombre donné de générations. La transmission du succès reproducteur au cours du temps modifie la part contributive génétique entre les individus. Par exemple, dans la population du Québec, la contribution génétique relative des fondateurs mariés avant 1700 est de 78.8 % chez les hommes alors qu'ils ne représentent que 40 % des fondateurs, alors que les fondateurs mariés après 1700 ont une contribution génétique proportionnellement trois fois plus faible (Vézina et al. 2005). Cette variabilité de contribution génétique entre fondateurs a été attribuée à la présence d'un « cœur » et d'une « frange » de la population, les individus qui s'installaient plus tardivement dans la population laissaient moins de descendants que les individus appartenant à des lignées déjà présentes.

La mesure peut être complétée par l'analyse des transmissions des ADN mitochondriaux (c'est-à-dire par le parcours des lignées Mère-Filles) et des chromosomes Y (c'est-à-dire par le parcours des lignées Père-Fils) au sein des populations grâce aux généalogies. Tremblay et Vézina (Tremblay & Vézina) ont montré que la population du Québec était constituée d'un nombre plus important de lignées patrilinéaires que de lignées matrilinéaires (c'est-à-dire plus d'hommes que de femmes parmi les fondateurs initiaux) et

que les lignées matrilinéaires dernières présentaient une contribution génétique moins variable que lignées patrilinéaires. Ces deux phénomènes pourraient s'expliquer par des différences de comportement migratoire entre hommes et femmes. En revanche, dans la population du Saguenay-Lac-Saint-Jean, la distribution des contributions génétiques des fondateurs était équivalente entre génome nucléaire, mitochondrie (Heyer 1995).

D'autres approches ont été utilisées pour analyser l'impact de la structure des généalogies d'individus sur la diversité génétique neutre. Par exemple, la méthode de l'« allèle dropping » (Heyer 2009; MacCluer et al. 1986) utilisée sur des généalogies de la population du Saguenay-Lac-Saint-Jean a montré que la transmission du succès reproducteur augmentait les associations entre allèles (Austerlitz & Heyer 2000). Dans la population de Campora dans la région du Cilento (Italie), la comparaison de la consanguinité et de l'exogamie montrait que plus le pourcentage de mariages exogames était important, plus la consanguinité moyenne de la population augmentait, ceci pouvant être expliqué là-encore par l'existence d'un cœur et d'une frange de population (Colonna et al. 2007). Une autre mesure utilisée était l'apparentement au sein des généalogies qui montraient une corrélation positive entre l'apparentement des conjoints et leur succès reproducteur chez les Islandais (Helgason et al. 2008a; Helgason et al. 2008b; Labouriau & Amorim 2008).

L'ensemble de ces mesures n'a été formalisé que partiellement dans le cadre de généalogies ascendantes et dans celui de la transmission du succès reproducteur pour des généalogies ascendantes. Dans un premier temps nous allons formaliser ces mesures dans ce cadre avec le modèle décrit en partie I.B. Nous allons aussi les comparer entre elles et proposer d'autres mesures qui pourraient être un signe de la transmission du succès reproducteur. Enfin nous confronterons ces résultats à des données de la population du Cilento (Italie).

I.D.ii Approche par l'outil de modélisation

I.D.ii.a Matériels et méthodes

Généalogie

Les généalogies d'individus d'une population consistent en un ensemble de relations de filiation entre les individus. Dans une population, on peut avoir accès à deux types de généalogies : les généalogies de type descendantes où tous les individus ayant vécu dans la

population sont présents, même ceux n'ayant pas laissé de descendants, et les généalogies ascendantes où seuls les ancêtres ayant laissé au moins un individu dans la population sont présents. Nous distinguerons les individus de la génération la plus récente, que nous appellerons descendants, des individus qui n'appartiennent pas à cette génération, que nous appellerons ancêtres. Pour les ancêtres, nous ferons une distinction pour les individus n'ayant pas de parents dans la généalogie et que nous nommerons « fondateurs » (qu'ils proviennent ou non de la population).

Modélisation des généalogies.

Nous modélisons des populations en utilisant le modèle diploïde décrit dans la partie I.B. Nous simulons des populations de taille constante de 500 individus. Pour chaque généalogie, nous récupérons tous les individus de la dernière génération et leurs ancêtres sur 16 générations ; ce sont donc des généalogies ascendantes puisque nous ne considérons que des individus ayant laissé au moins un descendant. Dans ces populations, nous simulons plusieurs types de transmission du succès reproducteur : patrilinéaire, matrilinéaire ou biparentale. Nous faisons également varier l'hétérogénéité du succès reproducteur entre les individus (a=1, a= ∞) et le niveau de transmission du succès reproducteur (α=0, α=1, α=1.4). Pour chaque jeu de paramètres, nous répétons 500 fois le scénario sauf indication contraire.

Mesure de la contribution génétique

La contribution génétique (*gc*) d'un individu est la part du génome autosomal des individus de la génération actuelle qui a été hérité de l'individu considéré (Heyer & Tremblay 1995; O'Brien et al. 1988). Pour calculer la contribution génétique d'un individu *i* donné, nous considérons les *p* descendants de cet individu à la génération actuelle. Pour chacun de ces descendants *j*, nous considérerons le nombre n_c de générations qui séparent l'individu *i* de son descendant *j* pour chacun des *c* chemins généalogiques liant l'ancêtre *i* au descendant *j*. La contribution génétique totale gc_i de l'individu *i* se calcule ainsi :

$$gc_i = \sum_p \sum_c \frac{1}{2}^{n_c} \quad \text{Eq. I.D-a}$$

Analyse des données démographiques

Nous calculons plusieurs types de corrélation : les corrélations entre le nombre d'enfants et le nombre de frères et sœurs de tous les individus, les corrélations entre le nombre de frères et le nombre de fils des hommes (Père-Fils) et les corrélations entre le nombre de sœurs et le nombre de filles des femmes (Mère-Fille). Nous analyserons ces valeurs dans les généalogies ascendantes, et nous comparerons ces valeurs à des généalogies descendantes pour les généalogies simulées.

Analyse des lignées

Pour chaque généalogie ascendante, nous avons extrait les arbres des lignées uniparentales. En suivant, à partir de chaque fondateur, la lignée continue mère-fille ou la lignée continue père-fils, on extrait les arbres correspondant à la transmission des gènes mitochondriaux ou à celle des gènes du chromosome Y. A partir de l'ensemble des lignées Y extraites de la généalogie (une par fondateur ayant laissé son chromosome Y), nous calculons sur chaque nœud le déséquilibre à partir de la formule I.B-b et calculons la moyenne du déséquilibre sur l'ensemble des nœuds. Le même calcul sera fait sur l'ensemble des lignées mère-fille. De plus, nous regarderons l'influence sur le déséquilibre de la hauteur minimale des nœuds (nombre de générations moyennes entre l'ancêtre et ses descendants) utilisée pour calculer la moyenne du déséquilibre dans la généalogie.

I.D.ii.b Résultats

Analyse démographique

Analyse des corrélations intergénérationnelles sur la taille des fratries entre généalogies ascendantes et généalogies descendantes dans des généalogies simulées.

Nous avons comparé les corrélations moyennes entre tailles de fratrie des parents et des enfants, suivant que ces corrélations sont calculées sur les généalogies descendantes ou sur les généalogies ascendantes correspondantes. Nous faisons varier l'intensité de la transmission du succès reproducteur pour différents scénarios de transmission (biparentale, patrilinéaire et matrilinéaire) avec plus ou moins d'hétérogénéité du succès reproducteur. Nous observons (Figure I.D.i-1) que les valeurs moyennes des deux types de corrélation sont hautement corrélées (r=0.99) et que les valeurs moyennes des corrélations obtenues sur des généalogies ascendantes sont environ 27% inférieures à celles observées sur des généalogies descendantes. Les mêmes résultats sont observés pour les corrélations entre nombre de frères d'un homme et de son nombre de fils, comme pour les corrélations entre nombre de sœurs d'une femme et de son nombre de filles (résultats non montrés).

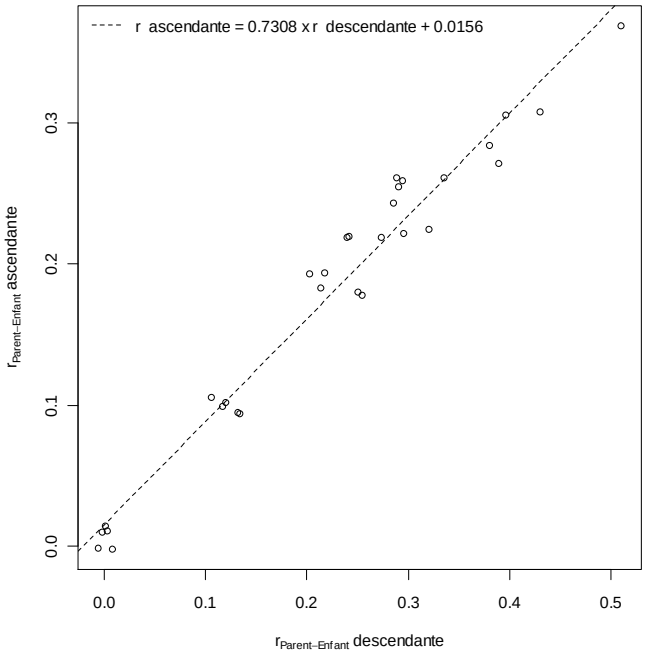
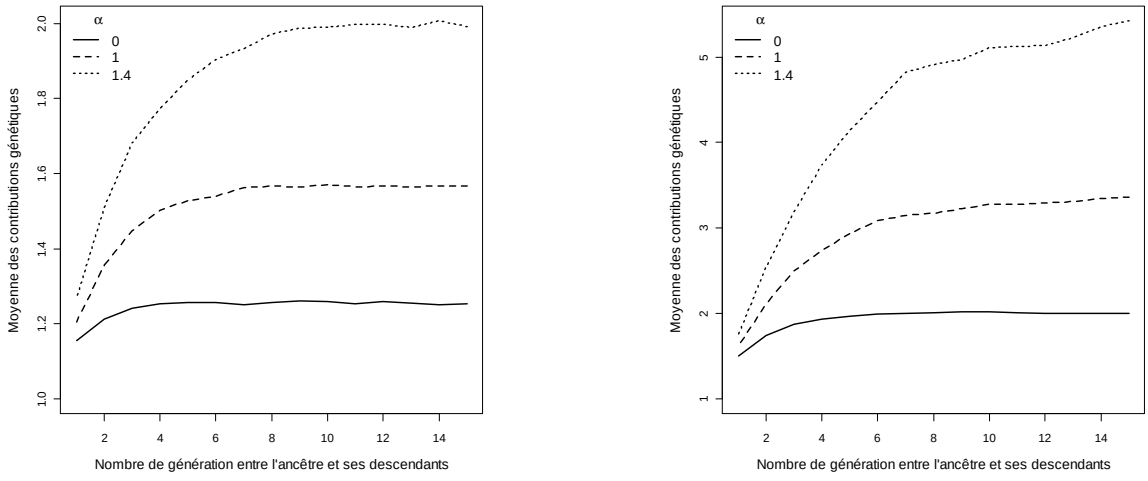


Figure I.D.i-1: Comparaison des moyennes des corrélations entre la taille de fratrie d'un enfant et celle d'un parent entre des généalogies ascendantes et descendantes, pour plusieurs valeur de α (0, 0.4, 0.8, 1.0, 1.4, 1.8), plusieurs valeurs d'hétérogénéité (a=1, a=∞) pour un scénario biparental, patrilinéaire et matrilinéaire. Chaque point est une moyenne sur 100 simulations. La droite de régression linéaire dont l'équation est précisée dans la légende est figurée en pointillés.

Contribution génétique

Moyennes des contributions

Nous analysons la contribution génétique moyenne des ancêtres de la population actuelle en fonction du nombre de générations qui les séparent de la génération actuelle, pour différents niveaux de transmission du succès reproducteur et différents niveaux d'hétérogénéité. On observe que la contribution des ancêtres augmente avec le nombre de générations qui les séparent de la génération actuelle, jusqu'à atteindre un plateau. L'augmentation est d'autant plus élevée et l'atteinte du plateau tardive que la transmission du succès reproducteur et l'hétérogénéité de celui-ci sont élevées. En moyenne, sans transmission et sans hétérogénéité du succès reproducteur, lorsque le nombre de générations est supérieur à 4 on observe une contribution moyenne des ancêtres lointains de 1.25 par individu (courbe $\alpha = 0$, $a = \infty$, Figure I.D.i-2a) qui peut augmenter jusqu'à 2 lorsqu'il y a une forte hétérogénéité (courbe $\alpha = 0$, $a = 1$, Figure I.D.i-2a.). La transmission du succès reproducteur augmente la contribution génétique de ces ancêtres qui peut dépasser 2 pour un fort niveau de transmission mais une faible hétérogénéité ($\alpha = 1.4$, $a = 1$, Figure I.D.i-2a), et monter jusqu'à 5 pour un très fort niveau de transmission associé à une forte hétérogénéité ($\alpha=1.4$, $a = \infty$, Figure I.D.i-2b). Au final, plus la transmission et l'hétérogénéité du succès reproducteur sont élevées, plus la contribution des ancêtres lointains est élevée par rapport à celle des ancêtres récents.

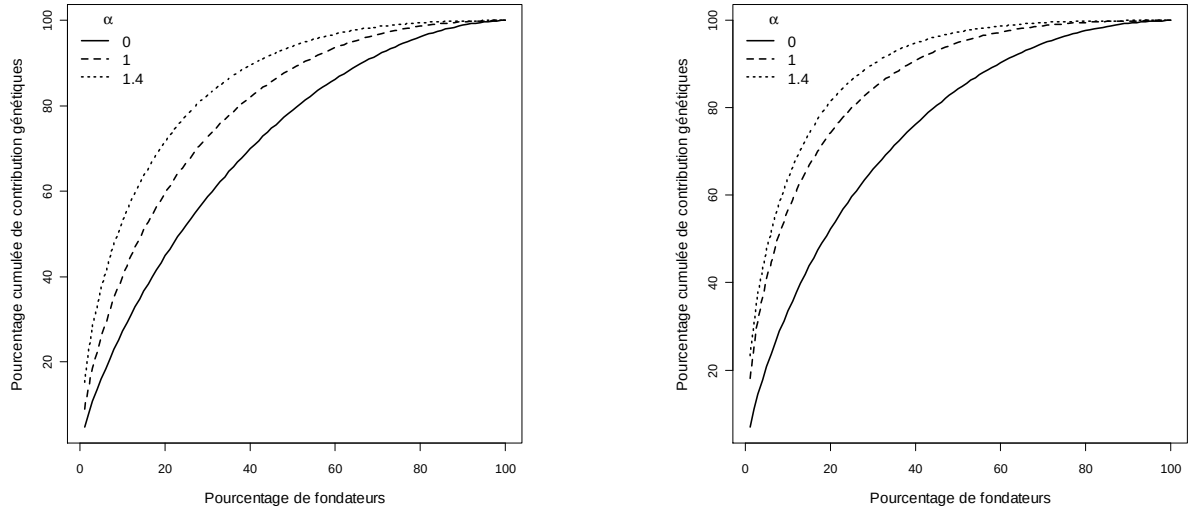


A *B*

Figure I.D.i-2 : Contribution génétique moyenne des individus à la génération actuelle en fonction du nombre de générations qui sépare l'ancêtre aux descendants, sans hétérogénéité du succès reproducteur ($a=\infty$) (a) ou avec hétérogénéité ($a=1$) (b) et pour différents taux de transmission biparentale du succès reproducteur ($\alpha=0$, 1, 1.4). Toutes les courbes correspondent à des moyennes calculées sur 100 simulations indépendantes.

Pourcentages des contributions

L'analyse du pourcentage cumulé de la contribution génétique des fondateurs montre que l'hétérogénéité du succès reproducteur et la transmission du succès reproducteur modifient sensiblement la distribution des contributions génétiques entre les fondateurs. Par exemple, sans transmission et hétérogénéité du succès reproducteur, 10% des plus gros contributeurs représentent 27% de la contribution génétique (cf. Figure I.D.i-3a) alors que si la population est soumise à une forte hétérogénéité, cette valeur augmente à 33% (cf. Figure I.D-3b). De même, lorsque la population est soumise à une transmission modérée du succès reproducteur ($\alpha = 1$) mais sans hétérogénéité supplémentaire ($a = \infty$), 10% des fondateurs représentent 40% de la contribution génétique totale, ce pourcentage grimpant même à 52% dans le cas d'un niveau de transmission du succès reproducteur plus élevé (α=1.4, Figure I.D.i-3a). Si la population est de plus soumise à une forte hétérogénéité du succès reproducteur ($a = 1$), les valeurs sont respectivement de 56 % et 63 %.



a *b*

Figure I.D.i-3 : Pourcentage cumulée de la contribution génétique rangée par ordre décroissant à la génération actuelle pour les 500 fondateurs, avec de l'hétérogénéité ajoutée (a=1) (b) ou sans hétérogénéité (a=∞) (a) et en fonction de différents taux de transmission du succès reproducteur (α=0, 1, 1.4), chaque simulation est répétée 100 fois.

Analyse du déséquilibre des arbres généalogiques

Nous avons extrait les généalogies des 500 individus de la population simulée sur 16 générations, récupéré les lignées mitochondriales et Y dans chacune de nos généalogies et calculé le déséquilibre dans chaque lignée. Nous observons que sans hétérogénéité et sans transmission du succès reproducteur ($\alpha = 0$, $a = 1$), nos généalogies sont plus équilibrées qu'attendues dans un arbre de Kingman (la valeur attendue est de 0.5) (Tableau I.D.i-1). Au

contraire, avec hétérogénéité mais toujours sans transmission, le déséquilibre est celui observée dans un arbre de Kingman.

Tableau I.D.i-1 Déséquilibre calculé dans des lignées père-fils (Y) et mère-filles extraites de généalogies avec de l'hétérogénéité (a=1) ou sans (a=∞) et sans transmission du succès reproducteur.

	Sans Hétérogénéité	*Avec Hétérogénéité*
Y	0.4284	0.5003
M	0.4324	0.5045

Lorsqu'un niveau moyen de transmission du succès reproducteur est ajouté dans les populations (α = 1), mais sans hétérogénéité du succès reproducteur (a = ∞), la valeur du déséquilibre est augmentée mais reste inférieure à celle attendue dans un cadre neutre sur les arbres (cf. Tableau I.D.i-1). Elle est par contre supérieure à 0.5 dans le cas où il y existe à la fois transmission et hétérogénéité du succès reproducteur (α=1.0, a = 1). Nous avons montré sur les arbres de gènes que lors d'une transmission uniparentale du succès reproducteur, le compartiment qui est associé avec le type de transmission (la lignée Y pour une transmission patrilinéaire et la lignée mitochondriale pour une transmission matrilinéaire) est plus déséquilibré qu'attendu alors que le compartiment qui n'est pas associé ne l'est pas du tout (cf partie I.B). Ici, nous observons au contraire que la lignée associée au chromosome Y dans un scénario matrilinéaire et la lignée associée à la mitochondrie dans un scénario patrilinéaire ont des valeurs supérieures à la valeur de 0.5 (Tableau I.D.i-2), lorsqu'il y a une forte hétérogénéité et une transmission du succès reproducteur, même si le déséquilibre est tout de même plus élevé à chaque fois dans la lignée qui est associée à la transmission.

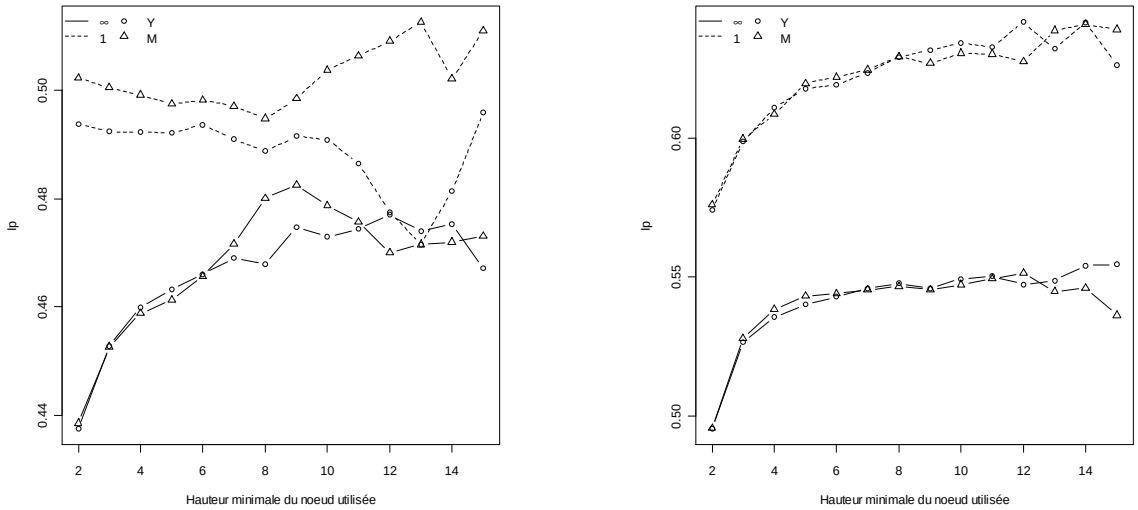
Tableau I.D.i-2 Déséquilibre calculé dans des généalogies ascendantes de 500 individus avec plus ou moins d'hétérogénéité et avec une transmission du succès reproducteur moyenne (α=1.0).

	Sans hétérogénéité			*Avec hétérogénéité*		
	Biparental	*Matrilinéaire*	*Patrilinéaire*	*Biparental*	*Matrilinéaire*	*Patrilinéaire*
M	0.471	0.5183	0.4668	0.5571	0.5845	0.5474
Y	0.4838	0.4669	0.5039	0.5664	0.5552	0.5915

Analyse du déséquilibre en fonction de la hauteur du nœud

Nous analysons le déséquilibre des arbres en fonction de la hauteur minimale du nœud utilisé pour calculer le déséquilibre. Lorsqu'il n'y a ni hétérogénéité du succès reproducteur (a = ∞), ni transmission du succès reproducteur (α = 0), nous observons que la valeur du déséquilibre augmente avec cette hauteur minimale, pour atteindre une valeur stable de 0.48 lorsque le nombre de générations entre la feuille et le nœud se situe entre 7 et 8 (Figure I.D-4a), donc légèrement en dessous de la valeur de 0.5 observée dans un arbre de Kingman. Lorsque les individus sont soumis à une forte transmission (a = 1), la valeur du déséquilibre

reste approximativement stable autour de 0.5. Lorsque la transmission du succès reproducteur est ajoutée, avec ou sans hétérogénéité, plus la hauteur minimale utilisée est haute plus le déséquilibre mesuré dans l'arbre va être important (Figure I.D.i-4b).



a b

Figure I.D.i-4, Moyenne des déséquilibres (I_p), en fonction de la hauteur du nœud calculée sur des lignées de généalogie Mère-Fille (M) et Père-Fils (Y) continues. Les généalogies sont constituées de 500 individus a) non soumis à de la transmission du succès reproducteur ou b) avec de la transmission du succès reproducteur (α=1) avec une transmission bi parentale, pour deux valeurs d'hétérogénéité (a=1 et a=∞)

I.D.ii.c Conclusion / Discussion.

Nous avons analysé différentes mesures sensibles à la transmission du succès reproducteur dans les populations et nous montrons que ces valeurs peuvent dépendre du type de données qui sont analysées. La mesure la plus directe consiste à mesurer la corrélation intergénérationnelle des tailles de fratrie dans les généalogies ascendantes qui incluent tous les individus de la population. Cependant on ne dispose en général que de généalogies ascendantes où seuls les individus qui ont des descendants dans la population actuelle sont présents. Dans ce cadre, nous avons montré que les corrélations entre le succès reproducteur des parents et celui des enfants augmentent avec le niveau de la transmission du succès reproducteur (α), que l'on mesure cette corrélation dans les généalogies ascendantes ou descendantes. Cependant, l'utilisation de données ascendantes plutôt que descendantes amène à des corrélations environ 30% plus faibles. La diminution du signal vient probablement du fait que dans les généalogies ascendantes, tous les individus qui n'ont pas laissés de descendants ne sont pas pris en compte dans le calcul.

Une deuxième façon de détecter la transmission du succès reproducteur consiste à analyser la répartition de la contribution génétique pour les différents fondateurs calculée à

partir des généalogies. La contribution génétique a l'avantage de ne pas nécessité des généalogies descendantes. Nous nous sommes intéressés à la distribution de cette contribution génétique entre les différentes générations et à l'hétérogénéité de cette contribution entre les ancêtres au sein d'une même génération, en faisant varier le niveau de transmission du succès reproducteur et l'hétérogénéité de ce succès reproducteur entre les individus. Comme dans le cas des arbres de coalescence, nous observons une interaction entre les deux facteurs. L'augmentation de l'un comme de l'autre entraîne un déséquilibre entre les contributions génétiques des différentes générations, les générations les plus anciennes ayant la contribution la plus élevée. De plus, elles entraînent toutes les deux un déséquilibre plus important des contributions entre les individus au sein d'une même génération. L'effet cumulatif de ces deux paramètres entraînant les déséquilibres les plus élevés.

La variation de la contribution génétique semble donc fortement dépendante des paramètres démographiques de la population. La somme des contributions génétiques des individus à une génération donnée est égale au nombre de descendants de la population, si la généalogie est complète. Par le phénomène de dérive, un certain nombre d'individus ne vont pas se reproduire. Derrida et *al.* (2000) ont montré que dans une population de Wright-Fisher sans hétérogénéité et sans transmission du succès reproducteur, le nombre d'individus qui laissait un descendant dans la population à partir d'une génération suffisamment ancienne (de l'ordre du logarithme de la taille de population) représentait autour de 80% du nombre d'individus présents à cette génération. Dans le calcul de la contribution génétique, nous n'avons accès qu'aux individus qui ont au moins un descendant à la génération actuelle, et nous observons une stabilisation de la contribution génétique moyenne par individu autour de 1.25 (s'il n'y a pas d'hétérogénéité et de transmission du succès reproducteur, cf. figure I.D.i-2.a). Une contribution moyenne de 1.25 correspond à cette valeur de 80 % d'individus (80 % de 500 individus vont se répartir la contribution des 500 descendants, $gc_{mean} = N_{descendants}/(0.8*N_{Ancêtres})=1.25$). Lorsque les populations sont soumises à une forte hétérogénéité, nous observons que la contribution génétique moyenne est de 2, avec 50% des individus qui contribuent à la génération actuelle. De même, pour une transmission moyenne du succès reproducteur, seulement 50% des individus vont participer à la contribution génétique des individus actuels (cf. Tableau I.D.ii-S1). Nous voyons donc que cette valeur de 80% trouvée par Derrida et *al.* (2000) se trouve fortement diminuée en présence de transmission et d'hétérogénéité du succès reproducteur, deux phénomènes courants dans les populations humaines et animales (cf. I.B.i).

Enfin nous nous sommes intéressé au déséquilibre des arbres extraits des généalogies. Nos travaux présentés dans la partie I.B ainsi que ceux de Sibert et *al.* (2002) et Blum et *al.* (2006) ont montré que le déséquilibre des arbres de coalescence était un bon moyen de détecter la transmission du succès reproducteur à partir des données génétiques. Nous avons analysé directement le déséquilibre au sein des généalogies, en le calculant au sein des lignées généalogiques de la même façon que nous l'avions calculé sur les arbres de coalescence. Nous avons montré qu'effectivement ce déséquilibre était plus élevé dans une population soumise à la transmission du succès reproducteur, et d'autant plus si ce succès reproducteur est hétérogène entre les individus. Cependant nous avons montré que dans une population de Wright-Fisher, le déséquilibre est inférieur à celui attendu dans les arbres de coalescence dans un même scénario, où la valeur attendue est de 0.5. Ceci est dû à la présence des nœuds les plus bas dans la généalogie qui sont plus équilibrés qu'attendus. En effet, dans les généalogies, par la façon dont elles sont reconstruites, de nombreux individus sont apparentés (de nombreux couples frères-frères, sœurs-sœurs, cousin(e)s-cousin(e)s). Les arbres extraits des généalogies sont plus déséquilibrés qu'attendus puisqu'ils sont éloignés des arbres de Kingman attendus. Nous avons montré qu'une solution pour résoudre ce problème est de ne considérer que les nœuds suffisamment anciens dans l'arbre.

Aussi, nous montrons l'influence de la formation des couples sur le déséquilibre dans les lignées de gènes. En effet, lors d'une transmission du succès reproducteur de type matrilinéaire, nous observons que l'arbre du chromosome Y est plus déséquilibré que dans le cas où il n'y a pas de transmission du succès reproducteur, alors que sans formation de couples stables (c'est-à-dire que pour chaque individu on tire indépendamment son père et sa mère dans la population), la valeur du déséquilibre sur le Y est équivalente à celle observée sans transmission (résultats non présentés). De plus, les coalescents du chromosome Y ne sont pas déséquilibrés dans ce scénario (cf. partie I.B). La présence de déséquilibre au sein des lignées lorsqu'il y a une formation de couple peut être expliquée par des déséquilibres temporaires au sein de la lignée du Y.

Au final, même s'il reste à mieux clarifier l'impact de la transmission du succès reproducteur sur le déséquilibre des lignées, nous montrons que cette mesure ainsi que la mesure des corrélations intergénérationnelles au sein des lignées ascendantes et la mesure de la variation de la contribution génétique au sein et entre les générations sont toutes les trois sensibles à la transmission du succès reproducteur, et que cette sensibilité est accrue par la présence d'une forte hétérogénéité au sein des générations.

I.D.ii.d Informations supplémentaires

Tableau I.D.ii-S1, Pourcentage d'individus participant à la contribution génétique de la population après 16 générations sur des populations soumises à une forte hétérogénéité (a=1) ou sans hétérogénéité ($a=\infty$), sans transmission du succès reproducteur ($\alpha=0$) et avec deux valeurs de transmission du succès reproducteur ($\alpha=1,1.4$)

	Sans hétérogénéité	*Avec hétérogénéité*
0.0	79.5	50.1
1.0	56.8	25.0
1.4	51.2	21.0

I.D.iii Exemple du Cilento, Italie

Etude réalisée en collaboration avec Marina Ciullo (CNR, Naples), Anne-Louise Leutenegger et Catherine Bourgain (INSERM, Paris), les données brutes nous ont été fournies par M. Ciullo.

I.D.iii.a Contexte

Les résultats de la partie I.D.ii montrent qu'il est possible d'utiliser des informations comme le déséquilibre des lignées à transmission uni-parentale, la contribution génétique et/ou les corrélations intergénérationnelles sur les tailles de fratrie comme mesures permettant de détecter une transmission du succès reproducteur. Nous allons étudier ici ces différentes mesures sur les généalogies du Cilento pour observer si ces populations ont été soumises à des événements culturels et nous comparerons ces mesures aux résultats obtenus sur les généalogies reconstruites à partir de données de polymorphisme de séquence de l'ADN mitochondrial. Pour cela nous avons analysé le déséquilibre des arbres reconstruits par phylogénie et nous avons effectué des tests de déviation de ces séquences à la neutralité, car un déséquilibre des arbres et une déviation significative à la neutralité sont tous les deux indicateurs de la présence du succès reproducteur (Blum et al. 2006; Sibert et al. 2002)

Les généalogies étudiées sont issues de trois villages se situant au sein du Parc national du Cilento en Italie : Gioi, Cardile et Campora. Ces villages sont étudiés dans le cadre du projet « Genetic Park of Cilento and Vallo di Diano Project » dont le but est de mettre à profit leur statut d'isolats génétiques pour des études d'association entre gènes et maladies complexes comme l'hypertension, le diabète, l'obésité, le cancer ou les maladies neuro-dégénératives. Ce projet a d'ores et déjà donné lieu à la publication de nombreux articles en ce qui concerne notamment les facteurs génétiques impliqués dans la croissance vasculaire endothéliale (Ruggiero et al. 2011) ou les locus impliqués dans le comportement des fumeurs (Sorice et al.). A coté de ces études d'association génétique, une caractérisation précise de la structuration de ces villages a été entreprise. Par exemple, il a été montré que bien qu'ayant un fort passé commun (cf. « Histoire du Cilento»), les deux villages de Gioi et Cardile présentent une forte structuration génétique (Colonna et al. 2009). Sur la base de données génétiques, Colonna et *al.* ont montré par ailleurs que les nombres de lignées mitochondriales et de lignées du chromosome Y sont faibles dans le village de Campora, observation qui pourrait être dû à une réduction de la taille de la population lors d'épisodes de famine ou d'épidémie de peste (cf. contexte historique ci-dessous) (Colonna et al. 2007).

Les différents impacts observés sur la structure génétique du village de Campora pourraient cependant ne pas être dus qu'à des phénomènes démographiques simples. Par exemple, l'augmentation de la consanguinité avec l'augmentation des mariages exogames à Campora (Colonna et al. 2007) comme la forte structuration génétique entre les deux villages mitoyens de Gioi et Campora (Colonna et al. 2009) pourraient relever de phénomènes culturels. Nous allons chercher à déterminer à partir des généalogies et des séquences génétiques de l'ADN mitochondrial ce qui peut être expliqué par une structuration sociale, comme de l'hétérogénéité ou de la transmission du succès reproducteur.

I.D.iii.b Histoire du Cilento

Le Parc national du Cilento est une région italienne située au sud de la Province de Salerne (cf. Figure I.D.iii-1) et de Naples. Il s'agit d'une zone montagneuse entrecoupée de vallons. Les données historiques font état d'une alternance de périodes d'isolement et de périodes d'échanges avec les régions voisines. Initialement la région du Cilento a été occupée par les Grecs au VIII^{e} siècle avant notre ère. L'intérieur du territoire fut conquis par les Lucaniens (peuple originaire de la région Basilicate en Italie) au V^{e} siècle avant J.C. et reconquis par les Grecs par la suite. Au IIIe siècle, le territoire est intégré à Rome mais reste relativement isolé jusqu'au Moyen-âge. Des moines s'installent et exploitent le bord de mer de la région dès le VIII^{e} siècle de notre ère, mais à cause d'une série d'attaques des Sarrasins, ils sont obligés de se déplacer plus à l'intérieur des terres pour échapper à ces invasions. Les premières traces d'occupation du village de Campora datent de la période Lucanienne, mais aucune information n'est disponible avant le XI^{e} siècle. Il est probable que la population originelle ait été constituée d'un mélange de Grecs et de Lucaniens qui étaient employés pour l'agriculture par les moines. Le village fut frappé par une famine au XVI^{e} siècle, suivie par une épidémie de peste au XVII^{e} siècle et resta isolé jusqu'à la fin de la Seconde Guerre mondiale (Colonna

Villages de l'étude.
Carton : Italie avec Campanie et région de Salerno.
Fonds de carte : Google Maps, 2011. Carton : Wikicommons.

Figure I.D.iii-1, Cartes représentant la localisation de la région de Salerne en Italie et des trois villages du Cilento : Cardile, Campora et Gioi.

et al. 2007). Quant aux villages de Gioi et de Cardile, leurs histoires sont intimement liées puisque ce dernier dérive du premier. Les premières traces de Gioi remontent au IX^{e} siècle de notre ère, date à laquelle s'installent des Grecs byzantins et des moines latins. C'est au XI^{e} siècle que des habitants de Gioi fondent le village de Cardile à 6 km de Gioi. Au XVII^{e} siècle les deux villages subirent une famine et restèrent isolés jusqu'à la moitié du XX^{e} siècle (Colonna et al. 2009).

I.D.iii.c Matériels et méthodes

Les données généalogiques

Pour les trois villages, les données généalogiques utilisées ici sont de type ascendant et sont des sous échantillons de celles décrites par Colonna et *al.* (2007; 2009), (cf. Tableau I.D.iii-1 et I.D.iii-2). Les dates de naissance des individus pour les trois populations se répartissent entre le XVI^{e} et le XXI^{e} siècle. Les fondateurs, c'est-à-dire les individus dont l'information sur les parents est inexistante, se repartissent sur toutes les périodes. Néanmoins, nous ne savons pas si ces individus sans parents correspondent à de nouveaux arrivants ou bien s'il s'agit en fait de données manquantes. Les précédentes études supposaient cependant que l'information généalogique au sein de chaque population était complète (Colonna et al. 2007). La profondeur généalogique des données est respectivement de 14, 13 et 12 générations pour les villages de Campora, Cardile et Gioi.

Tableau I.D.iii-1, Nombre d'individus présents dans les données généalogiques ascendantes des trois villages étudiés (Campora, Gioi et Cardile)

	Campora	*Gioi*	*Cardile*
Nombre d'hommes	1526	2063	1171
Nombre de femmes	1535	2096	1199
Nombre total d'individus	3061	4159	2370

Tableau I.D.iii-2, Nombre d'individus actuels dans les données généalogiques ascendantes des trois villages étudiés (Campora, Gioi et Cardile)

	Campora	*Gioi*	*Cardile*
Nombre d'hommes	174	213	112
Nombre de femmes	161	238	136
Nombre total d'individus	335	451	248

La proximité géographique des deux villages de Gioi et de Cardile ainsi que la liaison historique entre leurs développements respectifs se retrouvent dans l'analyse des entrées communes entre les deux ensembles de données généalogiques. Au total, 31% des entrées de

la table des individus de Gioi se retrouvent dans celle de Cardile et inversement 55% des entrées de la table de individus de Cardile se retrouvent dans celle de Gioi. Ce chevauchement des généalogies des deux villages est par ailleurs assez ancien puisque 90% des individus en commun sont nés avant 1850. Ce chevauchement est également d'autant plus important que l'on considère une période ancienne. A titre d'exemple, les individus les plus anciens du corpus généalogique de Cardile sont presque tous présents dans celui de Gioi (Figure I.D.iii-2).

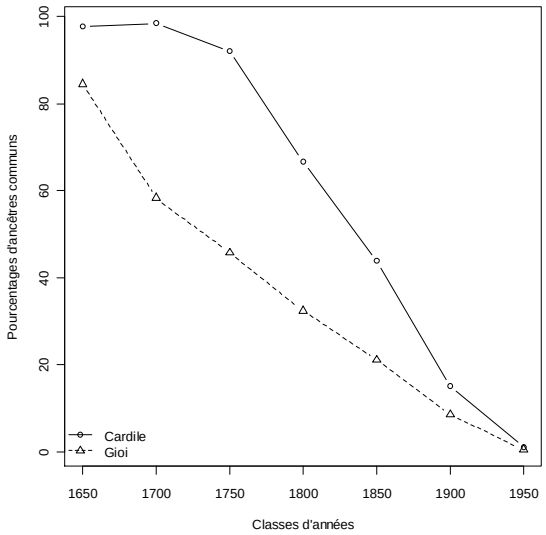


Figure I.D.iii-2: Pourcentage d'entrées en commun au cours du temps pour les deux généalogies du Cilento : Gioi (ligne en pointillés et triangles) et Cardile (ligne pleine et points)

Sur les généalogies des trois villages, nous avons analysé la contribution génétique des ancêtres, les corrélations intergénérationnelles entre les tailles de fratrie et le déséquilibre des généalogies de gènes pour les deux lignées uniparentales (matrilinéaires et patrilinéaires), selon les méthodes développées dans la partie I.D.ii. Pour la contribution génétique et les corrélations, nous avons aussi regardé si ces valeurs changeaient au cours du temps. Pour les corrélations, nous l'avons fait en découpant le pool des ancêtres en tranches de cinquante ans selon leur année de naissance. Pour les contributions génétiques, nous avons étudié comment la contribution génétique moyenne d'un ancêtre dépendait du nombre de générations qui le séparaient de la génération actuelle. Nous avons aussi regardé les différences entre hommes et femmes pour ces mesures.

Les données génétiques

Nous avons analysé les séquences de la région hyper-variable HVRI de l'ADN mitochondrial dans les populations de Gioi et de Campora : 45 séquences HVRI échantillonnées à Campora et 29 séquences échantillonnées à Gioi qui sont constituées de 312

et 363 bases respectivement. Chaque séquence appartient à une lignée généalogique Mère-Fille différente. Pour reconstruire les arbres généalogiques de ces séquences, nous utilisons une méthode de Neighbor-Joining avec un modèle de Jukes et Cantor (Jukes et al. 1969). Nous avons ensuite testé le déséquilibre de ces arbres selon la méthode décrite dans la partie I.B.iii

Par ailleurs, Sibert et *al.*(2002) ayant montré que la transmission du succès reproducteur entraînait des tests de déviation à la neutralité significatifs à la neutralité, nous avons effectué une série de ces tests de déviation à la neutralité sur ces données : D2* (Fu & Li 1993), D (Tajima 1989), et F / F* (Fu & Li 1993). Ceci a été réalisé avec l'interface web de Guillaume Achaz [3].

I.D.iii.d Résultats

La démographie

Corrélation entre les tailles de fratrie des parents et des enfants

L'analyse des corrélations intergénérationnelles des tailles de fratrie montre une variabilité selon la population considérée mais également selon l'époque considérée. Sur la totalité de la période temporelle, nous observons une corrélation positive entre le nombre d'enfants d'un individu et la taille de sa fratrie pour les villages de Cardile et de Gioi (Tableau I.D.iii-3) Dans le détail, cette corrélation reste significative, positive et relativement constante du XVIII^{e} siècle jusqu'à la période contemporaine pour la population de Gioi (I.D.iii-4), alors que pour le village de Cardile, la corrélation est plus variable. En ce qui concerne Campora, même si la valeur globale n'est pas significative, on observe que deux valeurs au XVIII^{e} et au XIX^{e} siècle sont significativement positives (Tableau I.D.iii-6). De plus dans la population de Gioi, on observe une corrélation significative entre le nombre de sœurs d'une femme et son nombre de filles, cette corrélation restant relativement constante au cours du temps (Tableau I.D.iii-3,4). Au contraire, pour le village de Cardile, on observe une corrélation significativement positive entre le nombre de frères d'un homme et son nombre de fils, mais avec de fortes fluctuations au cours du temps, puisque cette corrélation est significative pour seulement deux périodes sur cinq (Tableau I.D.iii-3,5).

[3] http://wwwabi.snv.jussieu.fr/achaz/neutralitytest.html

Tableau I.D.iii-3 Corrélations entre le nombre d'enfants d'un individu et la taille de sa fratrie (Parent-Enfants), le nombre de frères d'un homme et son nombre de fils (Frères-Fils) et le nombre de sœurs d'une femme et son nombre de filles (Sœurs-Filles). *p-value* : *** < 0.001, ** < 0.01, *< 0.05. Seuls les individus nés avant 1960 sont considérés.

	Campora	*Gioi*	*Cardile*
Parent-Enfants	0.0333	0.1395***	0.1866***
Frères-Fils	-0.0017	0.043	0.0829*
Sœurs-Filles	-0.0238	0.1343***	0.0327

Tableau I.D.iii-4, Corrélations entre le nombre d'enfants d'un individu et la taille de sa fratrie (Parent-Enfants), le nombre de frères d'un homme et son nombre de fils (Frères-Fils) et nombre de sœurs d'une femme et son nombre de filles (Sœurs-Filles), par périodes de 50 ans, pour le village de Gioi. *p-value* : *** < 0.001, ** < 0.01, *< 0.05. Seuls les individus nés avant 1960 sont considérés.

	<1700	*1700-1750*	*1750-1800*	*1800-1850*	*1850-1900*	*>1900*
Parent-Enfant	0.0004	0.1141*	0.1835***	0.0847*	0.226***	0.1189**
Frères-Fils	-0.0694	0.0049	0.2053**	0.0383	0.0731	0.0148
Sœurs-Filles	0.0863	0.179*	0.1207	0.0794	0.1482*	0.1449*

Tableau I.D.iii-5, Corrélations entre le nombre d'enfants d'un individu et la taille de sa fratrie (Parent-Enfants), le nombre de frères d'un homme et son nombre de fils (Frères-Fils) et nombre de sœurs d'une femme et son nombre de filles (Sœurs-Filles). *p-value* : *** < 0.001, ** < 0.01, *< 0.05. Seuls les individus nés avant 1960 sont considérés.

	<1700	*1700-1750*	*1750-1800*	*1800-1850*	*1850-1900*	*>1900*
Parent-Enfant	-0.0557	-0.0257	0.0787	0.2178***	0.0422	0.1261*
Frères-Fils	-0.162	-0.0346	0.224*	-0.0643	-0.0912	0.2144**
Sœurs-Filles	0.0416	0.1077	-0.1136	0.1591	-0.0719	-0.1447

Tableau I.D.iii-6, Corrélations entre le nombre d'enfants d'un individu et la taille de sa fratrie (Parent-Enfants), le nombre de frères d'un homme et son nombre de fils (Frères-Fils) et nombre de sœurs d'une femme et son nombre de filles (Sœurs-Filles). *p-value* : *** < 0.001, ** < 0.01, *< 0.05. Seuls les individus nés avant 1960 sont considérés.

	<1700	*1700-1750*	*1750-1800*	*1800-1850*	*1850-1900*	*>1900*
Parent-Enfant	-0.0698	0.1059	0.1378**	0.1021*	0.0017	-0.0111
Frères-Fils	-0.0111	0.1754*	-0.029	0.096	-0.088	-0.0383
Sœurs-Filles	-0.0223	0.0446	-0.0148	0.0915	-0.1432*	-0.0174

Moyennes et variances du nombre d'enfants

Notons par ailleurs que la moyenne du nombre d'enfants par femme est significativement différente entre les trois populations (test t de Student seuil de 5%), mais reste autour de 1.7 (Tableau I.D.iii-7). On observe qu'entre Cardile et Gioi la différence est assez faible (test t de Student *p-value* = 0.022). Au sein de la population de Campora, le nombre moyen d'enfants est supérieur à celui des deux autres populations. Dans les populations de Gioi et de Cardile, le nombre moyen de fils est significativement inférieur au nombre moyen de filles alors que cette différence n'est pas significative pour la population de Campora. Dans toutes les populations, la variance du nombre de filles est significativement supérieure à la variance du nombre de fils (test bilatéral Fligner-Killeen seuil de 5%).

Tableau I.D.iii-7, Moyennes et Variances entre parenthèse du nombre de filles, de fils et d'enfants pour les femmes au sein des populations de Campora, Gioi et Cardile.

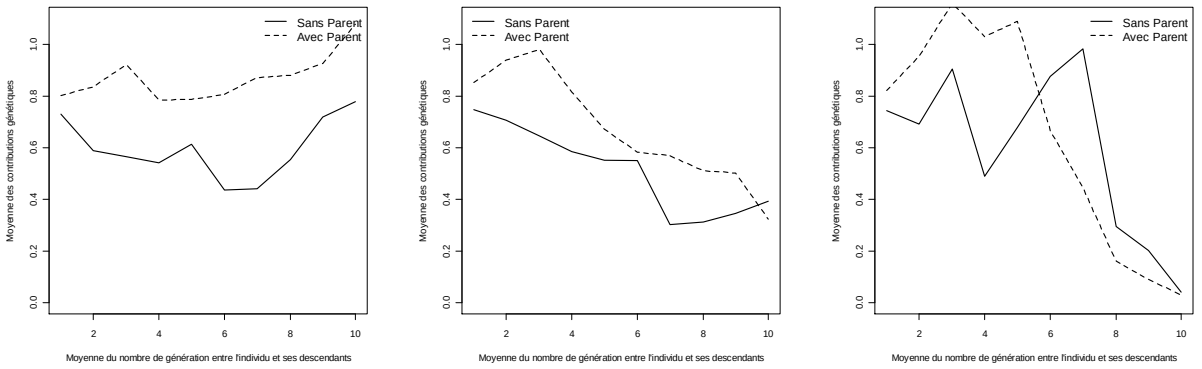
	Campora	*Gioi*	*Cardile*
Filles	0.9234 (0.799)	0.8226 (0.6796)	0.7818 (0.6582)
Fils	0.9327 (0.7005)	0.9035 (0.633)	0.8555 (0.615)
Enfants	1.8561 (1.2013)	1.7262 (1.0187)	1.6372 (0.9673)

Contribution Génétique

Contributions des individus pour les lignées autosomales

Nous comparons la moyenne des contributions génétiques entre un individu qui a au moins un parent dans les généalogies et un individu qui n'a pas de parents (fondateur) dans la généalogie en fonction du nombre moyen de générations qui séparent ces individus de leurs descendants finaux (profondeur généalogique moyenne). Pour les villages de Campora et de Gioi, on observe qu'un individu dont les parents sont référencés dans la base de données généalogique a en moyenne une contribution génétique plus importante qu'un fondateur (Figure I.D.iii-3). Dans le village de Cardile, cette tendance n'est observée que pour les individus de profondeur généalogique moyenne inférieure à 5. Par ailleurs, alors que la contribution génétique moyenne d'un ancêtre non fondateur de Campora est relativement stable quelle que soit sa profondeur généalogique, elle diminue dans les villages de Gioi et Cardile au fur et à mesure qu'on remonte dans le passé. Cette tendance étant plus marquée

pour Cardile que pour Gioi, la contribution moyenne des individus les plus anciens de Cardile était beaucoup plus faible que celle observée dans le village de Gioi.



Campora *Gioi* *Cardile*

Figure I.D.iii-3 : Moyenne des contributions génétiques pour les individus ayant des parents dans les généalogies (traits pointillés) et pour les fondateurs (ceux qui n'ont pas de parent) (traits pleins). Les individus sont regroupés en fonction de la valeur moyenne arrondie du nombre de générations qui les séparent de leurs descendants finaux.

Nous nous intéressons maintenant spécifiquement à la répartition des contributions génétiques des fondateurs (individus sans parents pris à n'importe quelle génération). Ces contributions ne sont uniformes ni dans le temps ni entre les individus. Dans la population de Cardile, nous observons que les individus sans parents qui participent le plus à la contribution génétique se trouvent entre 1750 et 1850 alors que plus de la moitié de ces individus sans parents se trouvent avant 1700 et que pourtant ils ne représentent que 10.4% de la contribution génétique (Tableau I.D.iii-8). Pour Campora, plus de la moitié de la contribution génétique est attribuable à des fondateurs nés avant 1750, ces fondateurs représentant la moitié des fondateurs. Par ailleurs, dans ce village, la moitié de la contribution génétique totale est due à 20% des plus gros contributeurs (Tableau I.D.iii-8). A Gioi, les contributions génétiques se répartissent plus régulièrement au cours du temps et la proportion cumulée de contribution génétique reste inférieure à la proportion de fondateurs correspondante. De plus, la contribution génétique des 20% plus gros contributeurs génétiques représente 56% de la contribution génétique totale (Tableau I.D.iii-9).

Tableau I.D.iii-8, Pourcentage cumulé de la contribution génétique des fondateurs, pour une transmission parent-enfant (A), mère-fille (M) et père-fils (Y), en fonction de l'année de naissance du fondateur, pour les trois populations. Seuls les individus nés avant 1960 sont intégrés dans les calculs de contribution génétique. Entre parenthèses se trouve le pourcentage cumulé des fondateurs correspondants.

	Campora			*Gioi*			*Cardile*		
	A	*M*	*Y*	*A*	*M*	*Y*	*A*	*M*	*Y*
<1700	40.4 (31.5)	45.5 (34.1)	55.8 (28.6)	34.3 (42.6)	28.2 (42)	61.6 (43.6)	10.4 (53.7)	4.7 (53.5)	12.7 (54)
<1750	55.9 (52.8)	57.8 (56.3)	62.6 (49)	43.7 (55.7)	39.2 (56.6)	63.1 (54.2)	28.1 (65.9)	39.1 (67)	18.6 (64.3)
<1800	66.7 (64.9)	68.2 (67)	70.6 (62.4)	58.4 (71.6)	51.5 (71.1)	68.5 (72.5)	70.2 (83.9)	57 (82.8)	83.3 (85.3)
<1850	77.2 (76.7)	75.3 (78.1)	74.8 (75.1)	66.4 (77.8)	63 (77.6)	72.4 (78.2)	77.6 (88.8)	61.7 (88.2)	91.2 (89.7)
<1900	85.4 (84.5)	83.1 (85.9)	84.7 (82.9)	77.4 (85.5)	74.4 (85.5)	82.8 (85.6)	85.8 (92.6)	80.5 (92.1)	94.1 (93.3)
<1960	100 (100)	100 (100)	100 (100)	100 (100)	100 (100)	100 (100)	100 (100)	100 (100)	100 (100)

Tableau I.D.iii-9 Pourcentage cumulé de la contribution génétique à la population actuelle en fonction du pourcentage cumulé des fondateurs (classés par ordre croissant de contribution génétique).

	Campora	*Gioi*	*Cardile*
10%	32.0	36.3	50.3
20%	49.9	56.0	72.6
30%	63.6	69.1	86.1
40%	73.8	79.7	94.8
50%	81.8	87.4	98.0
60%	88.9	92.9	99.0
70%	93.8	96.5	99.5
80%	97.4	98.7	99.8
90%	99.4	99.7	99.9

Contributions uniparentales

Lorsque nous analysons de quelle manière les individus sans parents contribuent génétiquement aux lignées matrilinéaires et patrilinéaires, nous observons une variabilité entre les trois populations.

Dans la population de Cardile, la majorité de la contribution génétique aux lignées patrilinéaires est due à des hommes nés avant le XVIII^{e} siècle, tandis que la contribution aux lignées matrilinéaires se fait plus régulièrement dans le temps (Tableau I.D.iii-8). Dans le village de Cardile, l'ensemble des lignées patrilinéaires et l'ensemble des lignées matrilinéaires dérivent chacune de seulement 15% des fondateurs hommes et des

fondateurs femmes ayant laissé au moins un descendant dans la population actuelle (Tableau I.D.iii-10).

Pour la population de Campora, la contribution des fondatrices et des fondateurs pour les lignées matrilinéaires et patrilinéaires se fait relativement progressivement dans le temps mais avec déjà plus de 50% de la contribution génétique dus à des fondatrices et des fondateurs nés avant 1750 (Tableau I.D.iii-8). Dans cette population, 20% des fondatrices et 25% des fondateurs sont respectivement parvenus à transmettre à la population actuelle leur ADN mitochondrial et leur chromosome Y (Tableau I.D.iii-10).

Pour Gioi la majorité des contributions génétiques aux lignées uniparentales est attribuable à des individus fondateurs nés avant 1900. Par ailleurs, on observe comme à Campora que 75% des lignées patrilinéaires et 80% des lignées matrilinéaires ont été perdues (Tableau I.D.iii-10).

Tableau I.D.iii-10, Pourcentage cumulé de la contribution génétique pour le compartiment bi-parental (A), matrilinéaire (M) et patrilinéaire (Y) à la population actuelle en fonction du pourcentage cumulé de femmes fondatrices pour la contribution génétique aux lignées matrilinéaires (M) ou du pourcentage cumulé d'hommes fondateurs pour la contribution génétique aux lignées patrilinéaires (Y).

	Campora		*Gioi*		*Cardile*	
	M	*Y*	*M*	*Y*	*M*	*Y*
2.5%	46.8	41.1	39.6	34.5	44.5	39.2
5.0%	62.3	57.1	58.1	52.7	65.6	63.7
7.5%	72.1	68.1	69.6	64.0	79.7	79.4
10.0%	81.2	77.3	79.7	72.9	89.1	87.3
12.5%	87.0	84.7	85.9	81.3	95.3	94.1
15.0%	90.9	88.3	92.1	85.7	100.0	100.0
17.5%	95.5	92.0	98.2	90.1	100.0	100.0
20.0%	100.0	95.7	100.0	94.6	100.0	100.0
22.5%	100.0	99.4	100.0	99.5	100.0	100.0
25.0%	100.0	100.0	100.0	100.0	100.0	100.0

Déséquilibre des généalogies de gènes des lignées uniparentales

Pour les trois populations nous avons étudié la forme des généalogies de gènes de l'ensemble des lignées matrilinéaires (transmission mère-fille) et de l'ensemble des lignées patrilinéaires (transmission père-fils) (Tableau I.D.iii-11). Les valeurs de déséquilibre des généalogies ne sont significatives dans aucun des trois villages et pour aucun des deux types de transmission. Néanmoins dans quatre cas sur six, l'indice de déséquilibre est sensiblement

supérieur à 0.5, suggérant donc une tendance vers des arbres déséquilibrés. Lorsqu'on compare au sein d'une même population le déséquilibre des généalogies matrilinéaires à celui des généalogies patrilinéaires, la situation est variable suivant les populations puisque dans le village de Campora, les lignées mitochondriales sont déséquilibrées alors que celles du chromosome Y ne le sont pas, alors que le signal opposé est observé dans le village de Gioi. Quant au village de Cardile, les deux types de généalogies présentent un signal de déséquilibre, l'indice de déséquilibre pour les lignées patrilinéaires étant légèrement supérieur à celui des lignées matrilinéaires.

Tableau I.D.iii-11, Valeur moyenne du déséquilibre des généalogies de gènes pour les lignées matrilinéaires (M) ou pour les lignées patrilinéaires (Y). n : nombre de nœuds sur lesquels est calculée la valeur du déséquilibre. p : *p-value* du déséquilibre estimée à partir de 5000 simulations selon la méthode par permutation de Purvis (2002).

	Y	*M*
Campora	0.4835 (n=25, p=0.553)	0.5759 (n=23, p=0.1976)
Gioi	0.5604 (n=24, p=0.2832)	0.4887 (n=25, p=0.5426)
Cardile	0.6126 (n=13, p=0.1786)	0.5846 (n=15, p=0.2978)

Analyse des séquences

A partir des séquences du mitochondriales échantillonnées dans les populations de Gioi et Campora, nous reconstruisons les arbres par une méthode de Neighbor-joining et analysons le déséquilibre dans les histoires de gènes reconstruites par phylogénie. Nous observons que les valeurs sont plus déséquilibrées que la valeur attendue de 0.5, mais les valeurs ne sont pas significatives. (Tableau I.D.iii-11)

Tableau I.D.iii-11, Déséquilibre, nombre de séquences analysées, nombre de nœuds sur lequel le déséquilibre a été calculé, *p-value* du déséquilibre calculé sur les arbres reconstruits avec une méthode de Neighbour-joining sur des séquences du HVRI de la mitochondrie dans deux populations : Campora et Gioi.

	Nombre de séquences (n)	*Déséquilibre* I_p	*Nombre de noeuds*	*P-value*
Campora	45	0.5883	14	0.207
Gioi	29	0.6828	10	0.1246

On observe que tous les tests de neutralité utilisés (D2* (Fu & Li 1993), D (Tajima 1989), et F / F* (Fu & Li 1993)) sont significativement négatifs à un seuil de 5% dans les deux populations (Tableau I.D.iii-12), le test D de Tajima ayant une *p-value* légèrement inférieure à celle des deux autres tests..

Tableau I.D.iii-12, Tests de neutralité sur les séquences de la région HVRI de l'ADN mitochondrial dans les populations de Campora et de Gioi. *n* : nombre de séquences. La *p-value* est calculée par un test unilatéral.

	Campora (n=45)		*Gioi (n=29)*	
Statistique	*Valeur*	*p-value*	*Valeur*	*p-value*
D	-1.8526	0.0112	-1.7586	0.01961
D2*	-1.7789	0.0466	-2.2462	0.02480
F*	-2.0147	0.0339	-2.2759	0.02486

I.D.iii.e Conclusion

Nous avons analysé les trois populations du Cilento dans plusieurs buts. Le premier était de comparer les résultats théoriques aux données réelles et d'analyser la structure sociale de ces populations. Dans cette perspective, une des difficultés de l'analyse à partir d'une base de généalogies ascendantes est celle de connaître la cause de l'absence d'informations sur les parents de certains individus. En effet, des fondateurs peuvent correspondre soit à des individus nouveaux arrivant dans la population, soit à des individus autochtones mais pour lesquels les informations parentales sont tout simplement inconnues.

Pour cela nous avons confronté l'étude de plusieurs variables qui ont, pour certaines, déjà été utilisées par d'autres auteurs, comme les corrélations intergénérationnelles entre taille de fratrie, et qui sont, pour d'autres, plus novatrices comme le déséquilibre des lignées généalogiques. Nous discutons ci-dessous les résultats que nous avons obtenus dans les trois villages.

Cardile

La population de Cardile semble avoir été soumise à une force sélective ou démographique, un très faible nombre d'individus ont participé à la majeure partie de la contribution génétique de la population actuelle, par exemple seulement 15% des fondateurs ont laissé leurs chromosomes Y et 15 % des femmes fondatrices ont laissé leurs mitochondries à la génération actuelle (on observe des valeurs environ deux fois plus grandes dans les deux autres populations). Le même constat peut être effectué pour la contribution génétique des autosomes et pour laquelle 10% des plus gros fondateurs sont à l'origine de plus de 50% de la contribution génétique totale, ce qui est supérieur à la valeur observée pour une population de taille constante, simulée dans un modèle de Wright-Fisher sur 16 générations (cf. partie I.D.ii). La population de Cardile semble donc avoir été soumise à une forte dérive avec certains individus ayant beaucoup plus contribué que d'autres au patrimoine génétique actuel de la population.

Cardile est un village qui a été fondé par des habitants de Gioi au XI^{e} siècle, à quelques kilomètres du village de Gioi. Cette proximité géographique et cet événement de fondation expliquent l'importance des migrations entre les deux villages. Par exemple, nous observons que 80% des fondateurs de Cardile apparaissent également dans les généalogies du village de Gioi. Lorsqu'on compare la différence de contribution de chacun de ces individus entre Gioi et Cardile, on constate qu'un individu qui a beaucoup contribué à Cardile a peu contribué à Gioi et vice versa. Par exemple, 10% des plus gros contributeurs de Cardile, parmi les fondateurs communs à Gioi et Cardile, représentent 22 % de la contribution génétique à Cardile et ces mêmes individus ne représentent que 2% de la contribution à Gioi. Donc ceci peut expliquer partiellement pourquoi des fondateurs de Cardile n'ont qu'une faible contribution à Cardile, et certains beaucoup plus. Ceci serait une trace de forte structuration entre les deux villages. Ceci est cohérent avec les résultats de Colonna et *al.* (2009) qui avaient montré que la structure de ces deux populations pouvait être détectée par l'utilisation de microsatellites. De même cela pourrait signifier que les nouveaux migrants arrivant de Gioi à Cardile vont avoir plus de mal à s'installer et à installer leurs enfants que quelqu'un qui appartient déjà à la population. Malheureusement nous n'avons pas les lieux de mariage pour étayer cette hypothèse d'un cœur et d'une frange de la population (Heyer 1993).

Certains de nos autres résultats pourraient appuyer cette hypothèse d'un différentiel de reproduction « utile » selon le degré « d'autochtonicité » des individus. Par exemple, les indices de déséquilibre des lignées matrilinéaires et patrilinéaires sont supérieurs à 0.5, même s'ils se sont tous révélés non significatifs, ce qui va dans le sens de ce qui est attendu dans le cas d'une transmission du succès reproducteur. Cependant nous avons montré dans la partie précédente que quand nous calculions le déséquilibre des arbres sur l'ensemble des nœuds, il était sous-estimé. Il conviendrait donc de calculer ce déséquilibre en ne tenant compte que des nœuds suffisamment hauts dans l'arbre. Aussi, on observe une corrélation positive entre le nombre d'enfants d'un individu et la taille de sa fratrie que l'on retrouve lorsque l'on considère non plus les individus mais les couples. (r=0.23). Autant les corrélations entre taille de fratrie et nombre d'enfants, les contributions fortes de certains individus et la présence de déséquilibre dans les lignées Père-Fils et Mère-Filles pourraient être expliqués par une transmission du succès reproducteur bi-parental.

Gioi

Une possible structuration sociale pourrait expliquer les résultats obtenus sur la population de Gioi. Tout d'abord, nous montrons que dès le XVIII^{e} siècle, il existe une corrélation positive entre la taille de la fratrie des individus et leur nombre d'enfants, qui va de pair avec une corrélation positive entre le nombre de sœurs d'une femme et son nombre de filles alors qu'aucune corrélation n'est observée pour les lignées pères-fils. Associé à ces corrélations, on observe que le pourcentage de femmes à l'origine des lignées mitochondriales actuelles est plus faible que le pourcentage d'hommes ayant transmis leur chromosome Y. Cette différence sexe-spécifique en ce qui concerne la dérive et la corrélation intergénérationnelle des tailles de fratries pourrait être expliquée par une migration plus importante de certaines lignées de femmes par rapport à d'autres. On peut également noter que la moyenne du nombre de filles est inférieure à celle du nombre de fils, la variance du nombre de filles étant plus importante que celle du nombre de fils. Un autre argument qui soutient l'existence d'une transmission de la mobilité chez les femmes est que dès le début de la période étudiée (c'est-à-dire pour les fondateurs nés avant le XVIII^{e} siècle), on observe une contribution génétique plus forte pour les lignées patrilinéaires que pour les lignées matrilinéaires (61.6 % et 28.2 % respectivement). Par ailleurs, nous observons que les arbres reconstruits sur les séquences de la région HVRI de l'ADN mitochondrial ont tendance à être déséquilibrés, bien que notre procédure de test ne démontre pas que ce déséquilibre est significatif. Enfin, les tests de neutralité sélective effectués sur les mêmes jeux de données ont tous les trois des valeurs négatives significatives, ce qui peut être soit le signal d'une sélection directionnelle, soit d'une expansion mais qui peut également être la trace d'une transmission du succès reproducteur (Sibert et al. 2002).

Campora

La population de Campora semble être constituée en cœur et en frange de la population. Les contributions génétiques sont stables dans le temps, sans qu'elles décroissent au contraire des deux autres généalogies (Gioi et Cardile). De plus, les ancêtres les plus anciens ont davantage contribué génétiquement que les ancêtres les plus récents à la population actuelle, par rapport à la part respective des ancêtres qu'ils représentent. Aussi, un individu qui est déjà inséré dans la population a une contribution plus importante qu'un individu arrivant dans la population. Même si nous n'observons pas une corrélation globale

entre les tailles de fratrie et le nombre d'enfants, elle est observée durant le XVII^{e} et XIX^{e} siècle.

Les derniers arguments qui sont en faveur de l'hypothèse d'une structuration sociale dans la population de Campora sont les fortes valeurs de déséquilibre calculées autant sur l'arbre reconstruit à partir des séquences de la région HVRI de l'ADN mitochondrial que sur les arbres des lignées matrilinéaires. Les valeurs ne sont pas significatives, ceci peut être expliqué par le faible nombre de nœuds des arbres généalogiques des gènes reconstruits, mais ce résultat peut être mis en relation avec la présence de valeur significative des tests de neutralité qui a été montré comme une trace de la transmission du succès reproducteur.

L'ensemble de ces éléments nous suggère une population structurée en cœur et frange de la population, où les individus qui sont insérés dans la population vont plus se reproduire que les individus qui arrivent dans la population. Colonna et *al.* (2007) montraient déjà que l'apparentement entre les individus augmentait avec le nombre de mariages exogames, et qu'il n'y avait qu'un nombre restreint d'haplotypes pour le chromosome Y et pour les mitochondries. Ce résultat pourrait être autant expliqué par le rôle d'un goulot d'étranglement dans l'histoire de la population que par une structuration de la population en plusieurs classes d'individus se reproduisant plus ou moins bien au sein de la population.

Conclusion

Les trois villages semblent avoir été influencés par certaines forces démographiques ou culturelles. En ce qui concerne les villages de Cardile et de Gioi, à la vue des données qui contiennent autant des individus de Gioi et Cardile, il serait nécessaire de connaître les lieux de mariage des individus pour séparer les individus qui sont de Gioi et ceux de Cardile, afin de vérifier notre conclusion de la présence d'un cœur et d'une frange de la population à Cardile et d'une possible migration sexe-spécifique entre hommes et femmes à Gioi qui entraîne dans les deux cas une transmission du succès reproducteur. Notre conclusion sur la population de Campora est plus nuancée. Dans cette population, il a été observé un apparentement qui augmente avec l'exogamie (Colonna et al. 2007). Nous y observons que les individus arrivant dans la population laissent moins leurs gènes dans la population que ceux déjà installés et des lignées plus déséquilibrées que dans les autres populations, mais il n'y a pas de corrélation entre les transmissions du succès reproducteur. Ces observations pourraient être dues à d'autres phénomènes sociaux qui demanderaient d'étudier l'histoire et les relations des individus de la population plus en détails.

I.D.iii.f Informations Supplémentaires

Tableau I.D.iii-S1, Nombre total d'hommes (H) et de femmes (H) dans les généalogies ascendantes des populations de Campora, Gioi et Cardile, selon leur année de naissance. Seuls les individus nés avant 1960 sont comptabilisés.

	Campora		*Gioi*		*Cardile*	
Naissance	*H*	*F*	*H*	*F*	*H*	*F*
<1700	173	187	364	408	281	308
1700-1750	182	179	271	322	135	160
1750-1800	219	234	311	390	152	166
1800-1850	238	250	296	372	147	172
1850-1900	214	233	246	288	126	148
>1900	311	318	330	388	203	230
Total	1337	1401	1818	2168	1044	1184

I.E Conclusion

De par son impact sur la diversité génétique, la transmission du succès reproducteur s'insère dans le domaine de la génétique des populations. Lorsque ce phénomène est dû à des facteurs culturels, son étude revêt un caractère transdisciplinaire puisqu'il faut non seulement analyser la diversité génétique des populations mais aussi comprendre les causes culturelles ou sociales qui modifient leurs paramètres démographiques (corrélation entre les tailles de fratrie, variance du nombre d'enfants). Notre travail consistait à comprendre les effets de la transmission du succès reproducteur et à tenter de le détecter.

Les raisons de la transmission du succès reproducteur peuvent être de deux ordres, socio-culturel ou génétique et sa détection est possible par des approches différentes (mais souvent complémentaires) : l'approche anthropologique, l'approche socio-démographique et l'approche génétique. Alors que la première tachera de reconstruire des arbres généalogiques à l'aide d'entretiens et d'apprendre les règles de mariage *etc.*, la seconde reposera sur des données démographiques historiques enregistrées et la dernière analysera la structure du polymorphisme génétique des populations actuelles. Dans notre cas, nous avons mobilisé les deux dernières approches et ainsi montré que les effets de la transmission du succès reproducteur sur la structure des généalogies et la diversité génétique dépendent de facteurs socio-culturels, et notamment que l'hétérogénéité du succès reproducteur et fonction des règles de transmission du succès reproducteur (patrilinéaire, matrilinéaire ou biparentale) et du choix du conjoint, notamment lorsqu'il y a homogamie pour la taille de la fratrie au sein des couples. L'ensemble de ces facteurs vont moduler les différents impacts démographiques et génétiques de la transmission du succès reproducteur, notamment les corrélations entre la

taille de la fratrie d'un individu et son nombre d'enfants, la variance du nombre d'enfants, l'effectif efficace de la population et le déséquilibre des arbres de gènes.

Intérêt et apport de l'extension du modèle

Notre modèle de simulation est une extension du modèle développé par Sibert et *al.* (2002). Il propose un cadre diploïde avec des individus sexués, de nouvelles règles culturelles (mariage avec ou sans homogamie pour la taille de la fratrie) et différents types de marqueurs (autosomes, chromosome X, ADN mitochondrial et chromosome Y). Cette complexification nous a permis d'analyser trois règles de transmission documentées dans les populations humaines (Kumar et al. 2006: biparentale, matrilinéaire et patrilinéaire). Chacun de ces phénomènes en association avec de la transmission du succès reproducteur va modifier la diversité génétique et les paramètres démographiques.

Règles de transmission

Pour les deux types de transmissions uniparentales du succès reproducteur lorsque nous avons analysé les corrélations de type mères–filles dans le cas d'une transmission matrilinéaire ou les corrélations pères–fils dans le cas d'une transmission patrilinéaire (cf. partie I.B), nous nous sommes retrouvé dans un cadre haploïde qui était celui des précédentes études (Blum et al. 2006; Sibert et al. 2002). Notre modèle nous permet cependant d'analyser l'impact de ce mode de transmission sur les autres compartiments (notamment le chromosome X et les autosomes). Par exemple, nous avons montré que les transmissions matrilinéaires, patrilinéaires et biparentales affectaient différemment la structure des arbres de coalescence des chromosomes X et des autosomes. Comme de nombreux marqueurs neutres peuvent être développés sur ces chromosomes, ceci ouvre des perspectives pour détecter ces différents types de transmission, en évitant les problèmes d'auto-stop génétique qui peuvent affecter les gènes du chromosome Y et de la mitochondrie (voir la partie « détection du phénomène » ci-dessous). Par ailleurs, un des résultats remarquables que nous avons observé est que la transmission par un seul des deux sexes affecte aussi la variance du succès reproducteur des individus de l'autre sexe et peut donc aussi amener une réduction de taille efficace pour l'autre sexe.

Les trois types de transmission (biparentale, matrilinéaire et patrilinéaire) sont retrouvés dans différentes populations. La transmission est par exemple matrilinéaire dans les populations de chasseurs-cueilleurs (Blum et al. 2006), biparentale dans la population du Saguenay Lac Saint Jean (Austerlitz & Heyer 1998) et patrilinéaire dans les populations

pastorales d'Asie centrale (cf . partie I.C.ii). Comme le prévoyaient les études théoriques autant dans les populations de chasseurs-cueilleurs que dans les populations pastorales d'Asie centrale, le déséquilibre est associé à une diversité réduite (Sibert et al. 2002). Il serait maintenant intéressant de contraster le chromosome X et les autosomes dans ces différentes populations.

Notons pour finir que plusieurs de nos hypothèses peuvent être discutées. Par exemple, pour la transmission biparentale nous avons supposé une transmission liée à la moyenne des tailles des fratries des deux parents, et dans le cas des transmissions uniparentales, un succès reproducteur lié à l'ensemble de la fratrie du parent (y compris les individus de l'autre sexe dans cette fratrie). Ces hypothèses pourront être relâchées dans des futurs développements du modèle.

Hétérogénéité du succès reproducteur

Indépendamment de la transmission du succès reproducteur, on peut observer une hétérogénéité du succès reproducteur, elle aussi, est liée à des phénomènes culturels ou génétiques : rang de naissance, âge de la mère à la naissance et survie du précédent enfant… (Cohen 1975; Nault et al. 1990; Pedersen 2000; Ronsmans 1995). L'impact de cette hétérogénéité, associée à de la transmission du succès reproducteur sur les paramètres démographiques ou génétiques, avait été partiellement analysé dans des études précédentes (Austerlitz & Heyer 1998; Sibert et al. 2002). Nous avons montré que cette hétérogénéité augmente l'effet de la transmission du succès reproducteur sur le déséquilibre des arbres, tant à l'échelle des lignées extraites des généalogies (cf. partie I.D) que pour des arbres de coalescence (cf. partie I.B). Elle entraîne aussi une diminution plus forte de la diversité génétique. De plus, nous avons souligné que les valeurs de déséquilibre prévues par nos modèles pour les niveaux de corrélations intergénérationnelles du succès reproducteur observées dans les populations humaines limitaient la détection de la transmission du succès reproducteur en utilisant le déséquilibre des arbres sans présence d'une hétérogénéité du succès reproducteur (cf. partie I.B).

Homogamie dans la taille de fratrie des conjoints

Finalement, nous avons analysé l'impact du choix non aléatoire du conjoint en simulant une homogamie pour la taille de la fratrie lors de la formation des couples. Dans l'introduction, nous avons souligné qu'une corrélation entre les tailles de fratrie des deux

conjoints avait été observée dans certaines populations humaines. Comme c'est le cas pour la transmission ou l'hétérogénéité du succès reproducteur, cette homogamie peut être liée à des phénomènes socio-culturels. Dans notre modèle biparental, cette homogamie entraîne une accentuation des effets démographiques et génétiques de la transmission du succès reproducteur. En effet, comme le succès reproducteur d'un couple dépend de la moyenne des tailles de fratries des deux conjoints, l'association préférentielle de deux individus provenant de fratries de grande taille ou de deux individus provenant de fratries de petite taille va très fortement jouer sur la variance du nombre d'enfants et aussi sur les corrélations entre taille de fratrie et nombre d'enfants. Pour finir, l'augmentation de ces deux variables démographiques est aussi associée à une augmentation du déséquilibre pour une transmission du succès reproducteur de niveau intermédiaire, ce qui indique donc que l'homogamie devrait permettre de faciliter la détection de cette transmission dans les populations. Dans ce contexte, il est intéressant de noter que les phénomènes qui entraînent un choix du conjoint non aléatoire ou une différence d'hétérogénéité du succès reproducteur, sont également ceux qui favorisent la transmission du succès reproducteur (alliance familiale, transmission du rang social, …).

Détection du phénomène

Nous avons utilisé deux types de support pour détecter la transmission du succès reproducteur : les généalogies reconstruites à partir des données de démographie historique et les informations génétiques.

Utilisation des généalogies

Les généalogies reconstruites des individus sont un bon moyen de détecter la transmission du succès reproducteur. Sur plusieurs générations, on peut mesurer directement une relation entre le succès reproducteur des parents et celui de leurs enfants grâce aux différentes mesures de corrélations (père-fils, mère-filles, parents-enfants). Mais nous montrons que si les généalogies sont de type ascendant plutôt que descendant le signal s'en trouvait diminué (cf. partie I.D).

Toujours dans le cadre de l'étude des généalogies, nous avons montré que l'observation d'un déséquilibre au sein des lignées d'individus extraites des généalogies est une trace de la transmission du succès reproducteur. Au contraire de la corrélation qui analyse le phénomène sur deux générations successives, le déséquilibre analyse l'ensemble de la généalogie. Nous avons testé cette approche dans la population de Cardile, où nous montrons des corrélations positives entre les tailles de fratrie des parents et de leurs enfants (mais aussi

entre le nombre de frères et de fils d'un homme). Nous mettons également en évidence un déséquilibre des arbres reconstruits à partir des lignées matrilinéaires et patrilinéaires plus forts qu'attendu dans un coalescent de Kingman, même si cette tendance est non-significative. Les deux mesures semblent donc aller dans le même sens.

Une autre piste intéressante est l'analyse de la distribution des contributions génétiques calculées à partir des généalogies des ancêtres de la population actuelle. Nous avons montré la dépendance de cette distribution à la transmission et à l'hétérogénéité du succès reproducteur. Par ailleurs, en comparant les contributions des néo-arrivants avec celles des individus déjà installés dans une population, comme nous l'avons fait dans les populations de Gioi, Cardile et Campora, nous avons pu montrer une différence de succès reproducteur entre les individus récemment installés ou non.

Utilisation des données génétiques

Malgré leur intérêt, les généalogies d'individus ne sont une source d'information disponible que pour un nombre limité de populations, car elles ne peuvent être réalisées que lorsqu'il y a la présence d'un état civil où sont enregistrés naissances, mariages et décès. De plus, la reconstruction des généalogies est un travail long et fastidieux (nécessité d'aller consulter et retranscrire de nombreux actes paroissiaux ou civils). L'utilisation des données génétiques permet de s'affranchir de certaines difficultés en se focalisant sur un échantillon actuel. Cela permet aussi de ne pas être limité comme pour les généalogies par la profondeur maximale de celles-ci.

Comme nous l'avons vu, les effets de la transmission du succès reproducteur sur la structure des arbres de coalescence sont divers. Nous les avons décrits dans l'introduction (cf. Partie I.A) et nous avons décidés d'utiliser le déséquilibre dans les arbres de coalescence pour détecter le phénomène. En effet, le déséquilibre des arbres de coalescence est l'un des signaux les plus spécifiques de la transmission du succès reproducteur même si dans certains cas de sélection purificatrice (Maia et al. 2004) ou de scénarios démographiques complexes (Blum et al. 2006), on peut s'attendre à observer du déséquilibre. Le déséquilibre ne peut toutefois pas être une mesure directe de l'intensité du phénomène car il est dépendant d'autres phénomènes comme l'hétérogénéité du succès reproducteur.

Comme le déséquilibre est un bon indicateur de la présence d'une transmission du succès reproducteur, il est nécessaire de savoir s'il peut être retrouvé à partir des données

génétiques échantillonnées dans les populations. Blum et *al.* (2006) ont montré que le déséquilibre dans des arbres reconstruits par phylogénie suivait globalement la même tendance que celui des arbres de coalescences originaux. Cependant, dans leur cas le déséquilibre était surestimé du fait de la méthode utilisée pour reconstruire l'arbre (PhyML) et du nombre insuffisant de sites polymorphes utilisés pour reconstruire l'arbre (cf. partie I.Ci et Blum et al. 2006). La conséquence de cette surestimation du déséquilibre est l'augmentation de la probabilité de conclure que la population est soumise à la transmission du succès reproducteur alors que ce n'est pas le cas. Pour prendre en compte ce biais, nous proposons une *p-value* fondée sur des simulations qui utilisent explicitement la méthode de reconstruction phylogénétique, le nombre de sites polymorphes présents sur les séquences et la taille de l'échantillon. Nous avons ainsi pu obtenir un test moins biaisé de la transmission du succès reproducteur.

La question est ensuite de savoir à quel type de données il est préférable d'appliquer ces méthodes. La reconstruction sur des séquences mitochondriales ou celles du chromosome Y permet de définir s'il y a une transmission soit matrilinéaire comme dans les populations de chasseurs-cueilleurs (Blum et al. 2006) soit patrilinéaire comme dans certaines populations d'Asie centrale (partie I.C.ii). Cependant plusieurs biais sont associés à ces chromosomes qui peuvent être soumis à sélection, comme cela a été montré notamment pour la mitochondrie (Balloux et al. 2009; Bazin et al. 2006; Elson et al. 2004; Ruiz-Pesini et al. 2004; Stewart et al. 2008). L'analyse seule du chromosome Y ou de l'ADN mitochondrial ne permet pas non plus de déterminer si la transmission est uniparentale ou biparentale. Pour finir, la mitochondrie et le chromosome Y ne représentent chacun qu'une seule histoire de gènes au contraire des autosomes, où chaque nouvelle recombinaison permet de découpler les généalogies des gènes. Pour ces raisons, nous avons décidé de détecter la transmission du succès reproducteur en utilisant les autosomes et le chromosome X, ces chromosomes offrant de nombreux locus neutres situés suffisamment loin des locus sous sélection. De part leur mode de transmission différent (biparental ou biaisé vers une transmission matrilinéaire), la comparaison des autosomes et des chromosomes X apparaît comme une piste très prometteuse pour détecter si la transmission est uniparentale ou biparentale.

Cependant un des problèmes associés à l'utilisation du polymorphisme de séquences issues de ces chromosomes est la présence de recombinaisons intragéniques au sein même de ces séquences. Plusieurs études ont montré que la non prise en compte cette présence de recombinaison pouvait entraîner des biais dans les reconstructions phylogénétiques (Arenas &

Posada; Posada & Crandall 2002; Schierup & Hein 2000) ou dans certaines analyses en génétique des populations (Anisimova et al. 2003; Ramírez-Soriano et al. 2008). Ces recombinaisons entraînent notamment un excès de déséquilibre dans les arbres reconstruits (Schierup & Hein 2000). Il est donc nécessaire de bien détecter les recombinaisons pour délimiter au sein des chromosomes les blocks non-recombinants sur lesquels peuvent être reconstruits les arbres.

Dans ce cadre, nous avons pu montrer à partir des données de HapMap (version 2) qu'il est possible d'isoler un certain nombre de blocks non-recombinants (entre 500 et 3500 selon les populations) ayant un nombre de sites polymorphes suffisants pour reconstruire des arbres. Dans les populations de HapMap, nous montrons qu'entre 10% et 20% des arbres sont en déséquilibre dans les populations d'Europe et d'Afrique, mais que cette proportion augmente à plus de 25 % dans les populations asiatiques. Nous n'avons aucune hypothèse sur l'origine de ce déséquilibre, à part un possible biais dû aux données. Par exemple, ce biais pourrait être dû à un manque d'allèles mineurs, ce qui biaiserait la *p-value* ou la détection des recombinaisons. Pour confirmer les résultats et trouver une explication à ce phénomène, il sera intéressant d'étendre nos travaux à des données comme HapMap version 3 ou le projet 1000 génomes (Via et al.), qui comprennent davantage de populations, avec notamment des pratiques culturelles différentes, en particulier de part la présence de populations de chasseur-cueilleurs.

Conclusion

Ce travail apporte donc des avancées dans la compréhension de l'impact de la transmission du succès reproducteur et des phénomènes qui peuvent lui être associés ou la moduler (hétérogénéité du succès reproducteur ; règle de transmission, uniparentale ou biparentale ; homogamie pour la taille de fratrie) sur la diversité génétique et la structure des généalogies d'individus et des arbres de coalescence des gènes. Il nous a permis aussi de déterminer dans quelles conditions ce phénomène peut être détecté dans les populations, tant sur la base de données généalogiques que génétiques. Nous avons pu de ce fait le mettre en évidence dans un certain nombre de populations en utilisant ces deux types de données. Comme ce phénomène de transmission du succès reproducteur est observé tant dans des populations humaines (Austerlitz & Heyer 1998; Blum et al. 2006; Bocquet-Appel & Jakobi 1993; Pluzhnikov et al. 2007) qu'animales (Frère et al. 2010; Kelly 2001; Whitehead 1998),

et qu’il a un impact fort notamment sur la diversité génétique et la fréquence des maladies génétiques (Austerlitz & Heyer 1998), il sera très utile de pouvoir déterminer quelle est l’étendue du phénomène dans les populations humaines ou celles d’autres espèces.

Partie II : Caractères à déterminisme complexe : impacts de la sélection sur des réseaux de gènes

II.A Introduction

Préambule

Ce chapitre s'axe principalement sur notre article publié dans Journal of Evolutionary Biology intitulé « Impact of selection on genes involved in regulatory network: a modelling study » (Rhone et al. 2011) qui étudie l'évolution sous l'action de la sélection naturelle d'un caractère quantitatif codé par un réseau de gènes (partie II.B). Dans la suite, nous rappelons brièvement les concepts utilisés dans le cadre de ce travail, notamment la notion de phénotype, les liens de celui-ci avec le génotype, plus particulièrement dans le cas des réseaux de gènes en interaction. Enfin nous décrivons les approches de modélisations utilisées dans ce cadre.

Caractère et phénotype

Le caractère, en biologie, est une propriété moléculaire, physiologique, anatomique, morphologique ou éthologique d'un organisme vivant qui peut être mesurée ; l'état particulier d'un caractère est un phénotype. Ces différents états du caractère vont dépendre du génotype, de l'environnement et de l'interaction entre les deux. Nous appellerons gènes les zones du génome qui participent aux variations de ces caractères. Ces zones peuvent être transcrites ou non, comme par exemple les zones de régulation ou les promoteurs.

Le nombre de gènes impliqués dans un phénotype, peut être variable allant d'un seul gène à plusieurs dizaines ou centaines de gènes dans le cas de caractères complexes quantitatifs comme la floraison chez *Arabidopsis thaliana* (Keurentjes et al. 2007 et Figure II.A-1) ou le risque de diabète de type II où de nombreux gènes à risques ont été détectés par des méthodes d'association (voir la revue de Travers & McCarthy 2011). Un certain nombre de méthodes ont été développées pour détecter les gènes impliqués dans les caractères. Par exemple les méthodes de détection de QTLs (pour « Quantitative Trait Locus » en anglais ou locus associés à des traits quantitatifs) permettent d'associer des régions du génome à la variabilité du caractère dans une population donnée. De même, les tests d'association permettent d'associer des allèles particuliers à un locus avec des caractères, ce qui est fait notamment dans le cadre des maladies polygéniques. D'après une récente revue de Roff (2007), le nombre moyen de QTLs détectés pour un caractère donné serait de l'ordre de 20, avec de fortes variations des effets de ces QTLs, quelques uns ayant un effet majeur et beaucoup un effet faible.

Une autre source permettant de relier le génotype au phénotype est l'étude des réseaux de régulation et des interactions protéines-ligands, cette approche permet de reconstruire sur des diagrammes l'ensemble des relations entre gènes, protéines, ARNs, ligands qui vont « construire » un caractère. Ces graphes de relation ont été beaucoup développés pour les voies métaboliques et sont notamment bien référencés dans la base de données Keeg (Kanehisa et al. 2008).

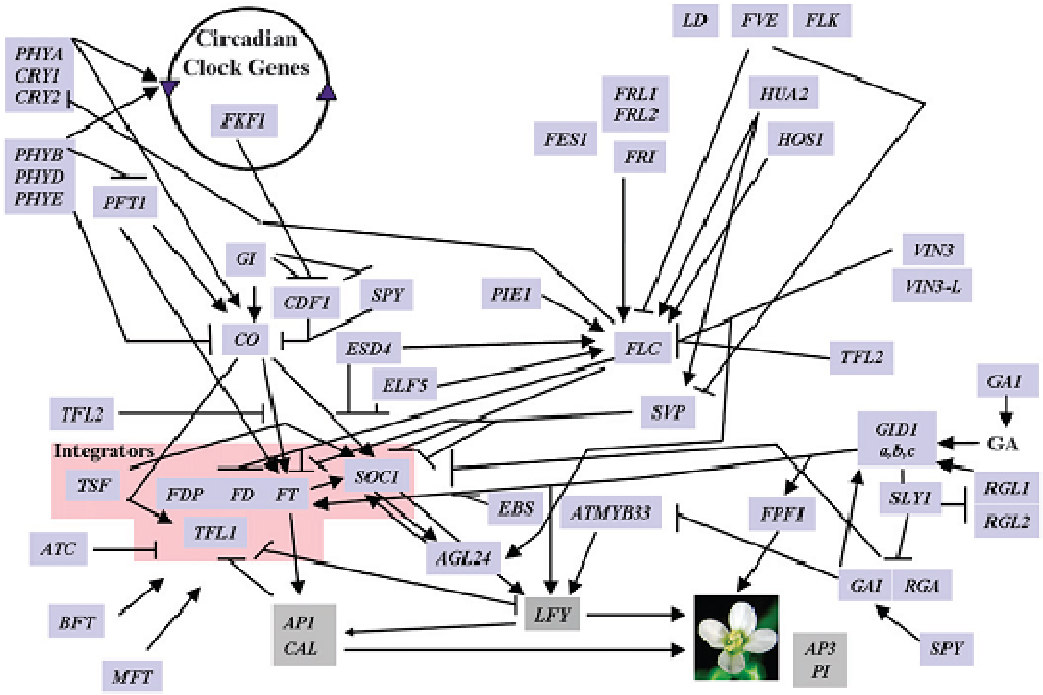


Figure II.A-1, Réseau de régulation de temps de floraison chez arabidopsis (extrait de Ehrenreich et al. 2009)

De la variation du génotype à la variation du phénotype

Le phénotype résulte de l'expression du génotype, de l'action de l'environnement et de l'interaction entre les deux. Si l'on se focalise sur l'impact du génotype sur le phénotype, on peut distinguer deux grands types de mutation qui vont affecter le phénotype. Le premier type concerne les mutations au sein de la séquence des gènes eux-mêmes. Il peut s'agir de mutations dans les exons qui vont modifier par exemple la structure des protéines ou de délétions qui vont modifier le cadre de lecture et rendre la protéine inactive. L'ensemble des phénotypes dépendant de cette protéine sera alors affecté à des degrés divers, selon le niveau d'implication de la protéine dans le caractère.

Un autre type de mutation qui va affecter les phénotypes sont les mutations affectant les régions régulatrices des gènes, qui vont modifier le niveau d'expression de ceux-ci. Plusieurs niveaux de régulation existent, au niveau de l'ADN, de l'ARN ou de la protéine. En

premier lieu, la régulation de la transcription de l'ADN en ARN se fait en général par l'accès ou l'affinité de l'ARN polymérase à l'ADN. Par exemple, la méthylation de l'ADN empêche la transcription comme c'est le cas pour l'un des deux chromosomes X chez les femelles des mammifères. En deuxième lieu, les facteurs *Cis* et *Trans* régulateurs modifient la régulation de l'ADN. Les facteurs *Trans* sont des protéines qui se positionnent sur des facteurs *Cis*, empêchant l'accès à la polymérase de manière plus ou moins forte selon l'affinité entre ces facteurs. Ces affinités dépendront de la séquence ADN présente au sein des promoteurs et des zones régulatrices comme de la structure des facteurs *Trans*. C'est ce type de régulation que nous avons modélisé dans le cadre de cette étude. Il est à noter qu'il existe aussi d'autres types de régulations, comme par exemple, au niveau de l'ARN, la régulation de micros ARN sur des ARN messagers complémentaires (Lagos-Quintana et al. 2001).

Modéliser l'impact de la sélection naturelle sur le génotype et le phénotype.

Certains individus, ayant un phénotype plus adapté dans un environnement donné, vont se reproduire davantage que d'autres, transmettant préférentiellement leur patrimoine génétique. De ce fait il est attendu une évolution des fréquences alléliques au sein des populations. Dans le cas où un caractère est associé à un seul locus ou à un très faible nombre de locus, des modèles analytiques sont disponibles depuis longtemps pour prédire l'action de la sélection naturelle sur les locus en question (par exemple Fisher 1930). La question devient cependant beaucoup plus complexe dans le cas des caractères codés par de nombreux locus. Les développements analytiques dans ce cadre ont été effectués en général dans le modèle infinitésimal utilisé en génétique quantitative (cf. par exemple Lande 1976), qui considère que le caractère est codé par un très grand nombre de gènes, chacun ayant individuellement un effet similaire et extrêmement faible sur le caractère. Le déterminisme du caractère suit un modèle additif, auquel sont ajoutées éventuellement des interactions de dominance et d'épistasie. Ce modèle n'est pas forcément très réaliste du point de vue biologique pour plusieurs aspects. Par exemple, il a été montré que les caractères sont certes codés par de nombreux gènes, mais que certains de ces gènes ont un effet beaucoup plus fort que d'autres sur le caractère (voir la revue de Roff 2007).

L'apport des outils de simulation informatique a permis de développer des méthodes pour étudier l'évolution des gènes impliqués dans les caractères complexes. Certains modèles considèrent un modèle additif mais à la différence du modèle infinitésimal, ils considèrent un nombre limité de locus et la possibilité d'avoir des locus ayant des effets plus ou moins forts

sur les caractères (Latta 1998; Le Corre & Kremer 2003). Ils ont permis notamment de montrer que la réponse adaptative des phénotypes se faisait autant par des changements de fréquences alléliques aux locus que par la création d'associations alléliques entre les différents locus. D'autres modèles se sont affranchis du modèle additif en considérant comme caractère quantitatif le flux à travers un réseau métabolique (Bost et al. 1999). Fondés sur la théorie du contrôle métabolique (Kacser & Burns 1981), ils permettent de prédire notamment la distribution des effets des gènes sur le caractère.

Ces modèles ne prennent pas en compte les régulations qui peuvent exister entre les gènes. Dans ce domaine, la dynamique des réseaux de régulation décrit comment vont s'exprimer les gènes au sein de ces réseaux, en fonction des interactions entre les gènes et de leurs propriétés biochimiques. Plusieurs types de modèles analytiques et dynamiques ont été développés, comme les modèles booléens ou des modèles utilisant des équations différentielles (voir la revue de Schlitt & Brazma 2007). Ces modèles s'intéressent à la dynamique et à la stabilité des réseaux de régulation mais uniquement dans le cadre du développement des individus. Ils ne s'intéressent pas à l'évolution au sein des populations de ces réseaux au cours du temps, sous l'effet de la sélection naturelle.

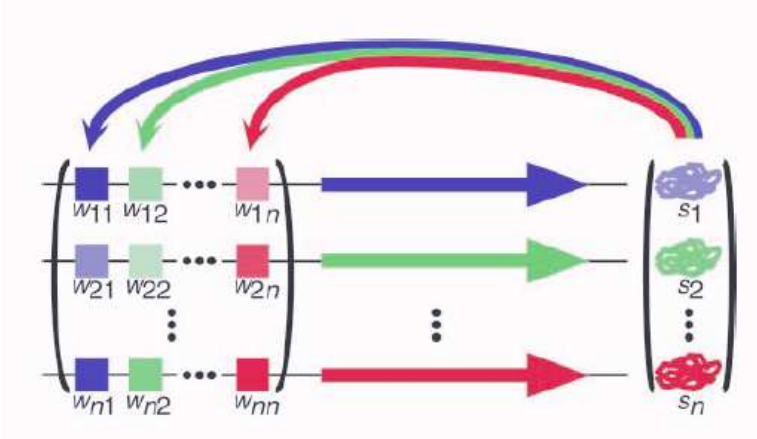


Figure II.A-2, Représentation d'un réseau de régulation de gènes. Chaque gène est régulé par le produit des autres gènes via des régions régulatrices en amont (carré). L'intensité et le type de la régulation (positive ou négative) dépendent à la fois de la région régulatrice et de l'abondance des produits des gènes correspondants. Le Génotype est la matrice, w, les forces de régulations, S, les produits. Figure extraite de (Siegal & Bergman 2002)

Cette approche évolutive des réseaux de régulation a été développée par Wagner (1996) et Siegal et Bergman (2002), sur la base d'un modèle simple qui considère que le génotype d'un individu est une matrice donnant l'intensité des niveaux de régulation entre gènes : la matrice donne pour chaque gène le niveau de régulation qu'il exerce sur lui-même et les autres locus (cf. Figure II.A-2,). Ce modèle étant celui que nous avons étudié, il est présenté plus en détail dans la partie II.B. L'un des principaux résultats de ces auteurs a

été de montrer que ce type de réseaux de régulation évoluait de sorte à devenir robustes à l'apparition de mutations délétères sur certains de leurs gènes, dans le sens que ces mutations n'entraînaient pas de changement majeur de leur phénotype. Ceci se faisait grâce à des interactions épistatiques permettant d'annihiler ou du moins de limiter l'effet de ces mutations et constitue le phénomène de canalisation décrit par Waddington (1942).

Objectif de ce chapitre

Comme nous l'avons vu, la construction des caractères complexes résulte de processus biochimiques qui peuvent autant affecter la structure des protéines que les niveaux de régulations. La sélection va favoriser les phénotypes les mieux adaptés, et il est important dans ce cadre de comprendre comment cette action sur les phénotypes agit sur l'évolution de la diversité génétique sous-jacente et sur les niveaux de régulation entre les gènes. Ceci nécessite de définir un cadre théorique en passant par l'outil de la modélisation. Le modèle utilisé doit répondre à plusieurs contraintes : il doit permettre de définir un phénotype, des régulations entre les différents locus, de la variabilité entre les individus et l'action de la sélection. Nous avons adapté le modèle développé par Wagner (1996) et Siegal et Bergman (2002) qui permet de répondre à ces contraintes, grâce à son approche individu-centrée qui définit un phénotype à partir des interactions de régulation entre les différents gènes. Il répond donc à notre problématique, car il permet de mesurer et de relier les phénotypes, les niveaux d'expression, les régulations et la diversité génétique.

II.B Impact of selection on genes involved in regulatory network: a modelling study.

Bénédicte Rhoné[*], Jean-Tristan Brandenburg[*], Frédéric Austerlitz

Article publié dans Journal of Evolutionary Biology en 2011 (24:2087-2098)

*Les deux auteurs ont contribué autant au manuscrit.

II.B.i Abstract

Complex phenotypes are often controlled by many interacting genes. One question emerging from such organisation is how selection, acting at the phenotypic level, shapes the evolution of genes involved in regulatory networks controlling the phenotypes. We studied this issue through a matrix model of such networks. In a population submitted to selection, we simulated the evolution of a quantitative trait controlled by a set of loci that regulate each other through positive or negative interactions. Investigating several levels of selection intensity on the trait, we studied the evolution of regulation intensity between the genes and the evolution of the genetic diversity of those genes as an indirect measure of the strength of selection acting on them. We show that an increasing intensity of selection on the phenotype leads to an increased level of regulation between the loci. Moreover, we found that the genes responding more strongly to selection within the network were those evolving toward stronger regulatory action on the other genes and/or those that are the less regulated by the other genes. This observation is strongest for an intermediate level of selection. This may explain why several experimental studies have shown evidence of selection on regulatory genes inside gene networks.

Keywords: gene networks, quantitative traits interactions, selection

II.B.ii Introduction

Complex phenotypes or quantitative traits are generally controlled by many genes interacting through up and down regulatory interactions. Recent progress in molecular genetics has helped to elucidate the complex genetic architecture of quantitative traits. Gene regulatory networks are now well characterized for some traits (for example, flowering time in Arabidopsis thaliana (Ehrenreich et al. 2009; Keurentjes et al. 2007)). One question emerging from such organisation is how selection, acting at the phenotypic level, shapes the evolution of genes involved in the networks controlling the phenotype (Stern & Orgogozo 2008; 2009).

Various approaches based directly on sequence polymorphisms (Fu 1996; Tajima 1989) or on allelic frequencies distributions (Beaumont & Nichols 1996; Goldringer & Bataillon 2004; Vitalis et al. 2001) are now available to detect selection at candidate genes. However, molecular evolutionary analyses that succeed in identifying particular genes targeted by selection have previously focused on one or few genes in isolation (Le Corre 2005; Mullen & Hoekstra 2008; Tishkoff et al. 2007), whereas an increasing number of studies emphasize the necessity to consider genes evolution in the context of their interactions (Chouard 2008; Dalziel et al. 2009; Garfield & Wray 2010). Based on biosynthetic or metabolic pathways organisations, experimental and theoretical studies have tried to make some general predictions about the selective forces acting on genes involved in complex traits (reviewed in Cork & Purugganan 2004): (1) Some authors suggest that genes acting upstream in biosynthetic pathways should be submitted to strong stabilising selection because of the pleiotropic effects of those genes on other pathways, while downstream acting genes, whish should be less constrained, should evolve faster (Rausher et al. 1999)(2) Other studies predict that networks evolve primarily through adaptive changes in enzymes at major pathway intersections. Thus, genes that act at metabolic pathway nodes are the target of selection (Flowers et al. 2007). Those two expectations have been the subject of a vast debate (Jovelin et al. 2009; Livingstone & Anderson 2009; Ramsay et al. 2009; Vitkup et al. 2006; Yang et al. 2009). These predictions are based on biosynthetic or metabolic pathways, which describe the way metabolites derive from one another through the action of enzymes encoded each by a single gene, but provide no direct information on interactions/regulations among genes themselves. A growing number of studies however show that regulatory genes can be the

target of selection (reviewed in Fay & Wittkopp 2008; reviewed in Purugganan 2000; Wray 2007).

A theoretical framework is clearly needed to understand the genetic mechanisms underlying the adaptive response of traits governed by gene regulatory networks. Previous models have studied the evolution of quantitative traits and of the underlying genes coding for these traits (Latta 1998; Le Corre & Kremer 2003). However, they have considered only additive models, in which the phenotypic value of an individual for a given trait is the sum of the allelic values over all genes coding for the trait. They have shown that the responses of traits to selection in terms of phenotypic evolution stem both from changes of allelic frequencies for each gene and from the building of linkage disequilibrium between genes. By considering only an additive model, epistatic interactions are ignored that emerge within a regulation network and should yield dependencies between genes in their response to selection.

In the present work, we propose a simulation study of those processes, using a matrix model of complex regulatory gene network derived from Wagner (1996) and Siegal & Bergman (2002). This model assumes that each locus of the network codes for a product that regulates its own expression and the expression of the other loci inside the network. Evolution takes place through sexual reproduction and mutations that change the intensity of regulation of one gene toward another gene or its level of self-regulation. This model has been previously used to study several important aspects of evolutionary biology as the evolution of robustness and evolvability (Azevedo et al. 2006; Bergman & Siegal 2003; Ciliberti et al. 2007a; Ciliberti et al.; Huerta-Sanchez & Durrett 2007; MacCarthy & Bergman 2007; Siegal & Bergman 2002; Wagner 1996) or the role of network topology (Leclerc 2008; Siegal et al. 2007).

Here, we used a similar scheme to elucidate which genes are the targets of selection within the regulatory network. Investigating various levels of selection intensity, we first evaluated the impact of selection on the phenotypic distribution after evolution. Then, we analyzed the impact of selection at the genotypic level by studying the level of regulation between the different genes of the network, their level of expression and the evolution of their genetic diversity, as an indirect measure of selection acting on them within the network.

II.B.iii Materials and methods.

Constructing the phenotype from the genotype

We used a model derived from Wagner (1996) and modified by Siegal and Bergman (2002) that considers haploid individuals modelled as interaction networks of n genes (see Figure 1 in Siegal & Bergman 2002). The genotype of each individual is a $n \times n$ matrix, $\mathbf{W}=(w_{ij})$, where w_{ij} denotes the effect of the product of gene j on the expression level of gene i. The level of network connectivity is defined by the parameter c, the fraction of nonzero entries in the **W** matrix; (Siegal & Bergman 2002; Wagner 1996). Each row corresponds to a locus: the allele of an individual at a locus i is characterised by the regulatory effects $(w_{ij})_{1 \le j \le n}$ that each gene j exerts on this allele. They can be seen as the entire enhancers of gene i with all regulatory elements that affect the expression of this gene (Wagner, 1996).

The expression level S_i for each gene was deduced from the **W** matrix, by iterating the following equation until equilibrium:

$$S_{\mathrm{i}}(t+1) = f\left(\sum_{j=1}^{n} w_{\mathrm{ij}}\, S_{\mathrm{j}}(t)\right) \quad \text{Eq. II.B-a}$$

where f is a sigmoidal function $f(x)=2/(1 + e^{-x}) - 1$, which constrains the $S_i(t)$ values between −1 and 1. The initial states of genes expression $S_i(0)$ were randomly drawn within the values −1 and 1, with equal probability. If the system (Eq. II.B-a) did not reach equilibrium within 100 iterations, the individual was considered not viable and discarded. This is equivalent to Siegal and Bergman (2002)'s model, in which fitness 0 was assigned to such individuals. Indeed, selection is affecting viability in their model, while in our model it is affecting fecundity (see below). In Siegal and Bergman's model, the probability for an offspring to be retained in the population is its fitness. Thus, an individual who has not reached a stable equilibrium and has a fitness of zero will always be discarded and a new individual will be generated until a suitable one is produced, as population size is also constant in their model. Note also that the rate of rejected individuals never exceeded 1%. The same initial genes expression state $S_i(0)$ was used for all individuals within a simulation through time.

We considered that the phenotype Z of an individual was the sum of all its S_i values. Our model differs from Wagner's (1996) and Siegal and Bergman's (2002) model, which

considered the whole vector of the S_i's as the phenotype. They were considering developmental networks in which each gene had to reach a specific level of expression S_i to reach the optimal phenotype. However, in a quantitative genetics context, the same level for a phenotypic trait can be reached through different combinations of alleles at various loci. For example, this has been shown in lake whitefish (*Coregonus clupeaformis*), where different alleles at different loci have been selected to reach a small size phenotype in different lakes (Campbell & Bernatchez 2004; Rogers & Bernatchez 2005), in rock pocket mice (*Chaetodipus intermedius*) for polymorphism at pelage colour (Hoekstra & Nachman 2003; Nachman et al. 2003) or in *Drosophila* species for pigmentation (Wittkopp et al. 2003).

We assumed an additive model, as it allows that different combination of alleles lead to the same phenotype. It is also the standard model in evolutionary quantitative genetics(Latta 1998; Le Corre & Kremer 2003). Moreover, this model was convenient to model selection in particular, since the natural steady state of the system was $Z = 0$. Indeed, as the w_{ij} and the $S_i(t)$ were drawn in distributions centered on zero and the $S_i(t)$ values were constrained between −1 and 1, the mean standard value expected for Z was zero, which constituted the mean steady state for the system in absence of selection on the phenotype. By increasing the value of Z_{opt}, we could simulate selection towards more extreme phenotypes.

Population dynamics

The model considered a finite population of N individuals that evolved through:

(1) Stabilizing selection: At each generation, each individual was built by drawing its two parents in the previous generation. Each parent was chosen at random with replacement according to its fitness function $F(Z)$, which depended on its phenotype Z, as follows:

$$F(Z) = \exp\left(- \frac{\left(Z - Z_{opt}\right)^2}{\omega^2} \right), \quad \text{Eq. II.B-b}$$

where Z_{opt} is the optimum phenotypic value and $1/\omega$ denotes the intensity of selection, which corresponds to the classical stabilizing selection model (e.g. Latta 1998; Le Corre & Kremer 2003; Siegal & Bergman 2002)

(2) Reproduction: Because the model was haploid, new offspring received for each locus (i.e. a given row of the matrix, see above) either the allele of its first parent with

probability 0.5 or the allele from its second parent with probability 0.5. All loci were assumed to be independent: the parental allele was drawn independently for each locus.

(3) Mutation: Alleles could be transmitted unchanged or mutated. When a mutation occurred, it changed at random one of the w_{ij}'s, for which the new value was drawn in a normal distribution N(0,1). The individual was then tested with by iterating Eq. (II.B-a) 100 times. If it was viable (i.e. the S_i values had reached a stable equilibrium), it was kept in the population.

Parameters

The gene regulatory network was composed of five loci (n = 5). We considered here highly connected networks where c is equal to 1.0. We started all simulations with a population composed of 500 copies of the same individual. This individual was generated by drawing its w_{ij} coefficients into a normal distribution N(0,1). Its viability was tested by iterating Eq. II.B-b 100 times, this procedure being repeated until a stable individual was created. Then we let the population evolve through the process of mutation, sexual reproduction and selection for 10,000 generations to reach mutation-drift-selection equilibrium. Total mutation rate for the n genes of each individual was 0.01, meaning that the rate per locus was $0.01/n$ and the rate per value of the **W** matrix of $0.01/n^2$. We studied the impact of the intensity of selection ($1/\omega$), with values ranging from 0.01 (almost neutral case) to 25 (very strong selection) and explored different optimal phenotypic value (Z_{opt}) ranging from 0 (mean standard state for the system) to 4 (increasing expression level for the genes networks compared to the mean standard state). We also considered a model without selection for an optimum, referred hereafter as the "neutral model". In such model, all the individuals of the population had same fitness and the population evolved through drift and mutation.

Model outputs

We first investigated the impact of stabilising selection on the phenotypic evolution. To this end, we computed the mean phenotypic value ($\bar{Z}$) in the population by averaging the Z values over all individuals and calculated the standard deviation sd(Z) of this value across all individuals in the population. We also computed for each locus the mean level of its expression ($\bar{S}_i$) across individuals in the population and the standard deviation sd ($\bar{S}_i$) across loci of this quantity, as a measure of the level of heterogeneity of expression across loci.

We next investigated the impact of selection on genotypic evolution. For each locus i, we computed the average level of regulation of the locus on the other loci $\overline{w_{.i}}$ and the average level of regulation of the locus by the other loci $\overline{w_{i.}}$. $\overline{w_{.i}}$ (resp. $\overline{w_{i.}}$) was obtained by averaging w_{ij} (resp. w_{ji}) over all individuals and all locus j, with $j \neq i$. We also computed the average level of regulation within the population ($\overline{w_{..}}$) by averaging the w_{ij}'s over all individuals and all loci i and j, with $i \neq j$, and the average level of self-regulation of all loci ($\overline{w_{ii}}$) by computing the average w_{ii} value over all loci and individuals.

Finally, we estimated the average within-population diversity (H_{Si}) for each locus i using the standard formula $H_{si=}\ 1-\sum_{k=1}^{n_{ai}} p_{ik}^{\ 2}$, where p_{ik} denotes the allelic frequency of allele k at locus i, and n_{ai} the total number of alleles at this locus (this number varied through time, as we assumed an infinite-allele model). We also computed the mean diversity value (H_S) and its standard deviation (sd(H_{Si})) across all loci. Then, we computed the pairwise correlation coefficients within any of the quantities H_{Si}, $\overline{w_{i.}}$,$\overline{w_{.i}}$, $\overline{w_{ii}}$ and $\bar{S}_i$ using the loci as repetitions. We performed 500 replicates per parameter set. In each replicate, to reduce the noise, we averaged all values of the parameters and of their pairwise correlation coefficients over the last 1,000 generations. These values were averaged again over all replicates of a given parameter set.

II.B.iv Results.

Influence of selection on the phenotypic evolution:

The average phenotypic value ($\bar{Z}$) generally converged toward the optimal Z_{opt} value (See the result for Z_{opt}=4, Figure II.B-1) in almost all cases after 10,000 generations. The rate of convergence toward the optimal value depended on selection intensity. Convergence was fastest for an intermediate intensity ($1/\omega = 1$). In contrast, $\bar{Z}$ showed only a very limited increase for the lowest selection intensity ($1/\omega = 0.01$). The same was seen also for the highest selection intensity ($1/\omega = 25$) when selecting for the highest optimal phenotypic value (Z_{opt}=4). In the following, we always consider the values of all parameters after these 10,000 generations (which we will call equilibrium values for simplicity), as other parameters had also reached a stable value in most case after 10,000 generations (see the case of H_S in Figure II.B-S1). Thus, we will not consider the cases of $1/\omega = 25$ and Z_{opt}=4, as the simulations had not reached equilibrium under these settings.

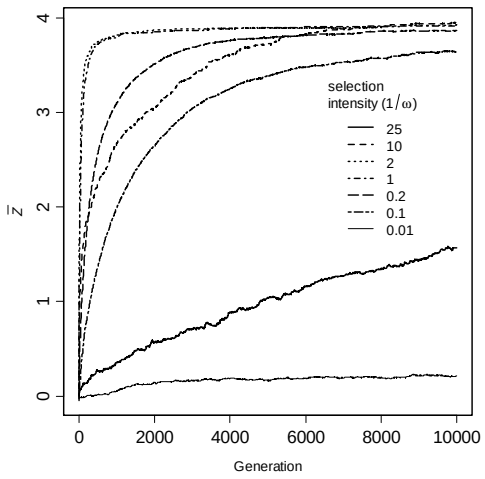


Figure II.B-1: Evolution of the phenotypic value ($\bar{Z}$) through the 10,000 generations as a function of selection intensity ($1/\omega$), for an optimal phenotypic value (Z_{opt}) of 4.

In most cases, the equilibrium $\bar{Z}$ values were very close to the optimal values whatever the Z_{opt} value, provided that selection was strong enough ($1/\omega \geq 0.2$). Conversely, for the lowest intensity of selection ($1/\omega = 0.01$), the $\bar{Z}$ values were close to zero, *i.e.* the level expected in absence of phenotypic selection (Figure II.B-2a.). The standard deviation sd(Z) among individuals at equilibrium (Figure II.B-2b), reflecting the within-population phenotypic variability, decreased with $1/\omega$ for Z_{opt}=0, while for the other Z_{opt} values it first increased toward a maximum reached for $1/\omega = 0.1$, and then decreased to reach a very low value for a high intensity of selection ($1/\omega \geq 1$).

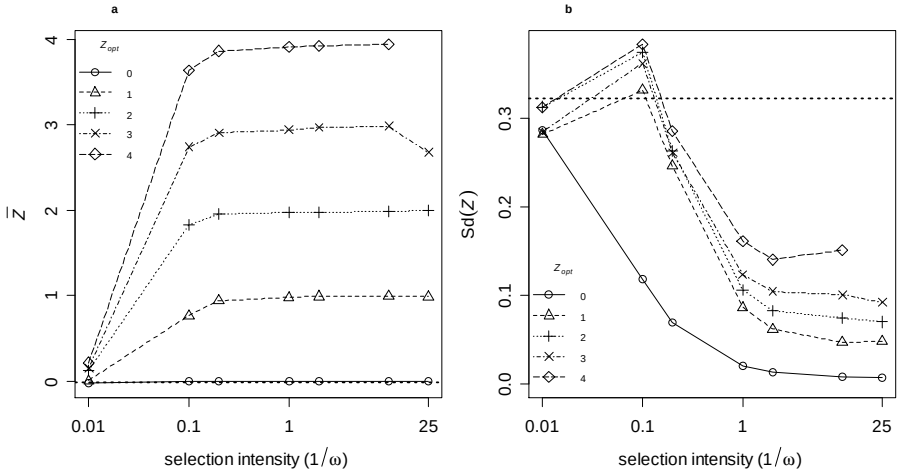


Figure II.B-2: Mean phenotypic value ($\bar{Z}$) (**a**) and standard deviation (sd(Z)) among individuals at equilibrium (**b**) after 10,000 generations of evolution, as a function of selection intensity ($1/\omega$), for several optimal phenotypic values (Z_{opt}). The values corresponding to $Z_{opt} = 4$ and $1/\omega = 25$ are not shown, as simulations did not converge in this case. The dotted line corresponds to the neutral case.

Effect of selection on the genotypic evolution:

The average level of cross-regulation between loci ($\overline{w..}$) was slightly negative whatever the intensity of selection (1/ω) for low Z_{opt} values (Z_{opt} = 0 or 1, Figure II.B-3a.). However, for higher Z_{opt} values, $\overline{w..}$ increased with 1/ω. The rate of increase was larger as Z_{opt} increased, making that $\overline{w..}$ reached values around 0.5 for Z_{opt} = 4. Thus, selection for high gene expression levels led to an increase in the average levels of cross-regulation between loci. The average level of self-regulation ($\overline{w_{ii}}$, Figure II.B-3b) was positive whatever the Z_{opt} value. It decreased with 1/ω for Z_{opt} = 0, while the opposite trend was observed for $Z_{opt} \geq 1$. The rate of increase was again larger as Z_{opt} increased, and the values reached by $\overline{w_{ii}}$ were higher then the values reached by $\overline{w..}$.

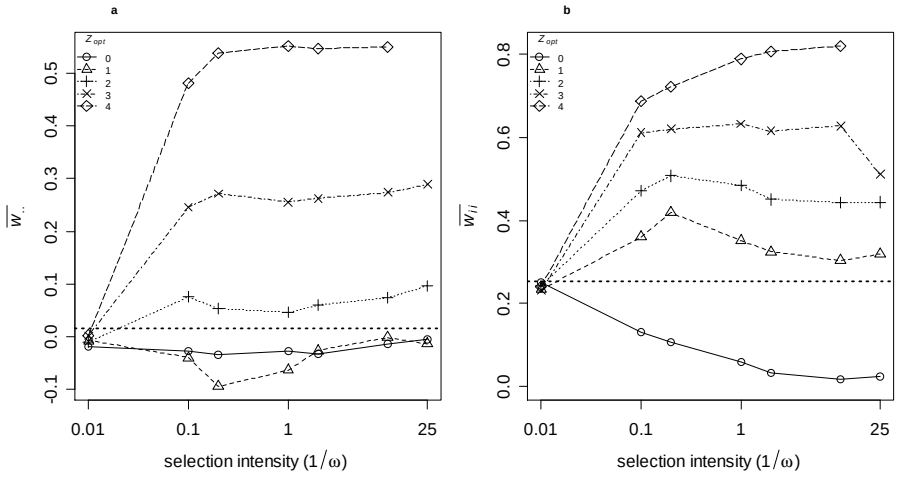


Figure II.B-3: Mean level of cross-regulation between loci $\overline{w..}$, **a.**) and mean level of self-regulation between loci ($\overline{w_{ii}}$, **b.**) after 10,000 generations of evolution, as a function of selection intensity (1/ω), for several optimal phenotypic values (Z_{opt}). The values corresponding to Z_{opt} = 4 and 1/ω = 25 are not shown, as simulations did not converge in this case. The dotted line corresponds to the neutral case.

Whatever the Z_{opt} value considered, the other main effect of selection at the genotypic level was that genetic diversity observed at equilibrium decreased as selection intensity increased (Figure II.B-4a.). While this tendency was limited for Z_{opt} = 0, it was very sharp otherwise, decreasing from ~0.6 (the neutral value) for 1/ω = 0.01 to ~0.1 for 1/ω = 10. The standard deviation of genetic diversity across loci (sd(H_{Si})) remained almost constant as 1/ω increased for Z_{opt} = 0, while for all other Z_{opt} values it increased first to decrease then afterwards (Figure II.B-4b). This shows that for moderate selection intensity, the standard deviation across loci increased, while the mean genetic diversity decreased; i.e. the effect of selection became very heterogeneous across loci.

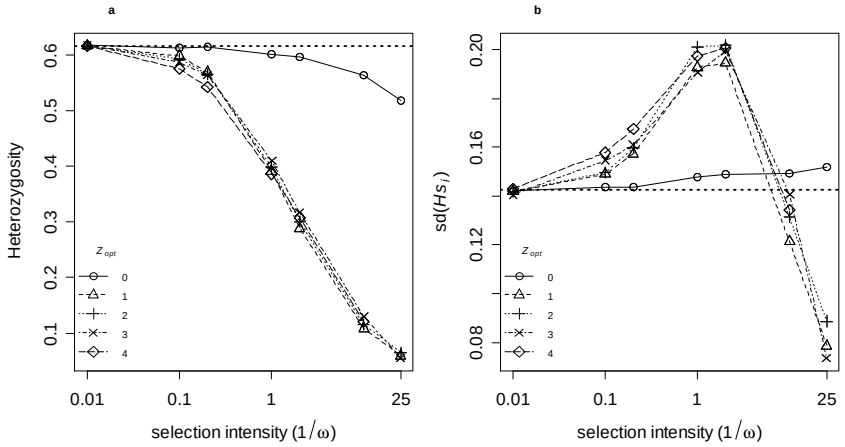


Figure II.B-4: **a.** Average within-population diversity (H_S) and **b.** standard deviation of H_S among loci after 10,000 generations of evolution, as a function of selection intensity (1/ω), for several optimal phenotypic values (Z_{opt}). The values corresponding to $Z_{opt} = 4$ and $1/\omega = 25$ are not shown, as simulations did not converge in this case. The dotted line corresponds to the neutral case.

To explain such heterogeneity, we computed both the correlation between the diversity of a given gene (H_{Si}) and the average level of regulation of this gene on the other genes ($\overline{w_{i.}}$, Figure II.B-5), and the correlation between the diversity of a given gene and the average level of regulation of this gene by the other genes ($\overline{w_{i.}}$, Figure II.B-6). In the two cases, we found no correlation for $Z_{opt} = 0$, whatever the selection intensity. Negative correlation between H_{Si} and $\overline{w_{.i}}$ was found for intermediate values of Z_{opt} (1 or 2) (Figure II.B-5a). Starting from zero for low 1/ω values, this correlation became more strongly negative as 1/ω increased and then increased back again: the most negative correlations were found for intermediate values of selection intensity ($1 < 1/\omega < 10$). For such values, the diversity at a given locus showed a clear negative relation with the intensity of regulation that it exerted on the other loci (Figure II.B-5c). For low level of selection intensity ($1/\omega < 1$), all the network genes had identical high level of diversity (Figure II.B-5b), whereas for the highest level of selection investigated ($1/\omega$=25), all the network genes had same low level of diversity (Figure II.B-5d). By contrast, the correlation between H_{Si} and $\overline{w_{.i}}$ was close to zero for all values of 1/ω, without a clear trend for the higher Z_{opt} values (3 or 4) (Figure II.B-5a).

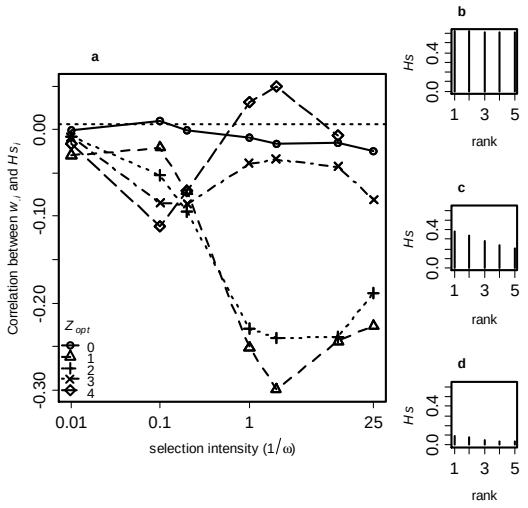


Figure II.B-5: **a.** Mean correlation over simulations between the level of within-population diversity of a given gene i (H_{Si}) and its level of regulation on the other genes network ($\overline{w}_{.i}$) after 10,000 generations of evolution, as a function of selection intensity (1/ω), for several optimal phenotypic values (Z_{opt}).. The values corresponding to Z_{opt} = 4 and 1/ω = 25 are not shown, as simulations did not converge in this case. **b., c. and d.:** Mean diversity of the loci as a function of their rank in the regulatory network (locus with rank 1 having the lowest mean value $\overline{w}_{.i}$ of regulation on the other genes), assuming Z_{opt} = 1, for weak selection intensity (1/ω = 0.1) (**b**.), intermediate selection intensity (1/ω = 2) (**c**.) or strong selection intensity (1/ω = 25) (**d.**), after 10,000 generations of evolution. The dotted line corresponds to the neutral case.

The pattern was quite different for the correlation between H_{Si} and the average level of regulation ($\overline{w_{i.}}$) of a gene by the other genes (Figure II.B-6). No correlation was found for intermediate Z_{opt} values (1 or 2), whereas positive correlation between H_{Si} and $\overline{w_i}$. was found for the higher Z_{opt} values (3 or 4); i.e. the less a gene was regulated, the more diversity it showed. Starting from a null value for 1/ω = 0.01, this correlation increased with 1/ω to reach values ~0.4 for 1/ω = 5 and then decreased again.

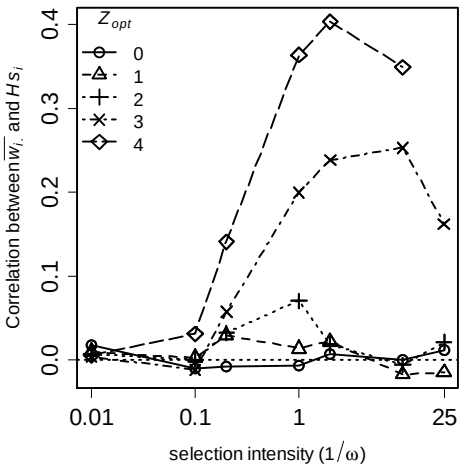


Figure II.B-6: Mean correlation over simulations between the level of within-population diversity of a given gene i (H_{Si}) and its level of regulation by the other genes network ($\overline{w_i}$.) after 10,000 generations of evolution, as a function of selection intensity (1/ω), for several optimal phenotypic values (Z_{opt}). The values corresponding to Z_{opt} = 4 and 1/ω = 25 are not shown, as simulations did not converge in that case. The dotted line corresponds to the neutral case.

We also investigated the level of heterogeneity in the response of the different loci by computing the standard deviation (sd($\overline{S_i}$)) of the level of expression ($\overline{S_i}$) across loci (Figure II.B-7). We found that for Z_{opt} between 1 and 3, this standard deviation increased with 1/ω until it reached a maximum value for 1/ω = 0.5, and decreased slightly afterwards. Conversely this standard deviation decreased with 1/ω for Z_{opt} = 0 or remained constant whatever the 1/ω value for Z_{opt} = 4.

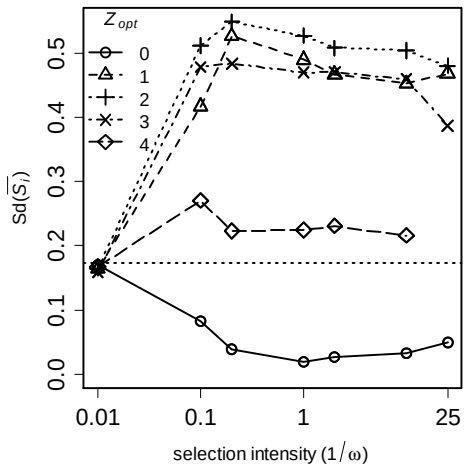


Figure II.B-7: Mean standard deviation between the level expression of locus ($\overline{S_i}$) for each individual after 10,000 generations of evolution, as a function of selection intensity (1/ω), for several values of optimal phenotypic distance (Z_{opt}). The values corresponding to Z_{opt} = 4 and 1/ω = 25 are not shown, as simulations did not converge in that case. The dotted line corresponds to the neutral case.

The correlation between expression level of a gene ($\overline{S_i}$) and its within-population diversity (H_{Si}) was also investigated (Figure II.B-8). A correlation level close to zero was

found for $Z_{opt} = 0$, whatever the intensity of selection. This correlation tended to be slightly negative for $Z_{opt} = 1$ or 2, meaning that the more a gene was expressed, the less diverse it was, while the opposite was true for $Z_{opt} = 3$ or 4. Finally, except for $Z_{opt} = 0$, we observed a high correlation between the level of regulation ($\overline{w}_{.i}$) that a gene exerts on the other genes and the level of expression ($\overline{S_i}$) of this gene (Figure II.B-9). In other words, the stronger the regulatory effect of a gene, the more it is expressed.

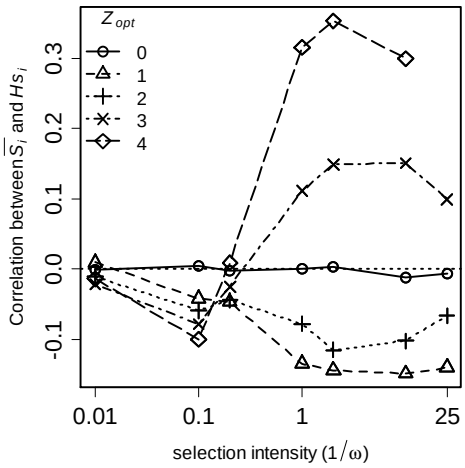


Figure II.B-8: Mean correlation over simulations between the level of within-population diversity of a given gene *i* (H_{Si}) and its expression level ($\overline{S_i}$) after 10,000 generations of evolution, as a function of selection intensity (1/ω), for several values of optimal phenotypic distance (Z_{opt}). The values corresponding to $Z_{opt} = 4$ and $1/\omega = 25$ are not shown, as simulations did not converge in that case. The dotted line corresponds to the neutral case.

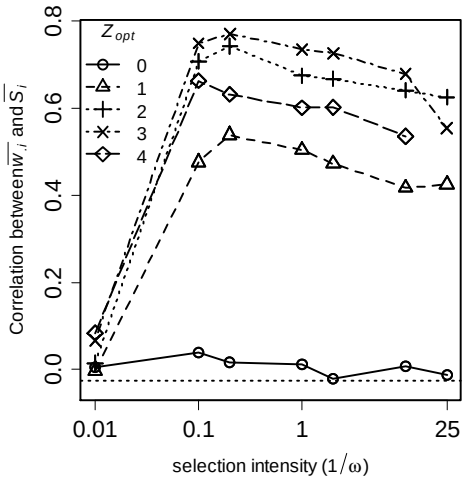


Figure II.B-9: Mean correlation over simulations between the level of regulation on the other genes network ($\overline{w}_{.i}$) and its expression level ($\overline{S_i}$) after 10,000 generations of evolution, as a function of selection intensity (1/ω), for several values of optimal phenotypic distance (Z_{opt}). The values corresponding to $Z_{opt} = 4$ and $1/\omega = 25$ are not shown, as simulations did not converge in that case. The dotted line corresponds to the neutral case.

II.B.v Discussion

Evolution of regulatory networks

Using a matrix model of regulatory networks, our simulation study provides new insights on the evolution of genes involved in complex regulatory networks under selection. We clearly showed that the response of gene networks to selection largely depends on the two parameters characterizing stabilizing selection in the model: the selection intensity ($1/\omega$), i.e. the narrowness of the fitness curve, and the optimal phenotypic value (Z_{opt}), i.e. the phenotypic value with maximal fitness. While increasing stabilizing selection intensity led to a drastic reduction of both phenotypic and genetic diversity (see also Kingsolver & Pfennig 2007), the impact of the optimal phenotypic value on the genes network evolution appeared to

be more complex. Three distinct cases must be considered, depending whether the optimal phenotypic value Z_{opt} was null, low or high.

When $Z_{opt} = 0$, selection favoured individuals with a phenotypic value close to the mean natural steady state of the system (see methods). As a consequence, this optimal phenotypic value was always reached, and reaching the optimal gene expression level did not constitute a selective pressure acting on the gene network during evolution. Genes were only constrained in the sense that alleles that would cause the phenotype Z to deviate from zero were eliminated. Several results were specific to this case. In particular, an increase in selection intensity led only to a moderate decrease of genetic diversity and did not affect the level of regulation between loci.

When $Z_{opt} > 0$, individuals with phenotypic value that deviated from the null value expected in absence of phenotypic selection were favoured. Individuals with a higher level of gene expression ($\bar{S}_i$) were selected. Thus, while for $Z_{opt} = 0$, individuals were only selected to a stabilising selection around the natural steady state of the system ($Z = 0$), reaching an optimal phenotypic value Z_{opt} different from zero constituted another level of selective pressure, as only individuals diverging from this standard steady state were selected. This additional level of selection is exemplified by the drastic reduction of genetic diversity observed as 1/ω increased. This type of selection had several impacts on evolution. First, we observed selection towards an increase in the average level of cross-regulation between loci ($\bar{w}_{..}$). Thus, the increase of gene level expression to reach the optimal phenotype was obtained by increasing the level of positive regulation between genes. Second, we found that loci were not affected with the same intensity by selection. The level of genetic diversity, which characterises the intensity of selection affecting those genes, was directly related to their capacity to regulate ($w_{.i}$) or to be regulated by the other genes ($w_{i.}$) in the network. This relation depended clearly on the optimal phenotypic value.

Indeed, in cases of selection for intermediate Z_{opt} values ($Z_{opt} = 1$ or 2), the negative correlation found between H_{Si} and $w_{.i}$ (Figure II.B-5) indicated that the more a locus regulated the others, the lower its genetic diversity, i.e. the more it was exposed to selection. In other words, stronger selection affected the more regulatory genes within the network. By contrast, in cases of selection for high Z_{opt} values ($Z_{opt} = 3$ or 4), the positive correlation found between H_{Si} and $w_{i.}$ (Figure II.B-6) indicated that selection affected the less regulated genes within the network.

This difference in the mechanism of selection can be understood as follows. When Z_{opt} = 1 or 2, this optimal phenotypic value could be reached either by moderately expressing all genes of the network (intermediate $\bar{S}_i$ values), or by strongly expressing one or two genes within the network ($\bar{S}_i$ values ~1.0 for these genes and ~0.0 for the other genes). The high variance between the expression level ($\bar{S}_i$) among loci observed for Z_{opt} = 1 or 2 (Figure II.B-7) and the fact that selection acts more on the genes with a high $\bar{S}_i$ value (negative correlation between the expression level of a gene and its genetic diversity after evolution, Figure II.B-8) show that the second hypothesis is correct: a few genes are strongly expressed and therefore are the ones targeted more strongly by selection. Going more into details, it is interesting to note that the genes that are the more expressed are also the more regulatory genes, *i.e.* the genes which have the higher impact on the other genes in the network (since the level of expression $\bar{S}_i$ of a given gene and the intensity of the regulation it exerts on the other genes are also positively correlated, see Figure II.B-9). Thus the system evolves into a dynamics where a few genes are strongly expressed and also strongly regulate the other genes. The optimal phenotypic value is reached through the expression of genes that simultaneously exert a control on other genes, keeping them at a lower level of expression to avoid any deviation from this optimal phenotype.

Selection for more regulatory genes reflects selection for genes that have the highest effect on the phenotype, given that these control the phenotype through a cascade of regulations. Therefore any mutation that will modify even slightly one of the w_{ij}'s for these genes is likely to modify strongly the phenotype. Indeed such a mutation can either affect the own expression of the gene through the w_{ii} coefficients or the expression of the other genes through the $w_{ij,\ i\neq j}$ coefficients. Thus any mutation on this gene is likely to be counterselected as it will strongly affect the phenotype. Conversely a mutation on a less regulatory gene will have less of an effect on the phenotype, as it can only affect its own expression and not much the expression of the other genes. Thus, more polymorphism can be maintained at equilibrium for such genes.

The situation was quite different when selecting for high Z_{opt} values (3 or 4). In this case, we did not observe that selection was stronger for the genes with the highest regulatory action ($w_{.i}$). The more regulatory genes were still the more highly expressed (positive correlation between $\bar{S}_i$ and $w_{.i}$, Figure II.B-9), however even the less regulatory genes needed to be expressed to some extent in order to reach the optimal phenotypic value. Indeed, we found lower variances for the level of expression $\bar{S}_i$ in that case, especially for Z_{opt} = 4 (Figure

II.B-7). Thus, selection had to be exerted on the genes independently of their regulatory effect, so that they reached the appropriate level of expression.

In this situation, selection acted more on the less regulated genes in the network, which may be a consequence of the level of regulatory interactions between the different genes becoming quite high so that they could reach these high levels of expression. Under these conditions, the expression of the most regulated genes became completely controlled by the regulation imposed by the other genes, and so selection could not act directly on them. Selection could thus just act only on the less regulated genes, since as they were less controlled by the regulatory interactions, only specific allelic forms for these genes could be consistent with the level of expression needed to reach the optimal phenotypic value.

Biological relevance

Scenarios with intermediate Z_{opt} values (Z_{opt} = 1 or 2) correspond to situations where a species is in an environment that is rather similar to the environment in which it has evolved for a long time, while the stronger Z_{opt} values correspond to species exposed to a novel environment that is very different from its previous environment. This corresponds for example to species submitted to a substantial climatic change or invading a new territory strongly different from their original habitat, as a result of an ecological expansion or an accidental transplantation. It is interesting to note that these different environmental conditions lead to quite different responses of the genes within the network. It might be interesting in that context to perform experimental studies comparing populations in moderately or highly stressful environments, for example at the centre vs. the edge of the species range.

Notice however that whatever the Z_{opt} value (provided Z_{opt} >0), the correlations found between genetic diversity and gene regulation intensity were maximal for intermediate values of selection intensity ($1<1/\omega<10$). Indeed, under very low selection intensity, the genes of the network were poorly affected by selection and the genetic diversities of the different loci were distributed around the mutation/drift equilibrium. Under very high selection, all genes of the network responded strongly to selection. By contrast, for intermediate intensities of selection ($1/\omega$ around 1.0), all genes did not reply similarly to selection, and depending upon Z_{opt}, either the more regulatory genes or the less regulated ones responded the more to selection. These high correlations found for intermediate values of selection intensity, tend to show that

selection should particularly target genes that are rather basal in the regulatory network, in the sense that they regulate many genes but are not themselves much regulated by other genes.

These results were found to be robust to several parameters of the model. For example, we found that increasing the network size had no effect on the level of correlations between genetic diversity and gene regulation intensity (Figure II.B-S2). Increasing the genetic variability (mutation rate or population size) reinforced the correlations (Figure II.B-S3, II.B-S4). By contrast, decreasing the connectivity level between the genes of the network (*c* parameter, see material and methods) tended to reduce these correlations to some extent (Figure II.B-S5, II.B-S6, II.B-S7), but they were still observed for a connectivity of 0.4 (two interactions per gene on average, see Figure II.B-S5). This last result is particularly important since Leclerc (2008) point out that gene networks are generally characterized by low levels of connectivity between network genes, allowing a mean of 1.5 to 2 direct interactions per genes within networks.

Our results contrast to some extent with those of Siegal *et al.* (2007). Indeed, investigating the importance of network topology in gene network evolution, they found no link between the level of polymorphism of a gene and its connectivity within regulatory networks, in contradiction with their expectation of finding lower polymorphism at the more connected genes. However, they did not investigate the impact of the deviation from the standard state of the system ($Z = 0$), and we have shown here that results are quite dependent upon the Z_{opt} value. Indeed, performing simulations under their model (selection for a specific value of each S_i), we found no correlation between the genetic diversity of a gene and the average level of regulation of a gene on the other genes or the level of regulation of this gene by the other genes (results not shown). Moreover, they measured the connectivity of a given locus as the number of loci with which it interacts. Thus, they did not consider the regulatory intensity involved in these connections. A gene that is connected to many loci but only through weak regulation effects (low w_{ij} values) will not be strongly exposed to selection, since it does not in practice have a strong effect on the expression of the products of the different genes. This may also explain why no significant relation was observed between connectivity and the level of selection in yeast (Evangelisti & Wagner 2004), as only the number of connection was considered in this empirical study.

Thus, the specificity of our study is that we have shown that selection will target more strongly basal genes within the evolutionary network. This is consistent with several

experimental studies that show that selection targets in particular regulatory genes (for a review, see Carroll 2000; Fay & Wittkopp 2008; Purugganan 2000; Wray 2007). For instance, Zhao *et al.* (2008) found some evidence that regulatory loci were more targeted by selection than other loci during the domestication process of maize. A similar process was observed in *Ipomea*, for which the adaptive response of flower colour is also controlled mainly by regulatory loci (Durbin et al. 2003). This is also the case of Hox genes in many animal species (for a review, see Carroll 2000), as with the gene *Ultrabithorax* in *Drosophila*, which has a strong impact on the evolution of morphological traits (Stern 1998). Similarly, *cis*-acting regulatory mutations in the gene *Pitx1* have an adaptive impact by causing a pelvic reduction in threespine stickleback fish (Shapiro et al. 2004). A similar pattern is observed the gene *tb1* in maize (Wang et al. 1999) as a consequence of the domestication process. Selection is also inferred in humans also on regulatory regions, for instance on the coagulation factor VII (Hahn et al. 2004) and on putative regulatory loci affecting the response to climate adaptation (Hancock et al.; Sun et al.).

Genome scans (Beaumont 2005; Luikart et al. 2003), which detect genes that shows an excess of differentiation between populations of a given species, offer a possibility to detect which genes are under adaptive selection. Several of these studies have shown that regulatory genes are the particular targets of selection. This was observed for regulatory genes involved in various developmental processes in *Picea* species (Namroud et al. 2010), in drought or temperature responses in black spruce (Prunier et al.), in growth in lake whitefish (Campbell & Bernatchez 2004; Rogers & Bernatchez 2005). Selection was also demonstrated for several genes in the gene network involved in floral development in *Arabidopsis thaliana* (Le Corre 2005; Olsen et al. 2002). In this context, studying 52 flowering time genes of *Arabidopsis*, Flowers et *al.* (2009) found traces of selection for several regulatory genes.

Thus, our model accounts well for the observation of selection at regulatory loci in many cases. However, a few studies showed selection on highly regulated genes rather than on regulatory loci (see Cork & Purugganan 2004). This was shown for instance for flowering time selection in wheat populations (Rhone et al. 2010). This difference with our model may be due to specific constraints on the genes of this network, connected in particular with pleiotropic interactions. As pointed out by Cork & Purugganan (2004), it should indeed be more difficult for a gene involved in several networks controlling different traits to be submitted to strong positive selection. Theoretical studies of gene networks controlling

several traits, including pleiotropic genes involved in several networks, will clearly be needed in the future.

Acknowledgments

We thank V. Orgogozo for helpful suggestions on the manuscript, J. Guglielmini for his implication on a preliminary version of this work and F. Palstra for his help with English usage. We thank the editor and two anonymous reviewers for their helpful comments and suggestions. This work was partially financed by an Agence Nationale de la Recherche grant (NUTGENEVOL, programme blanc 07-BLAN-0064) to FA. BR was supported by an "Attaché Temporaire d'Enseignement et de Recherche" grant from the French "Ministère de l'Enseignement Supérieur et de la Recherche". JTB was supported by a PhD grant from the French "Ministère de l'Enseignement Supérieur et de la Recherche". Simulations were run on the Linux cluster of the "Muséum National d'Histoire Naturelle".

II.B.vi Supplementary figures

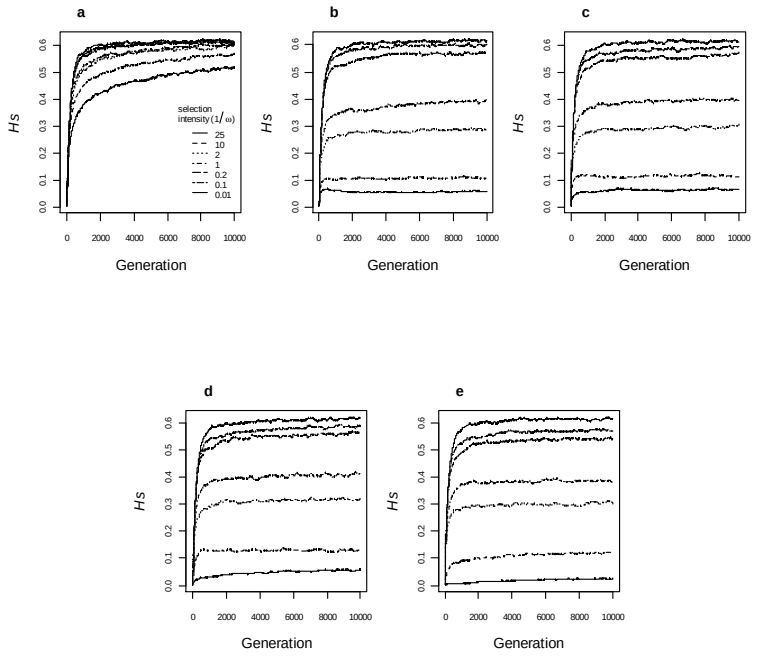


Figure II.B-S1: Evolution of genetic diversity (H_S) through the 10,000 generations as a function of selection intensity ($1/\omega$), for an optimal phenotypic value (Z_{opt}) of 0 (a), 1 (b), 2 (c), 3 (d) or 4 (e).

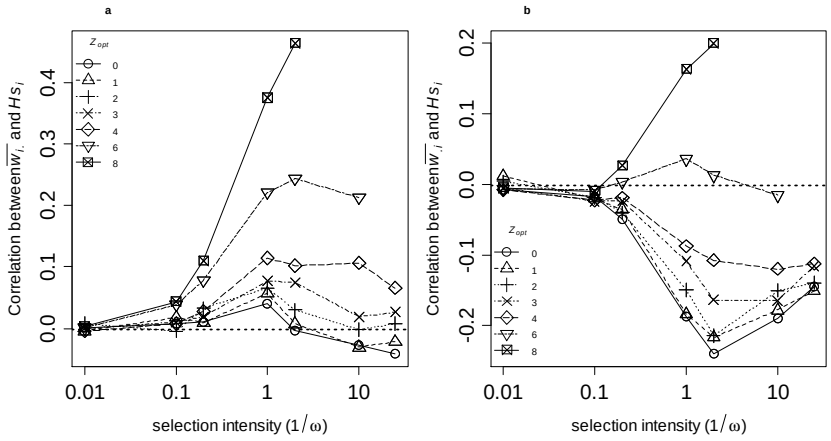


Figure II.B-S2: Results for a gene network of 10 loci and a mutation rate of 0.02. **a.** Mean correlation over simulations between the level of within-population diversity of a given gene *i* (H_{Si}) and its level of regulation by the other genes network ($\overline{w}_{i.}$). **b.** Mean correlation over simulations between the level of within-population diversity of a given gene i (H_{Si}) and its level of regulation on the other genes network ($\overline{w}_{.i}$). In both cases, the values are the final values after 10,000 generations of evolution. They are given as a function of selection intensity (1/ω), for several values of optimal phenotypic distance (Z_{opt}). The values for which simulations did not converge are not shown. The dotted line corresponds to the neutral case.

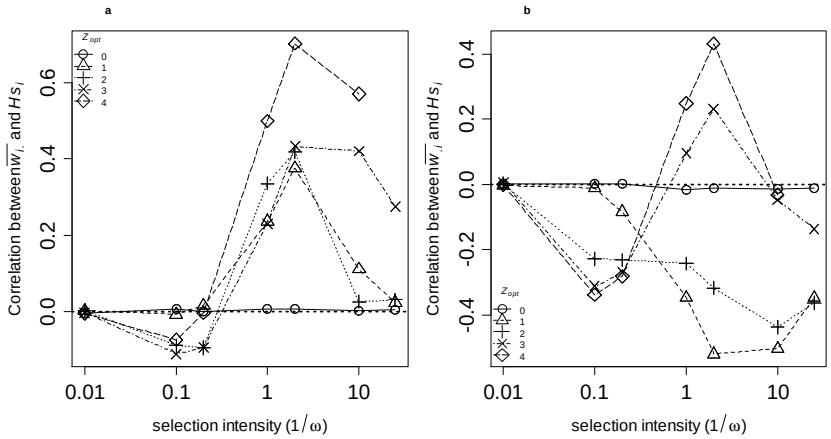


Figure II.B-S3: Results for an increased mutation rate (μ = 0.1). **a.:** Mean correlation over simulations between the level of within-population diversity of a given gene i (H_{Si}) and its level of regulation by the other genes network ($\overline{w}_{i.}$). **b.** Mean correlation over simulations between the level of within-population diversity of a given gene i (H_{Si}) and its level of regulation on the other genes network ($\overline{w}_{.i}$) after 10,000 generations of evolution. In both cases, the values are the final values after 10,000 generations of evolution. They are given as a function of selection intensity (1/ω), for several values of optimal phenotypic distance (Z_{opt}). The values corresponding to Z_{opt} = 4 and 1/ω = 25 are not shown, as simulations did not converge in that case. The dotted line corresponds to the neutral case.

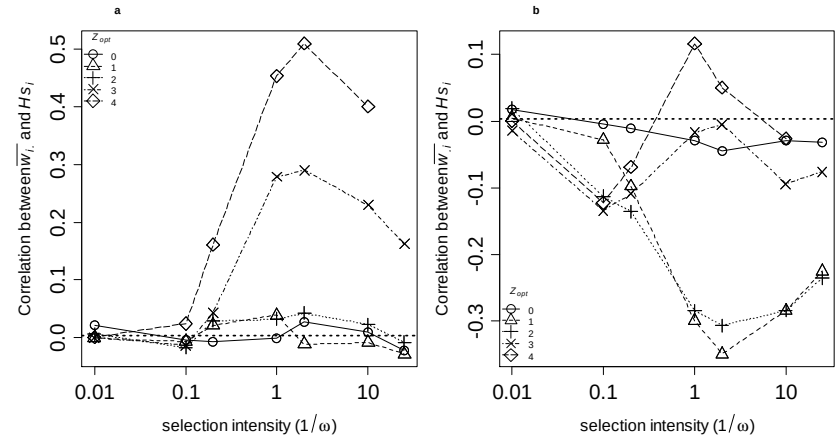


Figure II.B-S4: Results for a population size (N) of 1,000 individuals. **a.** Mean correlation over simulations between the level of within-population diversity of a given gene i (H_{Si}) and its level of regulation by the other genes network ($\overline{w}_{i.}$). **b.** Mean correlation over simulations between the level of within-population diversity of a given gene i (H_{Si}) and its level of regulation on the other genes network ($\overline{w}_{.i}$). In both cases, the values are the final values after 10,000 generations of evolution. They are given as a function of selection intensity ($1/\omega$), for several values of optimal phenotypic distance (Z_{opt}). The values corresponding to $Z_{opt} = 4$ and $\omega^2 = 25$ are not shown, as simulations did not converge in that case. The dotted line corresponds to the neutral case.

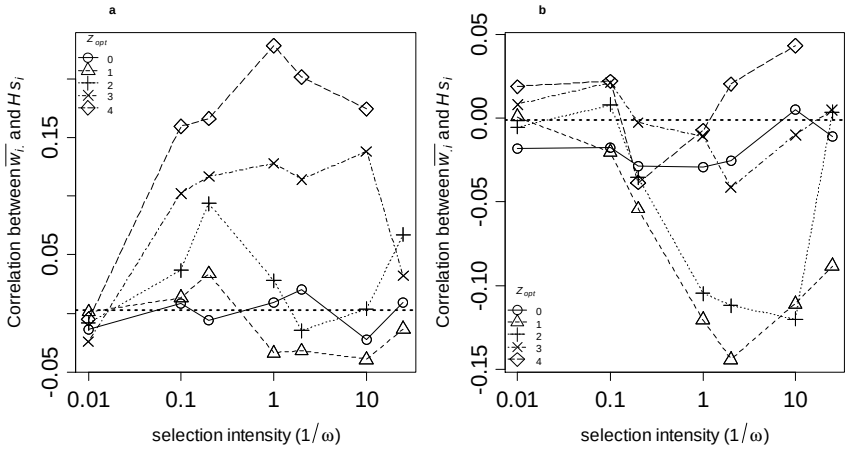


Figure II.B-S5: Results for a connectivity (c) of 0.4. **a.** Mean correlation over simulations between the level of within-population diversity of a given gene i (H_{Si}) and its level of regulation by the other genes network ($\overline{w}_{i.}$). **b.** Mean correlation over simulations between the level of within-population diversity of a given gene i (H_{Si}) and its level of regulation on the other genes network ($\overline{w}_{.i}$). In both cases, the values are the final values after 10,000 generations of evolution. They are given as a function of selection intensity ($1/\omega$), for several values of optimal phenotypic distance (Z_{opt}). The values corresponding to $Z_{opt} = 4$ and $\omega^2 = 25$ are not shown, as simulations did not converge in that case. The dotted line corresponds to the neutral case.

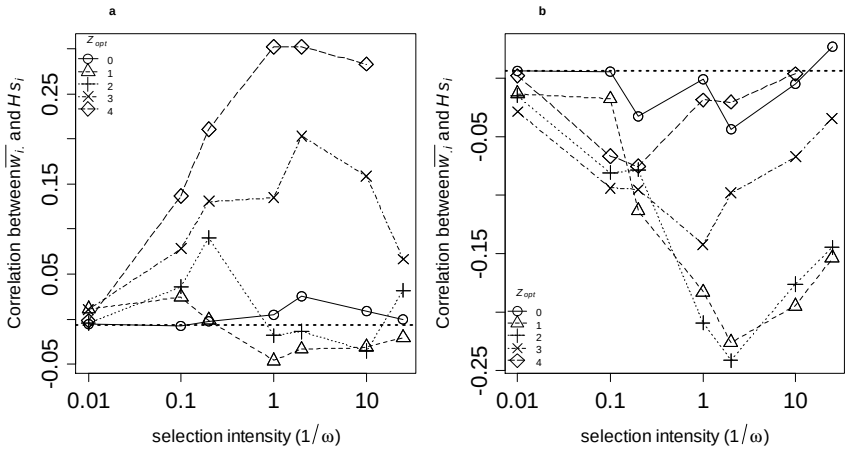


Figure II.B-S6: Results for a connectivity (*c*) of 0.6. **a.** Mean correlation over simulations between the level of within-population diversity of a given gene i (H_{Si}) and its level of regulation by the other genes network ($\overline{w}_{i.}$). **b.** Mean correlation over simulations between the level of within-population diversity of a given gene *i* (H_{Si}) and its level of regulation on the other genes network ($\overline{w}_{.i}$). In both cases, the values are the final values after 10,000 generations of evolution. They are given as a function of selection intensity (1/ω), for several values of optimal phenotypic distance (Z_{opt}). The values corresponding to $Z_{opt} = 4$ and $\omega^2 = 25$ are not shown, as simulations did not converge in that case. The dotted line corresponds to the neutral case.

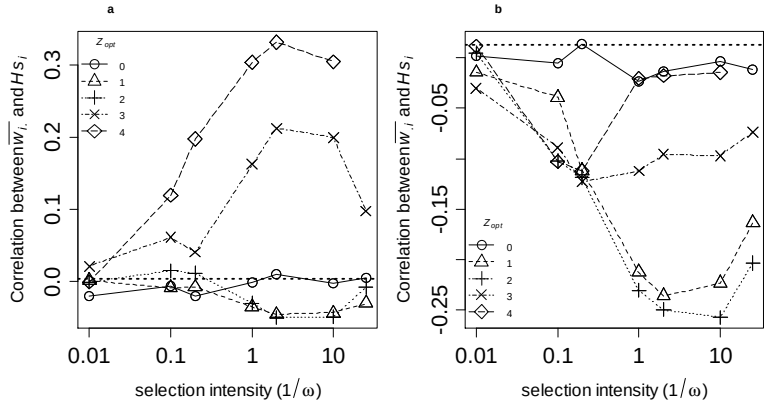


Figure II.B-S7: Results for a connectivity (*c*) of 0.8. **a.** Mean correlation over simulations between the level of within-population diversity of a given gene i (H_{Si}) and its level of regulation by the other genes network ($\overline{w}_{i.}$). **b.** Mean correlation over simulations between the level of within-population diversity of a given gene *i* (H_{Si}) and its level of regulation on the other genes network ($\overline{w}_{.i}$). In both cases, the values are the final values after 10,000 generations of evolution. They are given as a function of selection intensity (1/ω), for several values of optimal phenotypic distance (Z_{opt}). The values corresponding to $Z_{opt} = 4$ and $\omega^2 = 25$ are not shown, as simulations did not converge in that case. The dotted line corresponds to the neutral case.

Conclusion Générale

Les processus qui sont à l'origine de l'évolution du vivant sont complexes puisque celle-ci relève de processus génétiques, démographiques, stochastiques, culturels (pour les espèces sociales) et environnementaux. Autant la transmission du succès reproducteur que la sélection sur des caractères contrôlés par de nombreux gènes relèvent de phénomènes complexes qui vont modifier la diversité génétique, comme l'ont montré nos modèles individu-centrés, mais leurs effets dépendent fortement des interactions avec d'autres phénomènes environnementaux ou culturels.

Par exemple, nous montrons sur les caractères à déterminisme complexe que l'impact sur le réseau de régulation va dépendre de l'environnement, qui exerce la sélection naturelle sur les caractères. Nous avons montré dans ce cas que la réponse des gènes à cette pression de l'environnement dépendait du phénotype optimal dans cet environnement : (1) Si ce phénotype optimal correspond à l'état de stabilité naturelle du système, tel que prédit par les régulations entre gènes, la diversité génétique est faiblement réduite alors que la diversité phénotypique est très réduite. (2) Si le phénotype optimal dévie de façon intermédiaire par rapport à cet état de stabilité, les gènes les plus régulateurs et les plus exprimés seront sous plus forte sélection que les autres. Ils ont en effet un rôle primordial pour l'obtention du phénotype optimal et toute modification de leur expression va fortement déstabiliser le système, expliquant pourquoi ces gènes sont très contraints (3) Si le phénotype optimal dévie fortement de l'état de stabilité du système, tous les gènes doivent être fortement exprimés pour que le phénotype atteigne l'optimum. Dans ce cas les gènes qui sont les plus régulés sont moins sous la contrainte de la sélection car leur expression dépend plus des autres locus et les moins régulés vont être plus sous sélection pour atteindre l'expression du phénotype optimal.

Ces conclusions ont été obtenues à partir d'un modèle qui dérive de ceux de Wagner (1996) et Siegal & Bergman (2002) ; comme eux nous avons supposé que les gènes se régulaient les uns les autres, mais nous avons fait l'hypothèse que le phénotype d'un individu était la somme des produits des protéines.

Nous avons aussi comparé les résultats de notre modèle avec un modèle où la sélection doit amener le niveau d'expression de chaque gène à une valeur précise, comme l'avaient modélisé originalement Wagner (1996) et Siegal et Bergman (2002). Nous avons clairement

observé comme précédemment que la contrainte sélective ne se répartit pas uniformément entre les locus dans une certaine gamme de sélection. Mais nous n'avons pas pu corréler cette différence de diversité entre les locus avec les différents niveaux d'expression ou de régulation. Ce résultat montre que selon le type de mécanisme sous-jacent qui contrôle le caractère, ce ne sont pas les mêmes locus qui répondent à la sélection. En ce sens, d'autres types de relations entre le phénotype et le génotype pourraient être testées. Nous pourrions supposer par exemple que la sélection agit sur le produit d'un seul gène et non sur l'ensemble des produits comme supposé dans notre modèle ou ceux de Wagner (1996) et Siegal et Bergman (2002).

Aussi, le modèle utilisé considère un individu haploïde qui transmet aléatoirement la moitié des locus à son descendant, l'autre moitié étant donnée par son autre parent. Il serait intéressant de modéliser un système diploïde pour vérifier si nous retrouvons les mêmes impacts. Ceci nous permettrait notamment de regarder si les interactions de dominance entre les différents allèles interagissent avec les niveaux de régulations.

Nos conclusions montrent qu'il n'est pas possible de trouver une règle générale à propos de l'effet de la sélection sur la diversité génétique mais nous pouvons affirmer qu'elle ne se répartit ni uniformément, ni aléatoirement entre les locus d'un caractère soumis aux pressions sélectives de l'environnement. L'importance de la régulation dans l'évolution avait déjà été étudiée expérimentalement (Pour une revue lire Carroll 2000), mais nous montrons aussi qu'au niveau d'une population certains gènes ont un rôle prédominant par rapport aux autres, qu'ils soient faiblement régulés ou fortement régulateurs dans le réseau. Pour valider expérimentalement nos résultats il serait nécessaire d'étudier des populations dont un phénotype est soumis à des pressions de sélection différentes, en comparant les différences de niveau de diversité et de régulation de ces gènes (nombre de transcrits, affinité avec l'ADN…) dans le réseau de régulation.

La transmission du succès reproducteur modifie aussi la diversité génétique puisqu'en utilisant une approche par modélisation, Austerlitz et Heyer (1998), Sibert et *al.*, (2002), Blum et *al.* (2006) et nous-mêmes avons montré plusieurs impacts du phénomène, au niveau démographique (corrélation entre les tailles de fratrie, augmentation des variances du nombre d'enfants dans les populations) et au niveau des arbres de coalescence (diminution du TMRCA donc de la diversité, branches externes plus longues, arbres plus déséquilibrés). L'ensemble de ces facteurs affecte donc le niveau de polymorphisme pour des marqueurs

neutres et la répartition de ce polymorphisme entre les individus. Une des originalités de notre approche a été de ne plus s'intéresser à des arbres de coalescence de gènes mais à des lignées d'individus extraites de généalogies, pour lesquelles nous avons aussi montré que la transmission du succès reproducteur entraînait un déséquilibre. Par ailleurs nous avons pu montrer comment la transmission interagissait avec d'autres phénomènes tels que l'hétérogénéité du succès reproducteur, qui amplifie l'effet de la transmission sur les paramètres démographiques et la forme des arbres. Enfin nous avons pu montrer l'intérêt de considérer différents types de marqueurs (X, Y, mitochondriaux et autosomaux) car ils répondent différemment selon que la transmission est uniparentale ou biparentale. Enfin nous avons mis en évidence qu'il fallait tenir compte des mécanismes de formation non-aléatoire des couples, en particulier l'homogamie pour la taille de fratrie.

Nos modèles théoriques ont donc montré l'impact clair de la transmission du succès reproducteur sur le déséquilibre des arbres de coalescence et des arbres généalogiques. Comme nous l'avons souligné, il s'agit d'un phénomène caractéristique de la transmission du succès reproducteur (Blum et al. 2006) alors que d'autres conséquences de cette transmission sont moins spécifiques et peuvent résulter d'autres phénomènes génétiques ou démographiques. Reconstruire par des méthodes phylogénétiques les lignées des gènes et reconstituer leur déséquilibre apparaît donc comme la voie privilégiée pour détecter le phénomène de transmission du succès reproducteur à partir des données génétiques. Pour autant nous avons montré, en utilisant nos simulations, qu'il était probable que la transmission créerait un déséquilibre suffisamment élevé pour être détecté seulement s'il y avait aussi de l'hétérogénéité du succès reproducteur dans la population.

Ceci ouvre donc la possibilité de détecter ce phénomène dans les populations humaines. Déjà Blum et *al.* (2006), en calculant le déséquilibre des arbres reconstruits par méthode phylogénétique sur des données de polymorphisme de l'ADN mitochondrial, ont montré la présence d'une transmission matrilinéaire (ou biparentale) dans les populations de chasseurs-cueilleurs. En étendant ce travail nous avons montré qu'il est possible dans un premier temps de retrouver une transmission patrilinéaire en reconstruisant des arbres sur des données du chromosome Y (SNPs et microsatellites) par des méthodes de phylogénie, en tenant compte des spécificités mutationnelles de ces marqueurs.

Dans un second temps nous avons montré qu'il est possible de détecter le phénomène sur les autosomes, à partir des grands jeux de données de polymorphisme génétique de type

HapMap, en délimitant des blocks non-recombinants et en reconstruisant les arbres généalogiques des gènes sur les blocks suffisamment grands.

Enfin nous avons montré la possibilité d'utiliser les généalogies ascendantes d'individus pour inférer la transmission du succès reproducteur. Ceci est d'ailleurs la façon dont cette transmission avait originellement été détectée (Pearson et al. 1899). Nos approches par simulation et empiriques sur trois populations du Cilento (Italie) nous a permis de caractériser les différentes mesures permettant de détecter le phénomène, notamment via le déséquilibre des lignées reconstruites à partir des arbres généalogiques mais aussi via la répartition de la contribution génétique entre individus.

Pour finir, nous pouvons nous demander dans quelle mesure les deux phénomènes que nous avons étudiés ont des impacts similaires ou non. Nous constatons que la transmission du succès reproducteur et la sélection sur des caractères à déterminisme complexe modifient la diversité génétique via des réseaux de régulation, mais les impacts de ces phénomènes ne vont pas se répartir de la même façon sur le génome. Dans le cas d'une transmission culturelle, comme pour tout phénomène démographique, l'ensemble du génome est affecté, alors qu'au contraire pour une sélection sur des caractères complexes, la sélection agit sur les gènes impliqués dans le caractère. De plus, nous avons vu que dans ce cas les pressions de sélection ne s'exercent pas uniformément sur l'ensemble des locus, certains de ces locus étant plus affectés que d'autres selon les modalités d'action de cette sélection.

La modélisation en Biologie et plus particulièrement dans les Sciences de l'Évolution est constamment à la recherche d'approches élégantes, synthétiques et rapides permettant à la fois la prédiction de l'état futur d'un système selon les paramètres définissant les conditions de son évolution dans le temps mais également, et surtout, l'inférence des valeurs de ces mêmes paramètres à partir de jeux de données observées. A ce titre, la théorie du coalescent de Kingman ou le modèle infinitésimal ont par exemple permis des avancées exceptionnelles en Génétique des Populations ou en Génétique Quantitative. Dans bien des situations cependant, cette élégance se fait au prix d'une très grande difficulté voire d'une impossibilité à intégrer des aspects pourtant essentiels de la réalité physiologique de la construction du phénotype ou du fonctionnement parfois complexe des populations. La manière la plus simple de contourner ces obstacles méthodologiques est de développer des modèles individu-centrés qui permettent de considérer absolument tous les niveaux de complexité et de raffinement

imaginables, du moins en théorie. C'est ce que nous avons fait au cours de cette thèse pour un modèle de transmission du succès reproducteur et un modèle de sélection sur des réseaux de gènes. Pourvu qu'on se restreigne à un espace de paramètres relativement raisonnable, nous pensons avoir démontré dans les deux cas que ce type de modèles garde un grand pouvoir prédictif.

Les modèles individus-centrés sont donc des outils nécessaires pour comprendre l'impact d'un phénomène sélectif complexe bien déterminé. Il n'est cependant pas encore possible d'estimer les paramètres de ces modèles à partir des données observées, comme cela se fait pour des modèles de processus démographiques (expansions, goulot d'étranglement, migrations...). Ces estimations reposent en effet sur des méthodes « ABC » (Beaumont et al. 2002) utilisant des simulations par coalescence. Elles nécessitent la réalisation de très nombreuses simulations (en général plusieurs millions). Le nombre de paramètres qu'il faut explorer et la lenteur des simulations individu-centrées restent des limites à l'utilisation des simulations individu-centrées en ABC. Mais au vu de l'explosion de la rapidité des outils de calcul, il se peut que ce type de modèle devienne accessible aux méthodes ABC. Sinon l'autre solution, serait de développer des approches probabilistiques comme celles de la théorie de la coalescence permettant de modéliser rapidement des populations soumises à des processus sélectifs, mais le défi théorique reste à être relever.

Bibliographie

Agapow, P.-M. & Purvis, A. 2002 Power of eight tree shape statistics to detect nonrandom diversification: a comparison by simulation of two models of cladogenesis. *Systematic Biology* **51**, 866-872.

Anisimova, M., Nielsen, R. & Yang, Z. 2003 Effect of Recombination on the Accuracy of the Likelihood Method for Detecting Positive Selection at Amino Acid Sites. *Genetics* **164**, 1229-1236.

Arenas, M. & Posada, D. 2010 The Effect of Recombination on the Reconstruction of Ancestral Sequences. *Genetics* **184**, 1133-U429.

Austerlitz, F. & Heyer, E. 1998 Social transmission of reproductive behavior increases frequency of inherited disorders in a young-expanding population. *Proceedings of the National Academy of Sciences of the USA* **95**, 15140-4.

Austerlitz, F. & Heyer, E. 2000 Allelic association is increased by correlation of effective family size. *European journal of human genetics* **8**, 980-5.

Azevedo, R. B., Lohaus, R., Srinivasan, S., Dang, K. K. & Burch, C. L. 2006 Sexual reproduction selects for robustness and negative epistasis in artificial gene networks. *Nature* **440**, 87-90.

Balloux, F., Handley, L.-J. L., Jombart, T., Liu, H. & Manica, A. 2009 Climate shaped the worldwide distribution of human mitochondrial DNA sequence variation. *Proceedings of the Royal Society B-Biological Sciences* **276**, 3447-3455.

Barton, N. H. 2000 The effect of hitch-hiking on neutral genealogies. *Genetics Research* **72**, 123.

Bazin, E., Glémin, S. & Galtier, N. 2006 Population size does not influence mitochondrial genetic diversity in animals. *Science* **312**, 570-572.

Beaumont, M. A. 2005 Adaptation and speciation: what can F(st) tell us? *Trends Ecol Evol* **20**, 435-40.

Beaumont, M. A. & Nichols, R. A. 1996 Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London Series B-Biological Sciences* **263**, 1619-1626.

Beaumont, M. A., Zhang, W. & Balding, D. J. 2002 Approximate Bayesian Computation in Population Genetics. *Genetics* **162**, 2025-2035.

Bergman, A. & Siegal, M. L. 2003 Evolutionary capacitance as a general feature of complex gene networks. *Nature* **424**, 549.

Blum, M. G. B., Heyer, E., François, O. & Austerlitz, F. 2006 Matrilineal fertility inheritance detected in hunter-gatherer populations using the imbalance of gene genealogies. *PLoS Genetics* **2**, e122.

Bocquet-Appel, J. P. & Jakobi, L. 1993 A test of a path model of biocultural transmission of fertility. *Annals of Human Biology* **20**, 335-347.

Bost, B., Dillmann, C. & de Vienne, D. 1999 Fluxes and metabolic pools as model traits for quantitative genetics. I. The L-shaped distribution of gene effects. *Genetics* **153**, 2001-12.

Campbell, D. & Bernatchez, L. 2004 Generic scan using AFLP markers as a means to assess the role of directional selection in the divergence of sympatric whitefish ecotypes. *Mol Biol Evol* **21**, 945-56.

Cannings, C. 1974 The Latent Roots of Certain Markov Chains Arising in Genetics: A New Approach, I. Haploid Models. *Advances in Applied Probability* **6**, 260.

Carroll, S. B. 2000 Endless forms: the evolution of gene regulation and morphological diversity. *Cell* **101**, 577-80.

Cavalli-Sforza, L. L. & Feldman, M. W. 1981 *Cultural transmission and evolution: a quantitative approach.* Princeton, N.J.: Princeton University Press.

Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. 1994 *The history and geography of human genes*: Princeton University Press.

Chaix, R., Quintana-Murci, L., Hegay, T., Hammer, M. F., Mobasher, Z., Austerlitz, F. & Heyer, E. 2007 From social to genetic structures in central Asia. *Curr Biol* **17**, 43-8.

Chouard, T. 2008 Beneath the surface. *Nature* **456**, 300-3.

Ciliberti, S., Martin, O. C. & Wagner, A. 2007a Innovation and robustness in complex regulatory gene networks. *Proc Natl Acad Sci U S A* **104**, 13591-6.

Ciliberti, S., Martin, O. C. & Wagner, A. 2007b Robustness can evolve gradually in complex regulatory gene networks with varying topology. *PLoS Comput Biol* **3**, e15.

Clark, A. G., Hubisz, M. J., Bustamante, C. D., Williamson, S. H. & Nielsen, R. 2005 Ascertainment bias in studies of human genome-wide polymorphism. *Genome Research* **15**, 1496-1502.

Cohen, J. E. 1975 Childhood mortality, family size and birth order in pre-industrial Europe. *Demography* **12**, 35-55.

Colonna, V., Nutile, T., Astore, M., Guardiola, O., Antoniol, G., Ciullo, M. & Persico, M. G. 2007 Campora: A young genetic isolate in South Italy. *Human Heredity* **64**, 123-135.

Colonna, V., Nutile, T., Ferrucci, R. R., Fardella, G., Aversano, M., Barbujani, G. & Ciullo, M. 2009 Comparing population structure as inferred from genealogical versus genetic information. *Eur J Hum Genet* **17**, 1635-41.

Cork, J. M. & Purugganan, M. D. 2004 The evolution of molecular genetic pathways and networks. *Bioessays* **26**, 479-84.

Dalziel, A. C., Rogers, S. M. & Schulte, P. M. 2009 Linking genotypes to phenotypes and fitness: how mechanistic biology can inform molecular ecology. *Molecular Ecology* **18**, 4997.

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

Derrida, B., Manrubia, S. C. & Zanette, D. H. 2000 On the Genealogy of a Population of Biparental Individuals. *Journal of Theoretical Biology* **203**, 303.

Dobzhansky, T. 1937 *Genetics and the origin of species*: Columbia University Press.

Dulik, M. C., Osipova, L. P. & Schurr, T. G. 2011 Y-Chromosome variation in Altaian Kazakhs reveals a common paternal gene pool for Kazakhs and the influence of Mongolian expansions. *PLoS One* **6**, e17548.

Durbin, M. L., Lundy, K. E., Morrell, P. L., Torres-Martinez, C. L. & Clegg, M. T. 2003 Genes that determine flower color: the role of regulatory changes in the evolution of phenotypic adaptations. *Mol Phylogenet Evol* **29**, 507-18.

Ehrenreich, I., Hanzawa, Y., Chou, L., Roe, J., Kover, P. & Purugganan, M. 2009 Candidate Gene Association Mapping of Arabidopsis Flowering Time. *Genetics*.

Elson, J. L., Turnbull, D. M. & Howell, N. 2004 Comparative genomics and the evolution of human mitochondrial DNA: Assessing the effects of selection. *American Journal of Human Genetics* **74**, 229-238.

Evangelisti, A. M. & Wagner, A. 2004 Molecular evolution in the yeast transcriptional regulation network. *J Exp Zool B Mol Dev Evol* **302**, 392-411.

Fay, J. C. & Wittkopp, P. J. 2008 Evaluating the role of natural selection in the evolution of gene regulation. *Heredity* **100**, 191-9.

Fisher, R. A. 1930 The Genetical Theory of Natural Selection. *Clarendon Press*.

Flowers, J. M., Hanzawa, Y., Hall, M. C., Moore, R. C. & Purugganan, M. D. 2009 Population genomics of the Arabidopsis thaliana flowering time gene network. *Mol Biol Evol* **26**, 2475-86.

Flowers, J. M., Sezgin, E., Kumagai, S., Duvernell, D. D., Matzkin, L. M., Schmidt, P. S. & Eanes, W. F. 2007 Adaptive evolution of metabolic pathways in Drosophila. *Mol Biol Evol* **24**, 1347-54.

Frère, C. H., Krützen, M., Mann, J., Connor, R. C., Bejder, L. & Sherwin, W. B. 2010 Social and genetic interactions drive fitness variation in a free-living dolphin population. *Proceedings of the National Academy of Sciences of the USA* **107**, 19949-19954.

Fu, Y. X. 1996 New statistical tests of neutrality for DNA samples from a population. *Genetics* **143**, 557-70.

Fu, Y. X. & Li, W. H. 1993 Statistical Tests of Neutrality of Mutations. *Genetics* **133**, 693-709.

Fujita, P. A., Rhead, B., Zweig, A. S., Hinrichs, A. S., Karolchik, D., Cline, M. S., Goldman, M., Barber, G. P., Clawson, H., Coelho, A., Diekhans, M., Dreszer, T. R., Giardine, B. M., Harte, R. A., Hillman-Jackson, J., Hsu, F., Kirkup, V., Kuhn, R. M., Learned, K., Li, C. H., Meyer, L. R., Pohl, A., Raney, B. J., Rosenbloom, K. R., Smith, K. E., Haussler, D. & Kent, W. J. 2011 The UCSC Genome Browser database: update 2011. *Nucleic Acids Research*.

Fusco, G. & Cronk, Q. 1995 A new method for evaluating the shape of large phylogenies. *Journal of Theoretical Biology* **175**, 235-243.

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S. N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E. S., Daly, M. J. & Altshuler, D. 2002 The Structure of Haplotype Blocks in the Human Genome. *Science* **296**, 2225-2229.

Gagnon, A. & Heyer, E. 2001 Intergenerational correlation of effective family size in early Quebec (Canada). *American Journal of Human Biology* **13**, 645-659.

Garfield, D. A. & Wray, G. A. 2010 The Evolution of Gene Regulatory Interactions. *Bioscience* **60**, 15-23.

Geary, D. C., Vigil, J. & Byrd-Craven, J. 2004 Evolution of human mate choice. *Journal of Sex Research* **41**, 15.

Goldringer, I. & Bataillon, T. 2004 On the distribution of temporal variations in allele frequency: consequences for the estimation of effective population size and the detection of loci undergoing selection. *Genetics* **168**, 563-8.

Goldringer, I., Enjalbert, J., Raquin, A. L. & Brabant, P. 2001 Strong selection in wheat populations during ten generations of dynamic management. *Genetics Selection Evolution* **33**, S441-S463.

Griffiths, R. C. & Marjoram, P. 1996 Ancestral inference from samples of DNA sequences with recombination. *J Comput Biol* **3**, 479-502.

Griffiths, R. C. & Tavare, S. 1994 Sampling theory for neutral alleles in a varying environment. *Philos Trans R Soc Lond B Biol Sci* **344**, 403-10.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W. & Gascuel, O. 2010 New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology* **59**, 307-321.

Hahn, M. W., Rockman, M. V., Soranzo, N., Goldstein, D. B. & Wray, G. A. 2004 Population genetic and phylogenetic evidence for positive selection on regulatory mutations at the factor VII locus in humans. *Genetics* **167**, 867-77.

Haldane, J. B. S. 1932 *The Causes of Evolution.* London: Longmans Green.

Hancock, A. M., Clark, V. J., Qian, Y. & Di Rienzo, A. 2011 Population genetic analysis of the uncoupling proteins supports a role for UCP3 in human cold resistance. *Mol Biol Evol* **28**, 601-14.

Hardy, G. H. 1908 Mendelian Proportions In A Mixed Population. *Science* **28**, 49-50.

Hasegawa, M., Kishino, H. & Yano, T. 1985 Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* **22**, 160-74.

Helgason, A., Hrafnkelsson, B., Gulcher, J. R., Ward, R. & Stefansson, K. 2003 A populationwide coalescent analysis of Icelandic matrilineal and patrilineal genealogies: evidence for a faster evolutionary rate of mtDNA lineages than Y chromosomes. *American Journal of Human Genetics* **72**, 1370-88.

Helgason, A., Pálsson, S., Guðbjartsson, D. F., Kristjánsson, ó. & Stefánsson, K. 2008a An Association Between the Kinship and Fertility of Human Couples. *Science* **319**, 813-816.

Helgason, A., Pálsson, S., Guðbjartsson, D. F., Kristjánsson, ó. & Stefánsson, K. 2008b Response to Comments on "An Association Between the Kinship and Fertility of Human Couples". *Science* **322**, 1634.

Heyer, E. 1993 Population-Structure And Immigration - A Study Of The Valserine Valley (French Jura) From The 17th-Century Until The Present. *Annals Of Human Biology* **20**, 565-573.

Heyer, E. 1995 Mitochondrial And Nuclear Genetic Contribution Of Female Founders To A Contemporary Population In Northeast Quebec. *American Journal Of Human Genetics* **56**, 1450-1455.

Heyer, E. 1999 Les "enfants utiles": Une mesure démographique pour la génétique des populations. *Population* **4/5**, 677-691.

Heyer, E. 2009 One Founder/One Gene Hypothesis in a New Expanding Population: Saguenay (Quebec, Canada). *Human Biology* **81**, 645.

Heyer, E., Balaresque, P., Jobling, M. A., Quintana-Murci, L., Chaix, R., Segurel, L., Aldashev, A. & Hegay, T. 2009 Genetic diversity and the emergence of ethnic groups in Central Asia. *BMC Genet* **10**, 49.

Heyer, E., Sibert, A. & Austerlitz, F. 2005 Cultural transmission of fitness: genes take the fast lane. *Trends in Genetics* **21**, 234-9.

Heyer, E. & Tremblay, M. 1995 Variability of the genetic contribution of Quebec population founders associated to some deleterious genes. *Am J Hum Genet* **56**, 970-8.

Hill, W. G. 1976 Linkage Disequilibrium among Neutral Mutant Alleles in Finite Population. *Advances in Applied Probability* **8**, 10.

Hoekstra, H. E. & Nachman, M. W. 2003 Different genes underlie adaptive melanism in different populations of rock pocket mice. *Mol Ecol* **12**, 1185-94.

Hudson, R. R. 1990 Gene genealogies and the coalescent process. *Oxford Survey in Evolutionary Biology* **7**, 1-42.

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

Hudson, R. R. & Kaplan, N. L. 1985 Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**, 147-164.

Huelsenbeck, J. P. & Kirkpatrick, M. 1996 Do Phylogenetic Methods Produce Trees with Biased Shapes? *Evolution* **50**, 1418.

Huerta-Sanchez, E. & Durrett, R. 2007 Wagner's canalization model. *Theor Popul Biol* **71**, 121-30.

Huxley, J. 1942 *Evolution: The modern Synthesis*: Harper & brothers.

Imaizumi, Y., Nei, M. & Furusho, T. 1970 Variability and heritability of human fertility. *Annals of Human Genetics* **33**, 251-9.

Jacquesson, S. 2002 Parcours ethnographiques dans l'histoire des deltas ("Karakalpaks et autres gens de l'Aral: entre rivages et déserts"). *Cahiers d'Asie centrale [En ligne]* **10**, 51-90.

Jovelin, R., Dunham, J. P., Sung, F. S. & Phillips, P. C. 2009 High nucleotide divergence in developmental regulatory genes contrasts with the structural elements of olfactory pathways in caenorhabditis. *Genetics* **181**, 1387-97.

Jukes, T. H., Cantor, C. R. & Munro, H. N. 1969 *Evolution of Protein Molecules*. Evolution of Protein Molecules: Academy Press.

Kacser, H. & Burns, J. A. 1981 The molecular basis of dominance. *Genetics* **97**, 639-66.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T. & Yamanishi, Y. 2008 KEGG for linking genomes to life and the environment. *Nucleic Acids Research* **36**, D480-D484.

Kelly, M. J. 2001 Lineage loss in Serengeti cheetahs: Consequences of high reproductive variance and heritability of fitness on effective population size. *Conservation Biology* **15**, 137-147.

Keurentjes, J. J. B., Fu, J., Terpstra, I. R., Garcia, J. M., van den Ackerveken, G., Snoek, L. B., Peeters, A. J. M., Vreugdenhil, D., Koornneef, M. & Jansen, R. C. 2007 Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci. *Proceedings of the National Academy of Sciences of the USA* **104**, 1708-1713.

Kimura, M. 1968 Evolutionary Rate at the Molecular Level. *Nature* **217**, 624.

Kimura, M. 1983 *The neutral theory of molecular evolution.* Cambridge: Cambridge University Press.

Kingman, J. F. C. 1982a The coalescent. *Stochastic Processes and their Applications* **13**, 235-248.

Kingman, J. F. C. 1982b On the Genealogy of Large Populations. *Journal of Applied Probability* **19**, 27-43.

Kingsolver, J. G. & Pfennig, D. W. 2007 Patterns and Power of Phenotypic Selection in Nature. *BioScience* **57**, 561-572.

Kivisild, T., Shen, P. D., Wall, D. P., Do, B., Sung, R., Davis, K., Passarino, G., Underhill, P. A., Scharfe, C., Torroni, A., Scozzari, R., Modiano, D., Coppa, A., de Knijff, P., Feldman, M., Cavalli-Sforza, L. L. & Oefner, P. J. 2006 The role of selection in the evolution of human mitochondrial genomes. *Genetics* **172**, 373-387.

Kohler, H.-P., Rodgers, J. L. & Christensen, K. 1999 Is fertility behavior in our genes? Findings from a Danish twin study. *Population and Development Review* **25**, 253-288.

Kohler, H.-P., Rodgers, J. L., Miller, W. B., Skytthe, A. & Christensen, K. 2006 Bio-social determinants of fertility. *International Journal of Andrology* **29**, 46-53.

Kosova, G., Abney, M. & Ober, C. 2010 Heritability of reproductive fitness traits in a human population. *Proceedings of the National Academy of Sciences of the USA* **107**, 1772-1778.

Kumar, V., Langstieh, B. T., Madhavi, K. V., Naidu, V. M., Singh, H. P., Biswas, S., Thangaraj, K., Singh, L. & Reddy, B. M. 2006 Global patterns in human mitochondrial DNA and Y-chromosome variation caused by spatial instability of the local cultural processes. *PLoS Genet* **2**, e53.

Labouriau, R. & Amorim, A. n. 2008 Comment on "An Association Between the Kinship and Fertility of Human Couples". *Science* **322**, 1634.

Lagos-Quintana, M., Rauhut, R., Lendeckel, W. & Tuschl, T. 2001 Identification of novel genes coding for small expressed RNAs. *Science* **294**, 853-8.

Lande, R. 1976 Natural Selection and Random Genetic Drift in Phenotypic Evolution. *Evolution* **30**, 314.

Lansing, J. S., Watkins, J. C., Hallmark, B., Cox, M. P., Karafet, T. M., Sudoyo, H. & Hammer, M. F. 2008 Male dominance rarely skews the frequency distribution of Y chromosome haplotypes in human populations. *Proceedings of the National Academy of Sciences of the USA* **105**, 11645-11650.

Latta, R. G. 1998 Differentiation of allelic frequencies at quantitative trait loci affecting locally adaptive traits. *Am Nat* **151**, 283-92.

Le Corre, V. 2005 Variation at two flowering time genes within and among populations of Arabidopsis thaliana: comparison with markers and traits. *Molecular Ecology* **14**, 4181.

Le Corre, V. & Kremer, A. 2003 Genetic variability at neutral markers, quantitative trait loci and trait in a subdivided population under selection. *Genetics* **164**, 1205-1219.

Leclerc, R. D. 2008 Survival of the sparsest: robust gene networks are parsimonious. *Mol Syst Biol* **4**, 213.

Lewontin, R. C. 1964 The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. *Genetics* **49**, 49-67.

Livingstone, K. & Anderson, S. 2009 Patterns of Variation in the Evolution of Carotenoid Biosynthetic Pathway Enzymes of Higher Plants. *Journal of Heredity* **100**, 754-761.

Lonsdorf, E. V., Eberly, L. E. & Pusey, A. E. 2004 Sex differences in learning in chimpanzees. *Nature* **428**, 715.

Luikart, G., England, P. R., Tallmon, D., Jordan, S. & Taberlet, P. 2003 The power and promise of population genomics: from genotyping to genome typing. *Nat Rev Genet* **4**, 981-94.

MacCarthy, T. & Bergman, A. 2007 Coevolution of robustness, epistasis, and recombination favors asexual reproduction. *Proc Natl Acad Sci U S A* **104**, 12801-6.

MacCluer, J. W., VandeBerg, J. L., Read, B. & Ryder, O. A. 1986 Pedigree analysis by computer simulation. *Zoo Biology* **5**, 147.

Madrigal, L., Relethford, J. H. & Crawford, M. H. 2003 Heritability and anthropometric influences on human fertility. *Am J Hum Biol* **15**, 16-22.

Maia, L. P., Colato, A. & Fontanari, J. F. 2004 Effect of selection on the topology of genealogical trees. *Journal of Theoretical Biology* **226**, 315-20.

Mayr, E. 1942 *systematics and the origin of species*. New York: Colombia University Press.

McKusick, K. B., Schach, S. R. & Koeslag, J. H. 1990 Social mechanisms in the population genetics of Tay-Sachs and other lethal autosomal recessive diseases: a computer simulation model. *American journal of medical genetics* **36**, 178-82.

Moran, P. A. P. 1958 Random processes in genetics. *Mathematical Proceedings of the Cambridge Philosophical Society* **54**, 60-71.

Mullen, L. M. & Hoekstra, H. E. 2008 Natural selection along an environmental gradient: a classic cline in mouse pigmentation. *Evolution* **62**, 1555-70.

Murphy, M. 1999 Is the relationship between fertility of parents and children really weak? *Social Biology* **46**, 122-145.

Murphy, M. 2006 The role of assortative mating on population growth in contemporary developed societies. In *Agent-based computational modelling: applications in demography, social, economic and environmental sciences.* (ed. F. C. Billari, T. Fent, A. Prskawetz & J. Scheffran), pp. 61-84. Heidelberg: Physica-Verlag.

Murphy, M. & Knudsen, L. B. 2002 The intergenerational transmission of fertility in contemporary Denmark: the effects of number of siblings (full and half), birth order, and whether male or female. *Population Studies* **56**, 235-248.

Murray-McIntosh, R. P., Scrimshaw, B. J., Hatfield, P. J. & Penny, D. 1998 Testing migration patterns and estimating founding population size in Polynesia by using human mtDNA sequences. *Proceedings of the National Academy of Sciences of the USA* **95**, 9047-9052.

Nachman, M. W., Hoekstra, H. E. & D'Agostino, S. L. 2003 The genetic basis of adaptive melanism in pocket mice. *Proc Natl Acad Sci U S A* **100**, 5268-73.

Namroud, M. C., Guillet-Claude, C., Mackay, J., Isabel, N. & Bousquet, J. 2010 Molecular evolution of regulatory genes in spruces from different species and continents: heterogeneous patterns of linkage disequilibrium and selection but correlated recent demographic changes. *J Mol Evol* **70**, 371-86.

Nault, F., Desjardins, B. & Legare, J. 1990 Effects of Reproductive Behaviour on Infant Mortality of French-Canadians during the Seventeenth and Eighteenth Centuries. *Population Studies* **44**, 273-285.

Neel, J. V. 1970 Lessons from a "Primitive" People. *Science* **170**, 815.

Nei, M. & Murata, M. 1966 Effective population size when fertility is inherited. *Genetical research* **8**, 257-60.

Neuhauser, C. & Krone, S. M. 1997 The genealogy of samples in models with selection. *Genetics* **145**, 519-34.

Nielsen, R. 2001 Statistical tests of selective neutrality in the age of genomics. *Heredity* **86**, 641-647.

O'Brien, E., Jorde, L. B., Ronnlof, B., Fellman, J. O. & Eriksson, A. W. 1988 Founder effect and genetic disease in Sottunga, Finland. *Am J Phys Anthropol* **77**, 335-46.

Olsen, K. M., Womack, A., Garrett, A. R., Suddith, J. I. & Purugganan, M. D. 2002 Contrasting evolutionary forces in the Arabidopsis thaliana floral developmental pathway. *Genetics* **160**, 1641-50.

Orr, H. A. 2000 Adaptation and the cost of complexity. *Evolution* **54**, 13-20.

Paradis, E., Claude, J. & Strimmer, K. 2004 APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289-290.

Pavard, Samuel, Gagnon, Alain, Desjardins, Bertrand, Heyer & Evelyne. 2005 *Mother's death and child survival: The case of early Quebec*. Cambridge, ROYAUME-UNI: Cambridge University Press.

Pearson, K., Lee, A. & Bramley-Moore, L. 1899 Mathematical contributions to the Theory of Evolution. VI. Genetic (reproductive) selection: Inheritance of fertility in Man, and of fecundity in thoroughbred racehorses. *Philosophical Transactions of the Royal Society of London. Series A* **192**, 257-330.

Pedersen, J. O. N. 2000 Determinants of infant and child mortality in the west bank and Gaza strip. *Journal of Biosocial Science* **32**, 527-546.

Pluzhnikov, A., Nolan, D. K., Tan, Z., McPeek, M. S. & Ober, C. 2007 Correlation of intergenerational family sizes suggests a genetic component of reproductive fitness. *American Journal of Human Genetics* **81**, 165-9.

Posada, D. & Crandall, K. A. 2002 The effect of recombination on the accuracy of phylogeny estimation. *Journal Of Molecular Evolution* **54**, 396-402.

Prunier, J., Laroche, J., Beaulieu, J. & Bousquet, J. 2011 Scanning the genome for gene SNPs related to climate adaptation and estimating selection at the molecular level in boreal black spruce. *Mol Ecol* **20**, 1702-16.

Purugganan, M. D. 2000 The molecular population genetics of regulatory genes. *Mol Ecol* **9**, 1451-61.

Purvis, A., Katzourakis, A. & Agapow, P. M. 2002 Evaluating phylogenetic tree shape: two modifications to Fusco & Cronk's method. *Journal of Theoretical Biology* **214**, 99-103.

Ramírez-Soriano, A., Ramos-Onsins, S. E., Rozas, J., Calafell, F. & Navarro, A. 2008 Statistical Power Analysis of Neutrality Tests Under Demographic Expansions, Contractions and Bottlenecks With Recombination. *Genetics* **179**, 555-567.

Ramsay, H., Rieseberg, L. H. & Ritland, K. 2009 The correlation of evolutionary rate with pathway position in plant terpenoid biosynthesis. *Mol Biol Evol* **26**, 1045-53.

Rausher, M. D., Miller, R. E. & Tiffin, P. 1999 Patterns of evolutionary rate variation among genes of the anthocyanin biosynthetic pathway. *Mol Biol Evol* **16**, 266-74.

Rensch, B. 1947 *Neuere Probleme der Abstammungslehre. Die transspezifische Evolution*: Stuttgart.

Rhone, B., Brandenburg, J. T. & Austerlitz, F. 2011 Impact of selection on genes involved in regulatory network: a modelling study. *J Evol Biol* **24**, 2087-98.

Rhone, B., Vitalis, R., Goldringer, I. & Bonnin, I. 2010 Evolution of flowering time in experimental wheat populations: a comprehensive approach to detect genetic signatures of natural selection. *Evolution* **64**, 2110-25.

Richard, D. & Olivier, G. 2002 Fast and Accurate Phylogeny Reconstruction Algorithms Based on the Minimum-Evolution Principle. In *Proceedings of the Second International Workshop on Algorithms in Bioinformatics*: Springer-Verlag.

Rodgers, J. L., Kohler, H.-P., Kyvik, K. O. & Christensen, K. 2001 Behavior genetic modeling of human fertility: Findings from a contemporary Danish Twin study. *Demography* **38**, 29.

Roff, D. A. 2007 A centennial celebration for quantitative genetics. *Evolution* **61**, 1017-32.

Rogers, S. M. & Bernatchez, L. 2005 Integrating QTL mapping and genome scans towards the characterization of candidate loci under parallel selection in the lake whitefish (Coregonus clupeaformis). *Mol Ecol* **14**, 351-61.

Ronsmans, C. 1995 Patterns of clustering of child mortality in a rural area of Senegal. *Population Studies* **49**, 443-461.

Rousset, F. 2002 Inbreeding and relatedness coefficients: what do they measure? *Heredity* **88**, 371.

Ruggiero, D., Dalmasso, C., Nutile, T., Sorice, R., Dionisi, L., Aversano, M., Bröet, P., Leutenegger, A.-L., Bourgain, C. & Ciullo, M. 2011 Genetics of VEGF Serum Variation in Human Isolated Populations of Cilento: Importance of VEGF Polymorphisms. *PLoS One* **6**, e16982.

Ruiz-Pesini, E., Mishmar, D., Brandon, M., Procaccio, V. & Wallace, D. C. 2004 Effects of purifying and adaptive selection on regional variation in human mtDNA. *Science* **303**, 223-226.

Saitou, N. & Nei, M. 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**, 406-25.

Sastry, N. 1997 Family-level clustering of childhood mortality risk in northeast Brazil. *Population Studies* **51**, 245-261.

Schierup, M. H. & Hein, J. 2000 Consequences of recombination on traditional phylogenetic analysis. *Genetics* **156**, 879-891.

Schliep, K. P. 2010 phangorn: Phylogenetic analysis in R. *Bioinformatics*.

Schlitt, T. & Brazma, A. 2007 Current approaches to gene regulatory network modelling. *BMC Bioinformatics* **8 Suppl 6**, S9.

Segurel, L., Martinez-Cruz, B., Quintana-Murci, L., Balaresque, P., Georges, M., Hegay, T., Aldashev, A., Nasyrova, F., Jobling, M. A., Heyer, E. & Vitalis, R. 2008 Sex-specific genetic structure and social organization in Central Asia: insights from a multi-locus study. *PLoS Genetics* **4**, e1000200.

Shapiro, M. D., Marks, M. E., Peichel, C. L., Blackman, B. K., Nereng, K. S., Jonsson, B., Schluter, D. & Kingsley, D. M. 2004 Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature* **428**, 717-23.

Sibert, A. 2002 Héritabilité Non-Génétique de la fécondité: effet sur le polymorphisme, pp. 146. Paris: Muséum d'Histoire Naturelle.

Sibert, A., Austerlitz, F. & Heyer, E. 2002 Wright-Fisher revisited: the case of fertility correlation. *Theoretical Population Biology* **62**, 181-97.

Siegal, M. L. & Bergman, A. 2002 Waddington's canalization revisited: Developmental stability and evolution. *Proceedings Of The National Academy Of Sciences Of The United States Of America* **99**, 10528-10532.

Siegal, M. L., Promislow, D. E. & Bergman, A. 2007 Functional and evolutionary inference in gene networks: does topology matter? *Genetica* **129**, 83-103.

Simpson's, G. G. 1944 *Tempo and Mode in Evolution*. New York: Columbia University Press.

Slatkin, M. & Hudson, R. R. 1991 Pairwise comparisons of mitochondrial-DNA sequences in stable and exponentially growing populations. *Genetics* **129**, 555-562.

Sokal, R. R. & Michener, C. D. 1958 A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin* **28**, 1409.

Sorice, R., Bione, S., Sansanelli, S., Ulivi, S., Athanasakis, E., Lanzara, C., Nutile, T., Sala, C., Camaschella, C., D'Adamo, P., Gasparini, P., Ciullo, M. & Toniolo, D. 2011 Association of a variant in the CHRNA5-A3-B4 gene cluster region to heavy smoking in the Italian population. *Eur J Hum Genet* **19**, 593.

Stebbins, G. L. 1950 *Variation and Evolution in Plants*. New York: Columbia University Press.

Stern, D. L. 1998 A role of Ultrabithorax in morphological differences between Drosophila species. *Nature* **396**, 463-6.

Stern, D. L. & Orgogozo, V. 2008 The loci of evolution: how predictable is genetic evolution? *Evolution* **62**, 2155-77.

Stern, D. L. & Orgogozo, V. 2009 Is genetic evolution predictable? *Science* **323**, 746-51.

Stewart, J. B., Freyer, C., Elson, J. L. & Larsson, N. G. 2008 Purifying selection of mtDNA and its implications for understanding evolution and mitochondrial disease. *Nature Reviews Genetics* **9**, 657-662.

Sun, C., Southard, C., Witonsky, D. B., Kittler, R. & Di Rienzo, A. 2010 Allele-specific down-regulation of RPTOR expression induced by retinoids contributes to climate adaptations. *PLoS Genet* **6**, e1001178.

Tajima, F. 1983 Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437-60.

Tajima, F. 1989 Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics* **123**, 585-595.

The International HapMap Consortium. 2003 The International HapMap Project. *Nature* **426**, 789-96.

The International HapMap Consortium. 2005 A haplotype map of the human genome. *Nature* **437**, 1299.

The International HapMap Consortium. 2007 A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851.

Tishkoff, S. A., Reed, F. A., Ranciaro, A., Voight, B. F., Babbitt, C. C., Silverman, J. S., Powell, K., Mortensen, H. M., Hirbo, J. B., Osman, M., Ibrahim, M., Omar, S. A., Lema, G., Nyambo, T. B., Ghori, J., Bumpstead, S., Pritchard, J. K., Wray, G. A. & Deloukas, P. 2007 Convergent adaptation of human lactase persistence in Africa and Europe. *Nat Genet* **39**, 31-40.

Travers, M. E. & McCarthy, M. I. 2011 Type 2 diabetes and obesity: genomics and the clinic. *Hum Genet* **130**, 41-58.

Tremblay, M. & Vézina, H. 2010 Genealogical Analysis of Maternal and Paternal Lineages in the Quebec Population. *Human Biology* **82**, 179.

Vaupel, J. W., Manton, K. G. & Stallard, E. 1979 The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* **16**, 439-54.

Vézina, H., Tremblay, M., Desjardins, B. & Houde, L. 2005 Origines et contributions génétiques des fondatrices et des fondateurs de la population québécoise. *Cahiers québécois de démographie* **34**, 235-258.

Via, M., Gignoux, C. & Burchard, E. G. l. 2010 The 1000 Genomes Project: new opportunities for research and social challenges. *Genome medicine* **2**, 3.

Vitalis, R., Dawson, K. & Boursot, P. 2001 Interpretation of variation across marker loci as evidence of selection. *Genetics* **158**, 1811-23.

Vitkup, D., Kharchenko, P. & Wagner, A. 2006 Influence of metabolic network structure and function on enzyme evolution. *Genome Biol* **7**, R39.

Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. 2006 A Map of Recent Positive Selection in the Human Genome. *PLoS Biol* **4**, e72.

Waddington, C. H. 1942 Canalization of Development and the Inheritance of Acquired Characters. *Nature* **150**, 563.

Wagner, A. 1996 Does evolutionary plasticity evolve? *Evolution* **50**, 1008-1023.

Wang, R. L., Stec, A., Hey, J., Lukens, L. & Doebley, J. 1999 The limits of selection during maize domestication. *Nature* **398**, 236-9.

Weinberg, W. 1908 Über den Nachweis der Vererbung beim Menschen. *Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg*.

Whitehead, H. 1998 Cultural selection and genetic diversity in matrilineal whales. *Science* **282**, 1708-11.

Wittkopp, P. J., Williams, B. L., Selegue, J. E. & Carroll, S. B. 2003 Drosophila pigmentation evolution: divergent genotypes underlying convergent phenotypes. *Proc Natl Acad Sci U S A* **100**, 1808-13.

Wray, G. A. 2007 The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet* **8**, 206-16.

Wright, S. 1931 Evolution In Mendelian Populations. *Genetics* **16**, 97-159.

Yang, Y. H., Zhang, F. M. & Ge, S. 2009 Evolutionary rate patterns of the Gibberellin pathway genes. *BMC Evol Biol* **9**, 206.

Zerjal, T., Xue, Y., Bertorelle, G., Wells, R. S., Bao, W., Zhu, S., Qamar, R., Ayub, Q., Mohyuddin, A., Fu, S., Li, P., Yuldasheva, N., Ruzibakiev, R., Xu, J., Shu, Q., Du, R., Yang, H., Hurles, M. E., Robinson, E., Gerelsaikhan, T., Dashnyam, B., Mehdi, S. Q. & Tyler-Smith, C. 2003 The Genetic legacy of the Mongols. *American Journal of Human Genetics* **72**, 717.

Zhao, Q., Thuillet, A. C., Uhlmann, N. K., Weber, A., Rafalski, J. A., Allen, S. M., Tingey, S. & Doebley, J. 2008 The role of regulatory genes during maize domestication: evidence from nucleotide polymorphism and gene expression. *Genetics* **178**, 2133-43.

Abstract

Selective forces are one of the major determinants of the evolution of phenotypic diversity and genetic diversity, in neutral and coding zones of the genome. Selection can occur on genetically - or culturally - transmitted traits. This thesis considers these two selective processes. First, we studied the effects of intergenerational fertility transmission on neutral genetic diversity. Second, we considered the impact of selection on phenotypes coded by a gene network and on the polymorphism of genes within the network.

Fertility transmission is a cultural or genetic phenomenon, which is characterised by a positive correlation between the sibship size of an individual and that of its children. It was observed both in human and animal populations. Using a modelling approach, we show that its effects and the possibility to detect it depend both on the kind of studied data (genetic or genealogical data) and on the different kind of transmission (uniparental, biparental). We show that other phenomena, such as the heterogeneity of reproductive success between individuals, can affect its effects. We develop several tools allowing to infer this phenomenon of fertility transmission on genealogical data, as well as on genetic polymorphism data that follows different mutational models (microsatellites, sequences, SNPs) and different transmission modes (haploid or diploid, sex-linked or not). We applied in particular these tools to three human populations of the Cilento area in Italy (genealogical and mitochondrial DNA data), to Central Asian data (Y chromosome) and to HapMap data (autosomes).

The second part of this thesis deals with the modelling of the action of natural selection on traits coded by regulation networks and describes the impact of such selection on the evolution of the phenotype and of the underlying genes. A given phenotype is the result of the interaction between different genes and their products. We show that phenotypic selection will modify the gene network organisation, as well as the level of polymorphism of the genes involved in the network. For example, when the optimal phenotype corresponds to an intermediate level of gene expression, the most regulatory genes will lose much of their diversity. Conversely, if the optimal phenotype corresponds to a very strong expression of the genes, it will be the most regulated genes that will be the most constrained. This analysis allowed us to show the complexity of the relations between selection, regulation networks, phenotypes and the environment.

Résumé

Les forces de sélection sont un des moteurs de l'évolution de la diversité phénotypique et de la diversité génétique neutre et des zones codantes du génome. Cette sélection peut s'appliquer sur des caractères transmis génétiquement ou culturellement. Le travail effectué s'intéresse à ces deux processus de sélection. Nous avons étudié dans un premier temps les effets de la transmission intergénérationnelle de la fécondité sur la diversité génétique neutre puis dans un deuxième temps l'impact de la sélection sur des phénotypes codés par des réseaux de gènes sur le polymorphisme de ces gènes.

La transmission de la fécondité est un phénomène culturel ou génétique qui se caractérise par une corrélation positive entre la taille de fratrie d'un individu et la taille de fratrie de ses enfants. Il a été observé tant dans des populations humaines qu'animales. Nous montrons, par l'outil de la modélisation, que ses effets et la possibilité de le détecter dépendent autant du type de données étudiées (génétiques ou généalogiques), que des différents types de transmission (uniparentale, biparentale). Nous montrons que d'autres phénomènes, tels que l'hétérogénéité du succès reproducteur des individus, peuvent fortement moduler son impact. Nous développons un certain nombre d'outils permettant de détecter ce phénomène de transmission de la fécondité tant sur des données généalogiques que sur des données génétiques relevant de différents modèles mutationnels (microsatellite, séquences, SNPs) et de différents types de transmission (haploïde ou diploïde, lié au sexe ou non). Nous avons appliqué ces outils notamment à trois populations humaines du Cilento en Italie (généalogies et ADN mitochondrial), des données d'Asie Centrale (chromosome Y) et des données HapMap (autosomes).

La seconde partie de la thèse porte sur la modélisation de l'action de la sélection naturelle sur des caractères codés par des réseaux de régulation et décrit l'impact de ce type de sélection sur l'évolution du phénotype et sur la diversité des gènes sous-jacents. Un phénotype est le résultat des interactions entre différents gènes et leurs produits. Nous montrons que la sélection sur ce phénotype va modifier l'organisation du réseau de gènes ainsi que le niveau de polymorphisme des gènes du réseau. Par exemple, lorsque le phénotype optimal correspond à une expression médiane des gènes, les gènes les plus régulateurs vont être soumis à une plus forte perte de diversité. En revanche, si le phénotype optimal correspond à une expression très forte, ce sont les gènes les plus régulés qui vont être les plus contraints. Cette analyse a permis de montrer la complexité des relations entre sélection, réseaux de régulation, phénotypes et environnement.