Université de Montréal

Ramifications génétiques et démographiques de l'effet fondateur québécois

par

Claude Bhérer

Programmes de biologie moléculaire

Faculté de médecine

Thèse présentée à la Faculté de Médecine

en vue de l'obtention du grade de Docteur

en biologie moléculaire

Avril 2014

Résumé

Les événements fondateurs et les expansions territoriales peuvent promouvoir une cascade de changements génétiques et ont ainsi pu jouer un rôle important au cours de l'histoire évolutive de l'Homme moderne. Or, chez les populations humaines, les conséquences évolutives et la dynamique démographique des processus de colonisation demeurent largement méconnues et difficiles à étudier. Dans cette thèse, nous avons utilisé les généalogies de la population fondatrice canadienne-française ainsi que des données génomiques pour étudier ces questions. Les analyses génomiques et généalogiques, remarquablement concordantes, ont dévoilé un nouveau portrait détaillé de la structure de la population du Québec, incluant un continuum de diversité génétique dans l'axe ouest/est et des sous-populations significativement différenciées. L'analyse de l'immigration fondatrice a montré que virtuellement tous les Canadiens français sont métissés. Allant à l'encontre d'une prétendue homogénéité génétique de la population, nos résultats démontrent que le peuplement des régions a engendré une rapide différentiation génétique et expliquent certaines signatures régionales de l'effet fondateur. De plus, en suivant les changements évolutifs dans les généalogies, nous avons montré que les caractéristiques des peuplements fondateurs peuvent affecter les traits liés à la fécondité et au succès reproducteur. Cette thèse offre une meilleure compréhension du patrimoine génétique du Québec et apporte des éléments de réponse sur les conséquences évolutives des événements fondateurs.

Mots clés : Génétique des populations humaines, effet fondateur, Québec, Canadiens français, généalogies

Abstract

Founding events and range expansions can promote a cascade of genetic changes and may have played an important role in the evolutionary history of modern humans. Yet the evolutionary consequences and demographic dynamics of these colonization processes remain poorly documented and challenging to study in human populations. In this thesis, we used deep-rooted genealogies from the French Canadian founder population in addition to genomic data to address these questions. Genomic and genealogical analyses were remarkably concordant and revealed a new portrait of Quebec fine-scale population structure, including a continuum of genetic diversity in the west/east axis and sub-populations significantly differentiated. The analysis of the founding immigration showed that virtually all French Canadians are admixed. Contrary to the idea of homogeneity of the population, our results demonstrate that the regional settlement histories led to a rapid genetic differentiation and explain some regional signatures of the founder effect. By monitoring evolutionary changes in real genealogies, we show that founding events impact fertility traits and reproductive success. This thesis leads to a better understanding of the genetic heritage of Quebec and provides insights on how peopling of new territories shaped human evolution.

Key words: human population genetics, founder effect, Quebec, French Canadians, genealogy

Table des matières

Liste des figures

CHAPITRE I

CHAPITRE II

CHAPITRE III

CHAPITRE V

Liste des tableaux

CHAPITRE I

CHAPITRE II

CHAPITRE III

CHAPITRE IV

CHAPITRE V

Liste des abréviations

ACA : Acadians

ADN : Acide désoxyribonucléique

ANOVA : Analysis of variance

BRCA1 : Breast cancer 1 gene

BRCA2 : Breast cancer 2 gene

CMT4C : Charcot-Marie-Tooth Neuropathy Type 4C

CTR : Centre

CEU : Échantillon HapMap formé de résidents de l'Utah avec des origines du nord et de l'ouest de l'Europe

CEPH : Centre d'étude du polymorphisme humain

CI : Confidence intervals

ChSLSJ : Charlevoix Saguenay-Lac-Saint-Jean

DNA : Deoxyribonucleic acid

EFS : Effective family size

FC : French Canadian(s)

FS : Family size

FUN : Founder's uniform number

GC : Genetic contribution

GFC : Gaspesian French Canadians

HI : Homogeneity index

H : Hétérozygotie

HGDP : Human Genome Diversity Project

HLA : Human Leukocyte antigen

HWE : Hardy Weinberg equilibrium

LD : Linkage disequilibrium

LOY : Loyalists

MAF : Minor allele frequency

MDS : Multidimensional scaling

MON : Montreal

MTL : Montreal City area

NE : North-East

NMTL : North of Montreal

NW : North-West

N : Taille d'une population

N_e : Effectif efficace d'une population

NHSA2 : Névrite héréditaire et sensitive de type II

NS : North Shore

PCA : Principal component analysis

PC : Principal component

QUE : Quebec city area

QFP : Quebec founder population

RC : Range core

rGC : Relative genetic contribution

ROH : Runs of homozygosity

SAG : Saguenay

SLSJ : Saguenay-Lac-Saint-Jean

SMTL : South of Montreal

SNP : Single nucleotide polymorphism

WF : Wave front

WFI : Wave front index

WQC : Whole Quebec

Remerciements

Nombreux sont ceux qui m'ont accompagné dans ce périple scientifique sur le génome québécois. D'abord, je tiens à exprimer ma profonde gratitude à mes directeurs, Damian Labuda et Hélène Vézina. Merci de m'avoir accueillie au sein de vos groupes de recherche et associée à vos projets passionnants. Vos conseils et votre soutien autant d'un point de vue scientifique que non scientifique ont fait de moi une personne meilleure.

Merci à la grande famille du projet BALSAC. Avec cette thèse, j'espère avoir illustré comment le fichier de population est un outil de recherche privilégié en génétique des populations. Les travaux présentés ici ont été rendus possible grâce à votre travail monumental réalisé depuis plusieurs décennies. Un merci spécial à Michèle Jomphe et Ève-Marie Lavoie avec qui j'ai toujours beaucoup de plaisir à collaborer.

Merci à mes collègues passés et présents du labo Labuda. Je vous remercie pour votre présence, pour avoir su me conseiller et me partager votre science; vous avez fait de ces années une joyeuse aventure humaine. Un merci particulier à Elias Gbeha, Claudia Moreau et Vania Yotova. Pascale Gerbault, merci de m'inspirer avec ta science, merci pour ton aide, pour tout le plaisir que nous avons eu dans une année éclair et pour ton amitié trans-continentale. Merci à Véronique Ladret pour ton support et ton amitié.

Je n'ai pas suffisamment de mots pour remercier ma grande amie et collègue, Julie Hussin. Nos discussions infinies sur la science et la vie sont pour moi une source d'inspiration quotidienne et vitale.

Merci à Ted Bradley, Guillaume Lettre et Fanie Pelletier de prendre le temps d'examiner cette thèse. Merci aux membres de mon comité de thèse, en particulier Marie-Hélène Roy-Gagnon et Daniel Sinnett, qui m'ont soutenu à des moments cruciaux et qui m'ont guidée avec leurs conseils judicieux.

Merci à Laurent Excoffier, ton séjour à Montréal a été pour moi une période scientifique très stimulante. Je partage le crédit des travaux présentés ici avec tous mes co-auteurs et je remercie chaleureusement les participants aux projets. Je tiens aussi à remercier les assistantes du département de biologie moléculaire et du Centre de recherche du CHU Sainte-Justine. Merci à Vivianne Jodoin, Dominika Kozubska et Sandy Lalonde.

Cette thèse n'aurait pas été possible sans le soutien de l'amour de ma vie, Martin Dubreuil. C'est avec beaucoup d'émotion que je le remercie pour son aide, son soutien, son écoute et ses conseils. Merci de si bien t'occuper de Nicole et de moi. Merci de m'encourager à poursuivre ma carrière scientifique et merci de m'accompagner à l'étranger pour mon post-doctorat. Merci à ma fille Nicole Rose Dubreuil-Bhérer pour tous ces moments où tu as sagement laissé maman travailler.

Je tiens à remercier mes amis qui m'ont soutenue au cours de cette thèse. Un merci spécial à Karine Lacroix, Isabelle Gauvreau, Marie-Hélène Éthier, Annie Beaudoin, Gita Seaton, Pascal Lapointe, Isabelle Thiffault, Charles-Antoine Crête, Martine Zilversmit, Alex Laferrière et Marie-Ève Petit. Merci à ma grande amie Josée Bergeron, ton apport est immortalisé dans cette thèse. Merci à Louis Bhérer pour tes corrections.

La réussite de cette thèse tient aussi du soutien inconditionnel de ma famille et en particulier de mes parents, Marielle Ouimet et Dominique Bhérer. Merci maman pour ta force et tes conseils. Papa, je suis encore émue d'avoir pu vérifier tes hypothèses dans nos projets. Milles merci à Natacha, Gabrielle, Christian, André, Daphné et Gaïane. Merci à Odile Bhérer pour les photos de famille publiées dans Science.

Cette thèse a été réalisée avec le soutien financier des bourses d'excellence des Programmes de biologie moléculaire, de la Fondation de l'Hôpital Sainte-Justine et de la Fondation pour la recherche sur les maladies infantiles ainsi que du Fonds de recherche en santé du Québec.

à Nicole Rose Dubreuil-Bhérer

voici notre Québec, notre héritage

CHAPITRE I: Introduction

L’EFFET FONDATEUR CHEZ LES POPULATIONS HUMAINES: LE CAS DES CANADIENS FRANÇAIS DU QUÉBEC

INTRODUCTION

La fondation d'une nouvelle population est un phénomène démographique qui peut jouer un rôle déterminant dans le façonnement de son patrimoine génétique. Une population fondatrice est issue d'un événement migratoire, au cours duquel un groupe relativement limité de fondateurs migre, s'établit sur un territoire nouveau et donne ainsi naissance à une nouvelle population. Le bagage génétique unique apporté par chacun des fondateurs constitue le stock initial de diversité de la nouvelle population, qui sera ensuite remodelé, notamment par les comportements démographiques des générations suivantes. Des événements de fondation de diverses natures ont marqué l'évolution de l'Homme moderne. Depuis leur origine en Afrique, les humains ont progressivement peuplé toute la planète par une série de fondations de nouvelles populations et d'expansions territoriales. Au cours des cinq derniers siècles, les grands mouvements de colonisation européens sur les autres continents ont donné naissance à de nombreuses populations fondatrices. Comprendre les processus évolutifs en jeu lors de tels événements fondateurs et l'étendue de leurs conséquences est crucial pour comprendre comment l'Homme moderne a colonisé avec succès toute la planète et pour expliquer la variation génétique actuelle des populations humaines.

Chez l'humain, il existe une importante variation géographique d'une population à l'autre, à la fois dans les patrons de diversité génétique et dans les phénotypes. Notamment, on retrouve chez les populations fondatrices récentes une fréquence accrue de certaines maladies mendéliennes qui ne sont que rarement voire jamais observées chez d'autres populations, alors qu'à l'inverse, d'autres maladies y sont moins fréquentes ou absentes. Cette observation démontre que les différences inter-populationnelles dans les fréquences des variants génétiques, ou allèles, contribuent à la variation de l'incidence des maladies monogéniques d'une population à l'autre et suggère qu'elles pourraient aussi contribuer à la variation spatiale de l'incidence des

maladies plus communes. D'autre part, le bagage particulier de maladies mendéliennes observé chez les populations fondatrices récentes suggère que les événements fondateurs affectent la variation génétique et phénotypique, tel que postulé par le principe d'effet fondateur (Mayr 1942).

Les Canadiens français du Québec, au nord-est de l'Amérique du Nord forment un peuple récent dont les origines remontent au peuplement fondateur amorcé sous le Régime Français, il y a 400 ans. Ils sont les descendants de 8 500 fondateurs qui se sont établis en Nouvelle-France, entre 1608 et la Conquête britannique de 1760. Depuis quelques décennies, des études génétiques ont mis en lumière l'héritage génétique des Canadiens français, notamment son bagage particulier de maladies héréditaires mendéliennes. Dans cette thèse, j'utilise les généalogies des Canadiens français du Québec pour étudier de façon détaillée le peuplement fondateur du Québec et son impact sur le génome. La population canadienne-française du Québec représente une population modèle privilégiée pour étudier les effets fondateurs récents. Les résultats de ma thèse, en retour, ont le potentiel d'informer les politiques de santé publique en matière de génétique et de faciliter l'optimisation des études visant à découvrir les bases génétiques de susceptibilité aux maladies rares et communes.

Les évènements de fondation ont certainement façonné le génome humain, mais comment? Cette question, qui demeure ouverte, a des implications pour mieux comprendre l'évolution humaine et la santé des individus. Dans la première partie de ce chapitre, je propose une recension des écrits sur l'effet fondateur en lien avec l'évolution des populations humaines. Dans la seconde partie, je présente une synthèse des connaissances sur l'histoire génétique et démographique des Canadiens français du Québec. Enfin, dans la troisième partie, j'expose la problématique de ma thèse, ainsi que les objectifs, les hypothèses et les questions de recherche.

I. L'EFFET FONDATEUR ET L'ÉVOLUTION DES POPULATIONS HUMAINES

La fondation d'une population sur un territoire nouveau cause une rupture démographique entre la population source et la nouvelle. En conséquence, la nouvelle population peut se différencier, à la fois génétiquement et phénotypiquement, de la population source de laquelle elle dérive. En génétique des populations, ce phénomène est connu comme l'effet fondateur. De nombreuses études théoriques ont défini les conséquences génétiques de l'effet fondateur. Cependant, chez les populations humaines, la nature et l'ampleur des changements causés par l'effet fondateur restent à ce jour largement méconnus. Dans cette section, je définis l'effet fondateur et décris ses conséquences évolutives. Tout d'abord, je présente une recension des écrits portant sur les modèles théoriques de l'effet fondateur et leurs prédictions. Ensuite, je dresse un portrait des populations fondatrices humaines à travers le monde. Enfin, je propose un bilan des approches méthodologiques pour évaluer l'effet fondateur dans les populations naturelles.

1. L'effet fondateur et ses conséquences évolutives

En 1942, Ernst Mayr a postulé le principe de l'effet fondateur en ces termes :

« The reduced variability of small populations is not always due to accidental gene loss, but sometimes to the fact that the entire population was started by a single pair or by a single fertilized female. These "founders" of the population carried with them only a very small proportion of the variability of the parent population. This "founder" principle sometimes explains even the uniformity of rather large populations, particularly if they are well isolated and near the borders of the range of the species » (Mayr 1942, p. 237).

Mayr s'appuyait sur les travaux précurseurs de Wright démontrant que les populations de petites tailles sont sujettes à une plus forte dérive génétique[1], soit de plus grandes variations dans les fréquences alléliques d'une génération à l'autre, dues au hasard de la transmission des gamètes parentaux (Wright, 1931; 1938). Mayr soutenait que l'effet fondateur pouvait conduire à l'évolution rapide d'une population isolée, promouvoir une cascade de changements génétiques appelée « révolution génétique » et même conduire à la spéciation (Provine 2004), bien que cette dernière hypothèse demeure controversée (Barton and Charlesworth 1984; Charlesworth 1995; Slatkin 1996; Templeton 2008). L'effet fondateur est un mécanisme important en génétique des populations et en génétique humaine parce qu'il façonne les patrons de diversité génétique et peut ainsi affecter la diversité phénotypique des populations et des individus. De plus, les signatures génomiques laissées par l'effet fondateur offrent l'opportunité de reconstruire notre passé évolutif. Olson a récemment proposé que l'effet fondateur serait parmi les processus évolutifs les plus importants dans le façonnement de la diversité génétique et phénotypique interindividuelle chez l'humain, en combinaison avec l'équilibre mutation-sélection (Olson 2012). D'autres ont aussi avancé que l'effet fondateur aurait fortement influencé l'évolution des primates et des humains (Harris 2010). Tester ces hypothèses d'évolution humaine requiert de pouvoir différencier les conséquences génétiques de l'effet fondateur de celles des autres forces évolutives, un problème qui s'avère bien loin d'être trivial.

L'effet fondateur est souvent défini comme une forme de dérive génétique, correspondant à la catégorie « évènements accidentels » de la classification de Wright des mécanismes d'évolution purement neutralistes, n'impliquant aucune force de sélection (par exemple, Roberts 1968). Le hasard joue

[1] La dérive génétique est le processus évolutif qui implique la transmission aléatoire des variants génétiques des parents aux enfants. Dans une population de taille finie, la dérive génétique cause des changements stochastiques dans la composition génétique d'une génération à l'autre.

vraisemblablement un rôle primordial dans la détermination des allèles représentés dans la nouvelle population lors de l'échantillonnage des fondateurs. Par la suite, la dérive génétique est impliquée dans la redistribution de la diversité génétique d'une génération à l'autre. Or, lors d'un événement fondateur, il n'est pas exclu que la sélection naturelle puisse entrer en jeu, notamment lors de la sélection des fondateurs, de leur migration et de leur adaptation au nouvel environnement, tel que Mayr le croyait (Provine 2004). Il semble donc qu'une certaine ambiguïté dans la notion d'effet fondateur provienne de sa classification parmi les mécanismes évolutifs adaptatifs ou non-adaptatifs.

Dans cette thèse, j'adopterai une définition plus inclusive de l'effet fondateur, à l'instar de Mayr. L'effet fondateur sera défini comme l'ensemble des conséquences génétiques et phénotypiques de la fondation d'une nouvelle population par un nombre limité[2] de fondateurs et de son expansion dans les générations subséquentes. Ainsi, ce sont les changements génétiques ayant cours dans la nouvelle population qui seront qualifiés d'adaptatifs s'ils sont causés par la sélection et de non-adaptatifs ou neutres s'ils sont engendrés par des forces neutres, dépendantes de l'histoire démographique. Caractériser les forces neutres et non neutres agissant sur les populations fondatrices exigera alors, d'une part, de pouvoir estimer l'impact des composantes de l'histoire démographique sous-jacentes à l'effet fondateur, soit l'effet de la réduction de taille, l'effet de l'isolement ou de migrations subséquentes et l'effet de la croissance de la population dans les générations suivant la fondation et, d'autre part, de pouvoir comprendre la dynamique spatio-temporelle de l'expansion territoriale de la nouvelle population suivant sa fondation. Ces différentes composantes de l'effet fondateur ont été étudiées théoriquement par de nombreux auteurs. Dans la section qui suit, je décris les prédictions théoriques de ces études.

---

[2] Par « limité » j'entends un effectif de fondateurs substantiellement plus petit que la taille de la population source. Ceci est important puisque l'effet fondateur comprend une notion de réduction de taille de population.

2. Les prédictions théoriques de l'effet fondateur

Les conséquences génétiques des peuplements fondateurs dépendent largement de l'effectif de fondateurs et du nombre de générations depuis la fondation. Modulé selon ces facteurs, l'effet fondateur peut causer la réduction du nombre d'allèles et de l'hétérozygotie de la nouvelle population par rapport à sa population source, ainsi que l'altération de la distribution des fréquences alléliques et haplotypiques. L'effet fondateur peut aussi entrainer l'augmentation en fréquence de certains allèles rares, comme une mutation causale d'une maladie mendélienne. Par ailleurs, l'histoire démographique suivant la fondation, notamment les expansions démographiques et spatiales ou l'isolement ou non des sous-populations, peut faire varier l'impact de l'effet fondateur et entrainer des changements tout aussi significatifs sur les patrons de diversité génétique de la population. Dans cette section, je présente les études théoriques de l'effet fondateur et de leurs prédictions.

La réduction de taille d'une population

Lors d'un événement fondateur, seulement une fraction des individus provenant de la population source migre sur un nouveau territoire pour former une population nouvelle. L'effet fondateur implique donc une réduction dans la taille de la population. On parle aussi de « bottleneck » ou goulot d'étranglement démographique en référence à un événement de réduction de taille d'une population[3]. Sous l'hypothèse que les fondateurs sont tirés au hasard parmi la population source avec une probabilité uniforme, alors la fraction du génome tirée de la population source dépend du nombre de fondateurs et de la diversité génétique de la population source. En supposant

[3] Je souligne ici que l'effet fondateur n'est cependant pas synonyme à « bottleneck ». Alors qu'un mouvement migratoire est à l'origine de la fondation d'une nouvelle population, un goulot d'étranglement démographique peut avoir lieu sans que la population ne bouge de territoire, comme dans le cas d'une catastrophe par exemple. Ainsi, les phénomènes de migration, dispersion et même de sélection ne sont pas les mêmes pour les fondateurs arrivant dans un nouveau territoire que pour les survivants à un bottleneck.

que la population source est panmictique[4], les fondateurs apporteront un échantillon aléatoire des variants génétiques, ou allèles, présents à un locus donné parmi la population source (Figure 1). Plus le nombre de fondateurs est grand, plus la probabilité est grande que les allèles soient retrouvés parmi les fondateurs à une fréquence similaire que dans la population source. À l'opposé, plus le nombre de fondateurs est petit, plus la probabilité est grande que certains allèles ne soient pas échantillonnés parmi les fondateurs et que les fréquences alléliques soient altérées par rapport à la population source.

Réduction de la taille d'une population suivie d'une expansion démographique

En 1975, Nei, Maruyama et Chakraborty ont étudié les changements génétiques dus au hasard de la réduction de la taille d'une population et de la dérive génétique dans les générations suivantes (Nei et al. 1975). Dans un scénario classique incluant la réduction soudaine de la taille de la population de N_0 à N_1, suivie d'une expansion démographique soutenue par une croissance logistique de la population, ces auteurs ont démontré que la réduction du nombre d'allèles par locus est plus marquée que la perte d'hétérozygotie[5]. Néanmoins, pour obtenir une réduction drastique de la variabilité génétique, N_1, doit être très petit par rapport à N_0 et le taux de croissance doit demeurer relativement faible dans les générations subséquentes. Sinon, la réduction de l'hétérozygotie, ou inversement, l'augmentation de l'homozygotie, est faible. Lorsque la population croît dans les générations suivantes, le nombre d'allèles augmente (via de nouvelles mutations) alors plus rapidement que l'hétérozygotie. Ces auteurs ont aussi spéculé que les allèles rares seraient plus affectés par l'effet fondateur que les allèles à fréquence intermédiaire ou les allèles communs (Nei et al. 1975).

---

[4] Au sein d'une population panmictique, les unions entre individus sont aléatoires et les individus ont une chance égale de se reproduire, de sorte que les individus sont formés par le tirage indépendant des gamètes parentaux. On peut alors considérer que les allèles portés par les individus forment un pool d'allèles indépendants, dit « pool génique ».

[5] L'hétérozygotie est une mesure de diversité génétique qui décrit la probabilité de tirer au hasard deux allèles différents dans un échantillon.

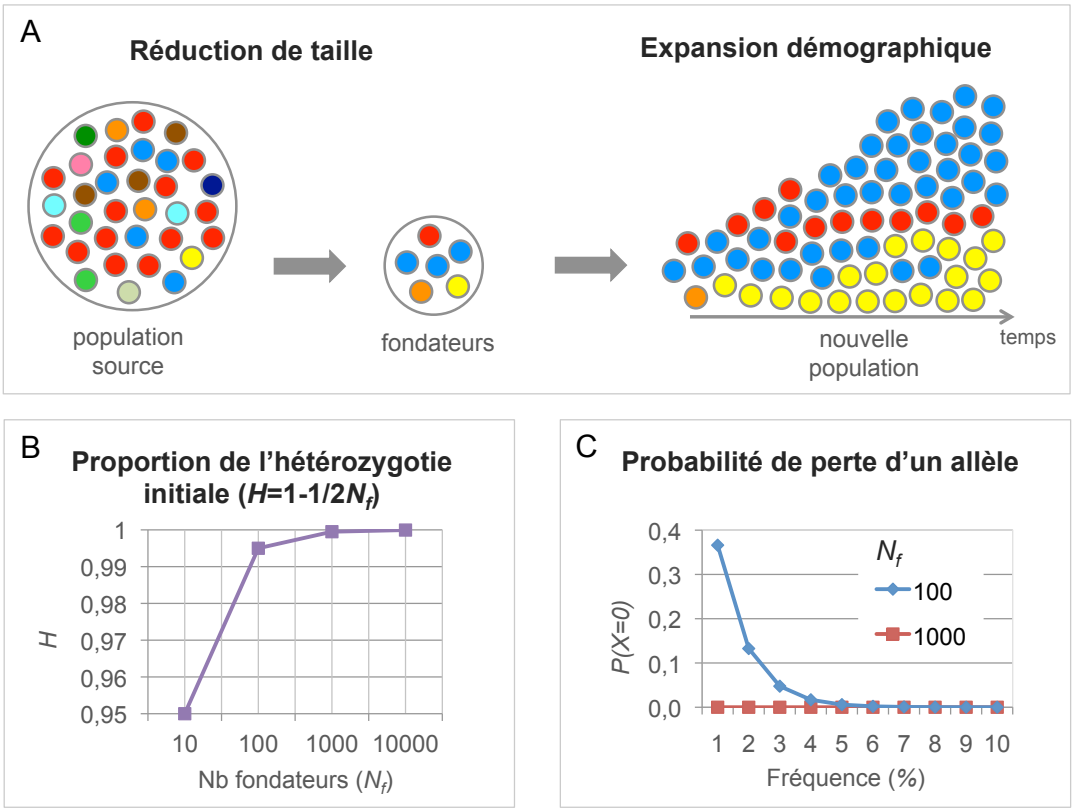

Figure 1. Effet de la réduction de taille d'une population.

(A) Représentation schématique de la fondation d'une nouvelle population par un tirage aléatoire d'allèles neutres parmi la population source. **(B)** Proportion de l'hétérozygotie de la population source (hétérozygotie initiale) attendue parmi N_f fondateurs échantillonnés au hasard (H=1-1/2N_f) (Allendorf 1986). **(C)** Probabilité de perte d'un allèle, selon sa fréquence dans la population source.

Cette prédiction a été confirmée par de nombreuses études théoriques qui ont illustré l'impact de l'effet fondateur sur le spectre de fréquences alléliques, défini comme la distribution du nombre de copies des allèles ou de la fréquence des allèles dans un échantillon (notamment Thompson and Neel 1978; Watterson 1984; Maruyama and Fuerst 1985a; Allendorf 1986; Tajima 1989; Cornuet and Luikart 1996; Luikart et al. 1998; Marth et al. 2004). Ainsi,

une réduction de taille sévère résultera en un déficit en allèles rares comparativement à une population idéale avec la même hétérozygotie sous l'équilibre mutation/dérive[6]. Ce déficit pourra persister longtemps (entre 2N et 4N générations) si la population ne croît pas suivant sa réduction (Maruyama and Fuerst 1985a), mais sera rétabli plus rapidement si la population subit ensuite une expansion démographique qui, à l'opposé, après un nombre suffisant de générations, entrainera un excès d'allèles rares (Maruyama and Fuerst 1984; Maruyama and Fuerst 1985a). Les allèles rares parmi les fondateurs pourront avoir une fréquence accrue par rapport à la population source et même augmenter dans les générations suivantes jusqu'à une fréquence élevée à cause de la dérive génétique (Thompson and Neel 1978; Luikart et al. 1998). Bien que les modèles théoriques présentés ci-haut aient portés sur des allèles neutres, ces résultats sont aussi attendus pour des allèles sélectionnés, puisqu'une réduction de taille diminue la capacité de la sélection naturelle à éliminer les allèles délétères et à fixer les allèles avantageux (Wright 1931; Robertson 1960; Otto and Whitlock 1997). Une croissance rapide suivant la fondation limitera aussi l'élimination des allèles délétères (Livingstone 1970; Otto and Whitlock 1997), de sorte que certains allèles délétères portés par les fondateurs pourront se propager parmi leurs descendants et croître en fréquence. Ceci explique que l'effet fondateur soit évoqué comme cause lorsqu'on observe la fréquence élevée d'une maladie mendélienne ou d'une mutation causale dans une population connue pour avoir été récemment fondée.

Un événement fondateur peut aussi altérer la distribution des haplotypes[7] et ainsi augmenter le déséquilibre de liaison[8]. Lorsqu'un allèle unique est introduit dans la nouvelle population sur un chromosome fondateur (par

[6] L'équilibre mutation/dérive est atteint lorsque la perte d'allèles par la dérive génétique est égale à leur accumulation par mutation.
[7] Un haplotype est une combinaison de variants génétiques à des sites adjacents sur un même chromosome.
[8] Le déséquilibre de liaison est défini comme l'association non-aléatoire entre les allèles situés à deux loci ou plus.

exemple, une mutation délétère), il sera lié génétiquement à la combinaison d'allèles présents sur ce chromosome. Au fil des générations, la recombinaison méiotique va progressivement briser ces associations et réduire le déséquilibre de liaison, en formant de nouvelles combinaisons des haplotypes parentaux lors de la production des gamètes. Différentes statistiques, notamment r^2 et *D'*, mesurent le déséquilibre de liaison entre les paires d'allèles à deux loci séparés par une distance génétique donnée (revu par Slatkin 2008). Le taux de recombinaison local, l'âge des allèles et l'histoire démographique déterminent notamment l'étendue du déséquilibre de liaison, qui décroît avec l'augmentation de la distance génétique entre les loci (Labuda et al. 1997; Pritchard and Przeworski 2001; Ardlie et al. 2002; Nordborg and Tavare 2002). Le modèle théorique de Kruglyak a démontré qu'une réduction de taille sévère peut augmenter le déséquilibre de liaison en moyenne sur l'ensemble du génome (Kruglyak 1999b; Kruglyak 1999a). Ceci peut s'expliquer par la réduction de taille qui cause une perte d'haplotypes lors de la fondation et par une dérive génétique accrue si la population reste de petite taille durant les générations subséquentes. L'augmentation du déséquilibre de liaison causé par l'effet fondateur se traduira par un plus grand nombre et de plus longs haplotypes partagés entre les paires d'individus au sein d'une population, hérités d'ancêtres communs depuis la fondation. De la même façon, chez un individu donné, la signature de l'effet fondateur sera un plus grand nombre de segments homozygotes, partagés entre les chromosomes homologues maternel et paternel. La taille des segments partagés par deux individus dépend du nombre de générations les reliant à l'ancêtre commun. L'âge de la population fondatrice a donc un impact sur le déséquilibre de liaison, mais de nombreux autres facteurs entre ici en jeu, notamment les mariages entre apparentés, le métissage et la stratification démographique (Clark 1999; Pritchard and Przeworski 2001). En particulier, à un locus donné, l'effet d'un événement fondateur sur l'étendue du déséquilibre de liaison dépend de la fréquence des allèles. Un fort déséquilibre de liaison devrait entourer un allèle initialement rare apporté par

un fondateur ou un allèle apparu dans les générations subséquentes (Labuda et al. 1996; Kruglyak 1999b; Kruglyak 1999a; Slatkin 2008). Cependant, le déséquilibre de liaison entourant les allèles communs ne devrait pas être substantiellement différent de la population source (Kruglyak 1999b; Kruglyak 1999a; Pritchard and Przeworski 2001). Même un petit nombre de fondateurs devrait apporter un échantillon suffisamment grand de chromosomes porteurs de l'allèle commun pour que les proportions haplotypiques ne soient pas altérées (Kruglyak 1999b; Kruglyak 1999a; Pritchard and Przeworski 2001).

Isolement versus migrations subséquentes et métissage

Les modèles d'effet fondateur présentés ci-haut peuvent être qualifiés d'extrêmes : on y suppose que les populations nouvelles sont formées à partir d'un seul mouvement de migration fondatrice et demeure isolées dans les générations subséquentes. Un tel scénario peut être plausible pour certaines populations insulaires, mais peut cependant être inapplicable pour d'autres populations fondatrices, comme par exemple pour les peuplements coloniaux récents qui ont apportés un flot continu de migrants ou les isolats religieux qui ne sont pas complètement fermés aux conversions religieuses. L'isolement d'une population conduira inévitablement à une perte de diversité due à la dérive génétique (Wright 1931) et à la sélection directionnelle (positive ou négative). Mais cette perte sera freinée par l'intégration de nouveaux migrants, qui augmenteront directement la taille de la population et introduiront de la diversité génétique (Barton 2008). Le flux génétique entrant dans la population fondatrice dépendra alors non seulement de la proportion relative du nombre de migrants sur l'effectif de la population, mais aussi des modalités de leur éventuelle intégration dans la population d'accueil ainsi que de leur succès reproductif (Slatkin 1987). Les scénarios simples d'événement fondateur supposent aussi que les fondateurs proviennent tous de la même population source. Or, la réduction de la diversité génétique peut être compensée par un métissage parmi les fondateurs ou leurs descendants, qui augmentera le nombre d'allèles, le niveau d'hétérozygotie, mais aussi

l'étendue du déséquilibre de liaison. L'effet du métissage sur les patrons de diversité dépend largement de la différenciation génétique entre les populations d'origine, mais aussi du nombre de populations d'origine et de la dynamique temporelle des événements (Jin et al. 2012).

Événements fondateurs successifs et expansion territoriale

Des événements fondateurs successifs ont lieu notamment lors de la colonisation successive de plusieurs îles (Clegg et al. 2002), ou des cycles successifs de croissance et de réduction de taille qui implique une suite de goulots d'étranglement démographiques (Thompson and Neel 1978; Maruyama and Fuerst 1985b). Une série d'événements fondateurs est aussi observée lors d'une expansion territoriale; i.e. lorsque la croissance de la population est accompagnée d'une expansion de son aire de répartition géographique. Les expansions territoriales peuvent être stimulées par les changements climatiques ou par l'ouverture de nouvelles niches (géographiques et/ou économiques). Les expansions territoriales sont notamment observées chez les espèces invasives et lors de certains mouvements de colonisation.

Des événements fondateurs successifs causeront une série d'effets fondateurs dont les conséquences seront consécutivement amplifiées. Ramachandran et ses collaborateurs ont proposé un modèle simple de « serial founder effect » qui montre que chaque nouvel événement fondateur devrait être accompagné d'une réduction de l'hétérozygotie et d'une augmentation de la différenciation génétique (Ramachandran et al. 2005). Thompson et Neel ont démontré que les allèles introduits par les fondateurs adultes auront une plus grande probabilité de survie et d'augmentation en fréquence que les mutations *de novo,* (les nouvelles mutations dans les lignées germinales transmises aux enfants), puisqu'une large fraction de celles-ci ne sera pas transmise à cause de la mortalité infantile (Thompson and Neel 1978). Plus récemment, les modèles théoriques d'événements

fondateurs successifs ont été élaborés en incluant les dimensions temporelle et spatiale des expansions territoriales.

Les expansions territoriales peuvent causer une structuration géographique des fréquences alléliques bien différente des patrons attendus pour une population à l'équilibre, ou même de modèles d'expansion démographique ne tenant pas compte de la dynamique temporelle et spatiale (Excoffier and Ray 2008; Slatkin and Excoffier 2012). En particulier, on s'attend à une réduction progressive de l'hétérozygotie plus on s'éloigne du lieu d'origine de l'expansion (Austerlitz et al. 1997; DeGiorgio et al. 2011; Slatkin and Excoffier 2012). Au front de la vague d'expansion, des allèles initialement rares peuvent se propager sur de longues distances et atteindre des fréquences élevées, un phénomène appelé « surfing génétique » (Edmonds et al. 2004; Klopfstein et al. 2006; Excoffier and Ray 2008; Slatkin and Excoffier 2012). Des allèles neutres, favorables ou même délétères peuvent augmenter en fréquence à cause du *surfing* (Klopfstein et al. 2006; Travis et al. 2007; Hallatschek and Nelson 2010; Hallatschek 2011; Slatkin and Excoffier 2012), suggérant que les populations au front de l'expansion peuvent porter des mutations avec un spectre de coefficients de sélection plus large que les populations au cœur du peuplement. Le *surfing* génétique, illustré à la Figure 2, est dû à une séquence d'événements fondateurs et une plus forte dérive dans les populations marginales au front de l'expansion qui ont une taille et une densité réduite (Klopfstein et al. 2006; Excoffier and Ray 2008; Slatkin and Excoffier 2012). La composition génétique de ces populations marginales déterminera aussi la diversité génétique propagée par les colonisateurs qui sont recrutés au front de l'expansion (Excoffier and Ray 2008). En outre, les expansions territoriales peuvent conduire à des différences drastiques entre la population au cœur et celle au front de l'expansion, qui prendront la forme de clines marqués dans les fréquences alléliques, imitant la signature de la sélection positive (Excoffier and Ray 2008).

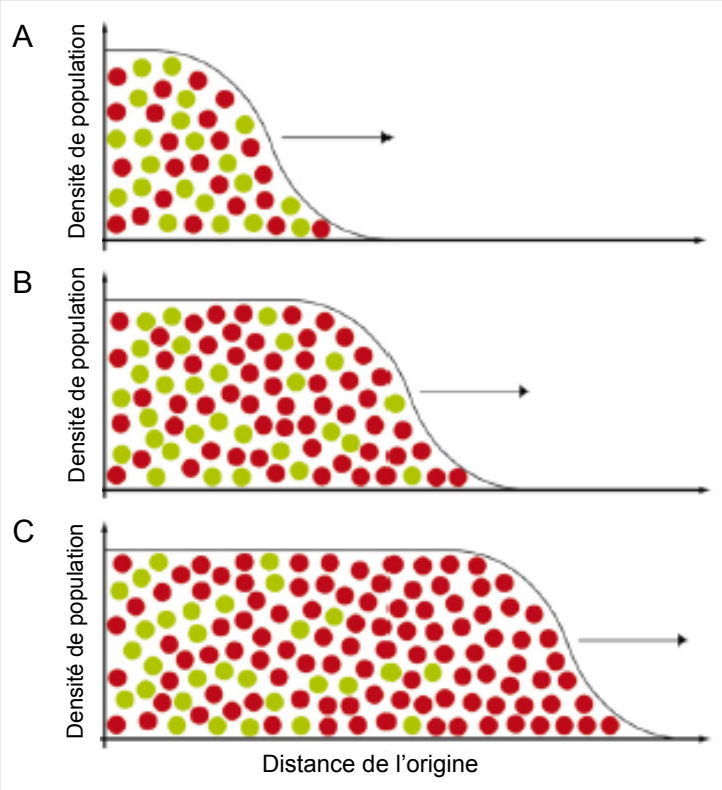

Figure 2. Dérive génétique et *surfing* génétique au front de la vague d'expansion territoriale.

(A) Population au début de l'expansion territoriale, initialement avec une fréquence égale de deux allèles (vert et rouge). **(B)** L'allèle rouge retrouvé par hasard au front de la vague d'expansion en (A) augmente en fréquence dû à la dérive génétique accrue par la faible densité au front. **(C)** L'allèle rouge est fixé par la dérive au front de la vague. Schéma redessiné à partir de Excoffier and Ray 2008.

Autres facteurs réduisant l'effectif efficace d'une population

En somme, les fluctuations de taille d'une population, tels qu'un événement fondateur ou une expansion démographique, affectent le nombre d'individus participants à la reproduction, qui module l'ampleur de la dérive génétique et l'efficacité de la sélection. Dans les populations naturelles, d'autres facteurs

démographiques peuvent aussi réduire localement l'effectif efficace d'une population, N_e, défini comme le nombre d'individus dans une population idéale pour lequel on aurait le même taux de dérive génétique que celui observé dans la population réelle étudiée (Wright 1931; Charlesworth 2009). Dans une population fondatrice, par exemple, un rapport de masculinité en déséquilibre parmi les fondateurs ou dans les générations subséquentes limitera le nombre d'unions possibles, si bien que la taille de recensement, N, sera plus grande que N_e (Charlesworth 2009). Les pratiques de mariage entre consanguins réduiront aussi la taille efficace de la population, à cause de la corrélation entre les génomes maternels et paternels, qui se traduira par une plus forte dérive génétique (Wright 1933). Une structuration géographique des unions aura le même effet, en subdivisant la population totale en de plus petites unités géographiques plus ou moins isolées (Wright 1943; Charlesworth 2009; Holsinger and Weir 2009). De plus, l'effectif efficace sera réduit par une variation non aléatoire dans la taille des familles, lorsque les individus n'ont pas tous une chance égale de se reproduire (Wright, 1938). Un succès reproducteur non aléatoire pourrait notamment être causé par la transmission intergénérationnelle de la fécondité : les individus nés de familles peu nombreuses sont eux aussi portés à avoir un petit nombre d'enfants alors qu'à l'inverse les individus nés de familles nombreuses ont tendance à avoir un grand nombre d'enfants. Une corrélation intergénérationnelle dans le nombre total d'enfants et dans le nombre d'enfants mariés (dit le nombre d'enfants utiles) est documentée chez certaines populations humaines et pourrait refléter une transmission culturelle ou biologique des comportements reproducteurs (Austerlitz and Heyer 1998; Sibert et al. 2002; Pluzhnikov et al. 2007; Brandenburg et al. 2012).

Ces facteurs démographiques affectent l'ensemble du génome, mais, à un locus donné, leurs effets dépendent du mode de transmission héréditaire : autosomique, lié au chromosome X, Y ou mitochondrial. Par exemple, puisque les hommes n'héritent que d'un chromosome X, l'effectif efficace du chromosome X est le ¾ de l'effectif efficace des autosomes sous l'hypothèse

de succès reproducteur égal entre les sexes (Vicoso and Charlesworth 2009; Labuda et al. 2010). De plus, l'effectif efficace varie le long du génome à cause de l'effet d'entraînement de la sélection à un site sur les sites adjacents, appelé « hitchhicking » génétique dans le cas de la sélection balancée ou positive (Maynard Smith and Haigh 1974) et de la sélection « background » dans le cas de la sélection purificatrice (McVean and Charlesworth 2000; Comeron et al. 2008). Ces facteurs génétiques modulent l'intensité de l'effet fondateur le long du génome. Pour obtenir une image globale de son impact sur le génome, il convient donc d'analyser conjointement de multiples loci, couvrant les autosomes, les chromosomes sexuels et l'ADN mitochondrial. De plus, la comparaison de la variation génétique neutre (non-codante) et la variation fonctionnelle facilite la distinction entre les effets dus aux facteurs démographiques des effets de la sélection, puisque celle-ci n'agit essentiellement que dans les régions fonctionnelles du génome.

3. Évaluer l'effet fondateur dans les populations naturelles

Les prédictions théoriques de l'effet fondateur ont depuis longtemps été démontrées expérimentalement en utilisant des organismes modèles, où les populations fondatrices sont créés artificiellement par la sélection de certains individus à l'origine de lignées de reproduction, notamment chez la drosophile (e.g. Dobzhansky and Pavlovsky 1957). Les études théoriques et expérimentales sont claires sur les changements génétiques qui peuvent avoir lieu sous différents scénarios extrêmes de fondation. Pourtant, la nature et l'ampleur des changements ayant été causés par l'effet fondateur au cours de l'évolution restent à ce jour largement méconnus. Le problème c'est que l'effet fondateur est difficile à mesurer dans les populations naturelles. Les événements de fondation peuvent être observés en temps réel chez les espèces avec un intervalle intergénérationnel court, notamment chez certaines populations invasives (Pysek and Hulme 2005; Estoup et al. 2010), pour lesquelles on a accès à la fois à la population pré-fondation et celle post-

fondation. Cependant, pour la majorité des espèces, comme pour la plupart des populations humaines[9], l'information génétique n'est disponible que sur les individus contemporains; les événements fondateurs sont ainsi plus souvent évalués à posteriori. Or, les patrons de variabilité génétique ne révèlent pas directement si un événement fondateur a eu lieu, ni même la nature et l'ampleur de ses conséquences. Les modèles d'effet fondateur et d'autres forces évolutives sont utilisés pour prédire quelle réduction de taille et quel taux de croissance subséquente doivent avoir eu lieu pour causer les patrons observés dans les données. Les méthodes indirectes pour évaluer l'effet fondateur reposent, d'une part, sur l'étude de maladies mendéliennes et des mutations causales ayant une fréquence accrue au sein d'une population fondatrice, et d'autre part, sur l'analyse de multiples loci du génome. Par ailleurs, l'étude des généalogies profondes offre l'opportunité d'évaluer directement les modalités démographiques des événements fondateurs et d'en estimer indirectement les conséquences génétiques.

L'étude des maladies mendéliennes

Au sein de nombreuses populations fondatrices, on observe une fréquence accrue de maladies mendéliennes[10] qui sont ailleurs plus rares, voire jamais observées (Arcos-Burgos and Muenke 2002). Pendant des années, les populations fondatrices ont été au cœur des études de cartographie et d'identification des mutations causales des maladies mendéliennes (Peltonen et al. 2000; Heutink and Oostra 2002; Chong et al. 2012). Les avantages potentiels principaux des populations fondatrices sont leur présumée plus grande homogénéité génétique (ce qui laisse supposer une moindre hétérogénéité génétique au locus causal, une fréquence plus élevée des

[9] En plus des données généalogiques et historiques, chez certaines populations humaines, il est possible d'étudier l'ADN ancien provenant de squelettes et/ou restes humains momifiés.
[10] Une maladie mendélienne présente, au sein d'une même famille, des ratios de ségrégation mendélien qui suggèrent une mutation dans un seul gène causal dont la transmission héréditaire peut être autosomique récessive, dominante, lié au chromosome X, ou mitochondriale.

mutations causales et le déséquilibre de liaison s'étendant sur de plus longues distances génétiques), leur environnement commun parfois moins variable (religion, culture, mode de vie partagés) et l'accès à des généalogies profondes (Peltonen et al. 2000; Shifman and Darvasi 2001; Arcos-Burgos and Muenke 2002; Heutink and Oostra 2002; Cannon-Albright et al. 2005). L'effet fondateur est souvent évoqué comme cause de la fréquence accrue des maladies mendéliennes, mais cette hypothèse a rarement été testé explicitement (Anderson and Slatkin 2007). En théorie, l'effet fondateur peut entrainer une augmentation en fréquence des mutations légèrement ou modérément délétères, tel qu'attendu pour les maladies récessives (qui ne sont pas délétères chez les porteurs hétérozygotes), les maladies dominantes n'affectant pas ou peu le succès reproducteur et les maladies ayant une apparition tardive.

Une mutation fondatrice est souvent définie comme une mutation partagée en copies identiques entre les patients atteints d'un désordre monogénique, héritée d'un ancêtre ou fondateur commun. Mais cette définition porte à confusion : bien que cette observation soit nécessaire, elle n'est pas suffisante pour confirmer l'effet fondateur. Une fois l'hypothèse de mutations récurrentes exclue, les copies identiques d'une mutation remontent nécessairement à un ancêtre commun qui peut être bien plus ancien que la fondation. En revanche, l'analyse des haplotypes entourant une mutation causale permet de tester l'hypothèse qu'une mutation causale a augmenté en fréquence à cause de l'effet fondateur. La caractérisation du nombre, de la fréquence et de la distribution géographique des haplotypes partagés permet d'évaluer si la mutation causale a été introduite par un seul ou par plusieurs fondateurs et si les descendants se retrouvent dans l'ensemble de la population ou sont limités à des sous-populations (Labuda et al. 1996; Slatkin 2004; Yotova et al. 2005). La cooccurrence de plusieurs haplotypes fondateurs suggérant de multiples introductions d'une mutation, peut à priori être interprétée comme contradictoire avec l'hypothèse de l'effet fondateur. Pourtant, un haplotype mineur a davantage l'occasion d'être détecté

lorsqu'un haplotype fondateur majeur démontre une fréquence accrue, un scénario qui est compatible avec l'effet fondateur (Yotova et al. 2005). D'autre part, différentes méthodes d' « horloge moléculaire » permettent d'estimer le nombre de générations écoulées depuis l'apparition d'une mutation sur un haplotype donné (Hastbacka et al. 1992; Labuda et al. 1996; Labuda et al. 1997; Thompson and Neel 1997; Risch et al. 2003; Slatkin 2004; Yotova et al. 2005). Ceci permet de vérifier si l'âge d'une mutation coïncide avec la chronologie historique ou archéologique d'un peuplement fondateur, dans quel cas l'hypothèse de l'effet fondateur est davantage supportée.

Les changements de fréquence observés chez les mutations causales de maladies mendéliennes présentent cependant un biais clinique : ils sont reconnus grâce à un phénotype particulier causés par des allèles rares. L'étude des maladies mendéliennes ne permet donc pas à elle seule d'estimer si ces événements sont rares et identifiés seulement à cause de leur impact observable sur le phénotype, ni de prédire si des changements sont aussi attendus pour les allèles plus communs. Ces questions sont importantes pour l'optimisation du design des études cherchant à identifier les déterminants génétiques des maladies complexes (Jorde et al. 2000; Peltonen et al. 2000; Shifman and Darvasi 2001; Heutink and Oostra 2002; Newman et al. 2004; Bourgain and Genin 2005; Kristiansson et al. 2008).

Les études de génétique médicale ont démontré qu'il peut exister une remarquable hétérogénéité des mutations dans un même gène causant une maladie mendélienne, au sein même de populations fondatrices (Chakravarti 1999; Scriver 2001; Zlotogora 2007; Ostrer and Skorecki 2013). Chez le peuple juif, par exemple, une ou deux mutations majeures par locus expliquent plus de 70% des cas de maladies mendéliennes (Ostrer and Skorecki 2013). La distribution ethnique des mutations et les estimations de leurs âges qui concordent avec les diasporas majeures suggère que l'effet fondateur est la cause de cette forte prévalence de ces maladies (Ostrer and

Skorecki 2013). Or, l'observation de multiples mutations causales a été interprétée pour certaines maladies comme suggérant un avantage adaptatif des porteurs hétérozygotes (revu par Zlotogora 2007), qui aurait favorisé leur maintien et leur augmentation en fréquence. Cette hypothèse de la sélection balancée est certainement séduisante, mais un avantage adaptatif des hétérozygotes n'a été démontré qu'en de rares occasions (Alves et al. 2012). Enfin, d'autres hypothèses doivent être considérées pour expliquer la prévalence accrue de maladies mendéliennes au sein d'une population fondatrice. Notamment, l'incidence de maladies mendéliennes peut être élevée à cause de pratiques répandues de mariage entre consanguins (Zlotogora et al. 2007), ou d'unions préférentielles entre individus partageant un même phénotype, tel qu'observé pour une forme autosomique récessive de surdité (Nance and Kearsey 2004).

Les études génomiques

De nombreux tests statistiques ont été développés pour détecter l'effet fondateur dans les données génomiques et estimer la nature et l'ampleur des changements résultants. Spécifiquement, il s'agit de tester si les patrons de variabilité génétique observés dans les populations actuelles sont conformes avec l'hypothèse d'équilibre ou plutôt avec les prédictions de modèles théoriques d'événement fondateurs (Gattepaille et al. 2013). Les tests de détection de l'effet fondateur sont basées sur le spectre de fréquences alléliques (ou certaines statistiques descriptives) (Watterson 1984; Tajima 1989; Cornuet and Luikart 1996; Luikart et al. 1998; Fay and Wu 1999; Marth et al. 2004; Gutenkunst et al. 2009), le déséquilibre de liaison (Thompson and Neel 1997; Reich et al. 2009; McEvoy et al. 2011) ou les patrons de diversité haplotypique, incluant les haplotypes partagés et les segments d'homozygotie (Depaulis et al. 2003; Lohmueller et al. 2009; Kirin et al. 2010; Gusev et al. 2012). De plus, à l'aide de ces patrons de variabilité, certains paramètres des événements fondateurs peuvent être approximés, tels que la réduction de taille, le taux de croissance démographique et le nombre de

générations depuis la fondation, à l'aide de méthodes d'inférences basées sur la coalescence (revu par Rosenberg and Nordborg 2002; Marjoram and Tavare 2006; Pool et al. 2010; Ho and Shapiro 2011). La théorie de la coalescence fournit une description statistique des relations ancestrales existant dans un échantillon de segments génomiques de façon rétrospective (Kingman, 1982; Wakeley, 2009) et permet de simuler des données génomiques sous différents scénarios démographiques, incluant des fluctuations de taille. Les méthodes de détection et d'inférence peuvent néanmoins être biaisées si on néglige la structure des populations étudiées ou le métissage parmi les fondateurs et/ou leurs descendants (Gattepaille et al. 2013). De plus, ces méthodes ne peuvent pas être appliquées rigoureusement aux populations fondatrices récentes puisque la coalescence est peu satisfaisante pour approximer le passé récent, pour moins de $\log_2(N_e)$ générations (Wakeley et al. 2012).

L'étude des généalogies profondes

Les paramètres démographiques d'un peuplement fondateur récent peuvent être directement estimés lorsque l'arbre généalogique d'une population est connu. La généalogie des membres d'une population représente l'ensemble des relations familiales les reliant à travers les générations. Les généalogies sont typiquement reconstruites à l'aide de la culture orale, des actes religieux, des actes de l'État civil et/ou de documents historiques. À l'aide de ces sources, le nombre de fondateurs, leurs dates d'arrivée et leurs origines géographiques peuvent être directement caractérisés. Cependant, la précision de ces estimations dépend de façon cruciale de la qualité et de la richesse des sources ainsi que de leur profondeur temporelle. Par exemple, on peut raisonnablement supposer que la tradition orale plafonne souvent aux ancêtres présents à la cinquième ou sixième génération et ne pourrait alors pas permettre de reculer jusqu'aux fondateurs d'une population, ce qui est nécessaire pour évaluer l'effet fondateur. Par ailleurs, les fondateurs n'ont pas une descendance égale dans la population contemporaine. Le nombre

de descendants de chaque fondateur et leur distribution géographique peuvent être étudiés à l'aide de la généalogie de la population. Pour caractériser un peuplement fondateur, de nombreux autres paramètres démographiques peuvent être étudiés à l'aide de données généalogiques, par exemple les trajectoires migratoires, le taux de croissance et la structure des familles, incluant les pratiques de mariage et la fécondité. Ces informations sont essentielles pour comprendre comment la généalogie d'une population est tissée et quels sont les rôles respectifs de la variation démographique et de la sélection dans l'assemblage des liens ancestraux. Étonnamment, ces questions demeurent largement méconnues à ce jour et n'ont reçu que très peu d'attention si on compare au nombre d'études s'intéressant à la généalogie des gènes décrite par la théorie de la coalescence (Barton and Etheridge 2011; Wakeley et al. 2012). Néanmoins, les estimations démographiques ne révèlent pas directement quelle diversité génétique a été introduite par les fondateurs, ni quelles sont les conséquences des événements de fondation sur la composition et la structuration de la variabilité génétique des populations, qui sont des éléments clés pour expliquer l'effet fondateur.

La généalogie d'une population contient les chemins de transmission des gènes. Les lignées généalogiques tracent directement les lignées génétiques maternelles (ADN mitochondrial) et les lignées paternelles (chromosome Y). Par contre, la transmission autosomique n'est pas directement retraçable dans les généalogies, puisqu'un parent ne passe que la moitié de son génome autosomique à chaque enfant. Les études généalogiques qui ont porté sur les conséquences génétiques d'un peuplement fondateur reposent essentiellement sur : (i) le coefficient d'apparentement et le coefficient de consanguinité, ainsi que (ii) les mesures de contribution génétique et (iii) les

simulations « allele dropping » de la transmission d'allèles fondateurs conditionnelle à la structure de la généalogie[11].

Les mesures de consanguinité et d'apparentement estiment la proportion du génome hérité d'ancêtres communs retracés dans la généalogie, respectivement chez un individu et entre les paires d'individus (Wright 1922; Malécot 1948; Thompson 1986). Elles permettent d'évaluer la perte de diversité causée par la dérive génétique depuis la fondation et distinguer si elle est due aux ancêtres partagés dans les générations proches ou éloignées. De nombreuses études ont ainsi caractérisé la consanguinité et l'apparentement de populations fondatrices à l'aide de généalogies extensives remontant jusqu'à la fondation (e.g. Martin 1970; Thompson and Roberts 1980; Tremblay et al. 2008). Elles ont notamment corroboré la plus grande homozygotie attendue pour les populations de plus petite taille, chez des sous-populations issues d'un plus petit nombre de fondateurs ou d'une plus faible croissance (Martin 1970; Tremblay et al. 2008), qui peuvent être préférées pour les études de cartographie par homozygotie (« homozygosity mapping »). Un écart a été observé entre les estimations génomiques de consanguinité et d'apparentement et l'espérance de ces quantités calculée exactement dans la généalogie, remettant en question la valeur des estimations généalogiques (e.g. Leutenegger et al. 2003; Carothers et al. 2006). Un tel écart sera observé même si les liens généalogiques connus sont exacts, parce que la recombinaison cause une variation dans la proportion du génome hérité d'un ancêtre donné (Donnelly 1983), il sera toutefois accentué par des faux liens généalogiques. Malgré cette limite, l'étude des généalogies offre l'avantage de pouvoir définir exactement quels sont les ancêtres partagés, ce qui est actuellement impossible avec les données génomiques au-delà de quelques générations.

---

[11] Bien qu'elles soient relativement plus rares, des études sur les mouvements migratoires ont aussi été réalisées au sein de populations fondatrices (ex. Gradie M, Jorde LB, Bouchard G. 1991. La structure génétique de la population du Saguenay. In *Histoire d'un génome Population et génétique dans l'est du Québec*, (ed. G Bouchard, M De Braekeleer), pp. 254-277. Presses de l'Université du Québec, Sillery, Québec.).

Avec l'hérédité mendélienne, la généalogie délimite le passage des gènes : la dérive génétique et la sélection peuvent agir uniquement à travers les différentes contributions des individus à la généalogie. La contribution génétique des fondateurs est une autre mesure utilisée pour évaluer l'effet fondateur à l'aide de généalogies. La contribution génétique est définie comme l'espérance du nombre de copies d'un allèle fondateur (ou la proportion d'un génome fondateur autosomal) transmis à ses descendants via l'enchevêtrement complexe des liens généalogiques (Roberts 1968; Barton and Etheridge 2011). Les études généalogiques de populations fondatrices basées sur cette mesure se sont surtout concentrées à décrire la contribution génétique différentielle des fondateurs (Roberts 1968; Edwards 1992; O'Brien et al. 1994; Heyer 1995; Heyer and Tremblay 1995; Labuda et al. 1996; Labuda et al. 1997). Roberts a notamment observé qu'une réduction de taille sévère et soudaine entraine un réarrangement de la contribution génétique des fondateurs (Roberts 1968). Cependant, ces études n'ont pas explicitement testé l'effet fondateur en comparant leurs résultats à un modèle de population à l'équilibre, vraisemblablement parce que la distribution théorique de la contribution génétique est encore peu étudiée (Barton and Etheridge 2011) et qu'à ma connaissance, aucun test statistique n'a encore été développé à cette fin. Enfin, à l'aide de simulations « allele dropping », des études ont exploré le destin d'allèles introduits parmi les fondateurs et évalué les changements de fréquences conditionnels à la généalogie de certaines populations (Thompson and Neel 1978; Heyer 1999; Chong et al. 2012). Ces études ont été réalisées dans le contexte des maladies héréditaires où l'on cherche spécifiquement à tester si la fréquence d'une certaine maladie mendélienne peut être expliquée par l'augmentation en fréquence d'une seule mutation apportée par un seul fondateur. Les simulations généalogiques sont prometteuses pour mieux comprendre les conséquences génétiques d'un peuplement fondateur pour les allèles rares et pour l'ensemble du spectre de fréquence (e.g. Pardo et al. 2005). Et bien que ce potentiel soit depuis longtemps évoqué (Edwards, 1968), les études de

génétique des populations qui ont explorées cette avenue sont étonnamment peu nombreuses, et un cadre théorique de la transmission des gènes à l'intérieur de la généalogie fixée d'une population n'a été élaboré que récemment (Wakeley et al. 2012).

4. Les populations fondatrices humaines

À cause du grand nombre de migrations qui ont eu lieu depuis son origine en Afrique, on peut raisonnablement supposer que l'effet fondateur est un phénomène commun au cours de l'évolution de l'Homme moderne. Cependant, chaque événement de fondation est unique et son impact dans le façonnement de la variation génétique observée aujourd'hui dépend de nombreux facteurs. Voici un très bref aperçu des grands effets fondateurs de l'ère préhistorique qui sera suivi d'un portrait sommaire d'une sélection de populations fondées récemment, soit au cours des derniers 2000 ans. Il apparaît important de souligner que d'une population à l'autre, les échelles de temps depuis la fondation varient énormément. De plus, les populations se distinguent entre elles par la nature et la géographie de leur peuplement. Ici, les populations sont présentées la plupart du temps selon la chronologie de leur fondation (mais avec des sauts irréguliers dans le temps). Je distingue d'abord les populations préhistoriques des populations fondées dans l'ère commune (ou historique). Parmi ces dernières, je présente les populations selon une grossière catégorisation géographique, soit, dans l'ordre, les populations fondatrices de l'Ancien Monde, les populations fondatrices insulaires, les isolats religieux dont les peuplements récents sont liés à des diasporas à travers le monde et enfin, les populations nées des grands mouvements de colonisation européens au Nouveau-Monde.

Depuis sa sortie d'Afrique, il y a environ 45 000 à 60 000 ans, l'Homme moderne a colonisé les autres continents par une série d'événements fondateurs (Cavalli-Sforza et al. 1994; Henn et al. 2012), d'expansions

territoriales et d'épisodes de métissage avec les populations d'Hominidés qui ont depuis disparu (Green et al. 2010; Reich et al. 2010a; Yotova et al. 2011). Les modèles d'effets fondateurs successifs et d'expansion territoriale en Europe, en Asie, en Océanie et en Amérique sont notamment soutenus par l'observation d'une réduction de l'hétérozygotie, de la présence de clines de fréquences alléliques et de l'augmentation du déséquilibre de liaison avec la distance des points d'entrées sur chaque continent (Prugnolle et al. 2005a; Ramachandran et al. 2005; Handley et al. 2007; Li et al. 2008; DeGiorgio et al. 2009; Deshpande et al. 2009; Reich et al. 2012). Néanmoins, les paramètres de ces expansions préhistoriques, tels que la datation et la localisation des points d'entrée sur les continents, de même que la sévérité des effets fondateurs et la dynamique des expansions territoriales, demeurent sujets à débat (Henn et al. 2012). En particulier, les inférences génomiques des expansions anciennes pourraient être biaisées si elles ne prennent pas en compte les événements fondateurs de l'histoire récente.

Depuis le début de l'ère commune, de nouvelles populations au cœur même des aires continentales de l'Ancien Monde ont été fondées. En Europe, c'est le cas de petites populations localisées dans les vallées de chaînes de montagne, potentiellement isolées par des barrières géographiques et parfois linguistiques, tels que la Vallée de la Valserine dans le Jura français (Lesca et al. 2008), le village de Campora en Italie (Colonna et al. 2007) et les isolats linguistiques Mocheni et Ladin dans les Alpes italiennes au Tyrol Sud et Trentino (Stenico et al. 1996), pour lesquels l'effet fondateur a été suggéré. L'effet fondateur a aussi joué un rôle dans la fondation de populations beaucoup plus grandes occupant aujourd'hui de larges aires géographiques, telles que le peuple Rom (Mendizabal et al. 2012; Moorjani et al. 2013) et la population finlandaise qui figure certainement parmi les populations fondatrices les plus étudiées. L'effet fondateur finlandais est notamment suggéré par l'occurrence de nombreuses maladies mendéliennes (Peltonen 1997; Peltonen et al. 1999; Kere 2001; Norio 2003a; Norio 2003b), dont la distribution géographique inégale entre les régions ou sous-populations

indique une structure populationnelle régionale, confirmée par des études génomiques (Jakkula et al. 2008; Sabatti et al. 2009). Par ailleurs, bien que les populations de l'Asie et de l'Afrique soient moins souvent étudiées, certaines ethnies de l'Inde formeraient aussi des populations fondatrices récentes, fondées il y a plus ou moins 30 générations et caractérisées par de fort taux d'endogamie (Reich et al. 2009).

De nombreuses populations insulaires ont aussi été formées au cours des deux derniers millénaires, notamment les Islandais (Williams 1993; Helgason et al. 2000), les Tasmaniens (Stankovich et al. 2005), les Aïnous et Ryukyu dans l'Archipel de la mer du Japon (Japanese Archipelago Human Population Genetics et al. 2012), les habitants de l'Île de Kosrae (Bonnen et al. 2006), de l'Île de Tristan da Cunha (Roberts 1968) et ceux de l'Île de Robinson Crusoë (Villanueva et al. 2010). Pour ces populations, l'effet fondateur est non seulement supposé *de facto*, mais aussi appuyé par la fréquence élevée de certains désordres mendéliens spécifiques. Les peuplements fondateurs insulaires sont parfois issus d'une migration fondatrice plus réduite que les populations fondatrices continentales, leur isolement géographique peut limiter les migrations subséquentes et leur expansion est limitée par une aire géographique réduite. Ces facteurs démographiques laissent supposer un grand potentiel de différenciation génétique des populations fondatrices insulaires.

Les diasporas de certains isolats religieux auraient causé une succession d'effets fondateurs notamment chez le peuple juif, dont les différents groupes se sont dispersés à travers le monde depuis 2 000 ans (Ostrer 2001; Ostrer and Skorecki 2013) et chez les Anabaptistes, Huttérites, Ménnonites et Amish (Puffenberger 2003; Boycott et al. 2008; Orton et al. 2008; Strauss and Puffenberger 2009), appelés « Plain People », qui, après avoir été relocalisés de nombreuses fois en Europe, ont établi des peuplements permanents en Amérique. Depuis les années soixante, de nombreuses maladies mendéliennes ayant une prévalence élevée ont été décrites et définissent,

chez ces populations, de nombreux effets fondateurs « cliniques » (McKusick et al. 1964; Motulsky 1995; Ostrer 2001; Risch et al. 2003; Strauss and Puffenberger 2009). Le peuple juif et le peuple des Plaines ont grandement contribué à la cartographie et à l'identification des mutations causales de maladies mendéliennes, grâce entre autres, à l'accessibilité des données médicales, génétiques et généalogiques (Arcos-Burgos and Muenke 2002; Strauss and Puffenberger 2009; Chong et al. 2012).

Au cours des derniers 500 ans, les grands mouvements de colonisation européens sur les autres continents ont remarquablement changé la distribution des humains sur la planète et ont donné naissance à de nombreuses populations. Des peuplements européens ont eu lieu en Afrique du Sud, au Nouveau-Monde et en Océanie. L'effet fondateur a été évoqué chez les Afrikaners d'Afrique du Sud, notamment pour expliquer l'incidence élevée de certaines maladies génétiques et par l'étendue du déséquilibre de liaison sur de plus longues distances (Hall et al. 2002; Abecasis et al. 2004; Greeff 2007). Les Garifunas forment une population fondatrice récente des Caraïbes, qui serait issue de naufrages et d'évasions durant la traite transatlantique des esclaves (Salas et al. 2005). Dans la vallée centrale du Costa-Rica et dans la région d'Antioquia en Colombie, l'effet fondateur a été soutenu pour la population d'origine hispanique, bien que sa signature soit brouillée par le métissage avec les peuples autochtones (Service et al. 2001; Carvajal-Carmona et al. 2003; Morera and Barrantes 2004). À Terre-Neuve, l'effet fondateur est appuyé par la prévalence accrue de maladies génétiques dans la population majoritairement d'origine britannique et irlandaise (Rahman et al. 2003). La population canadienne-française du Québec figure parmi les populations fondatrices les plus étudiées et les plus grandes, avec son effectif actuel de quelques six millions de personnes. L'histoire démographique et génétique de la population du Québec sera présentée en détail dans la seconde partie de ce chapitre.

II. L'HISTOIRE DÉMOGRAPHIQUE ET GÉNÉTIQUE DES CANADIENS FRANÇAIS DU QUÉBEC

Le Québec est l'un des premiers territoires d'Amérique du Nord colonisés par les Européens. L'histoire du peuplement du Québec est documentée par de nombreuses sources archéologiques, démographiques, généalogiques et historiques. Les Canadiens français du Québec forment une population fondatrice récente dont les origines remontent aux fondateurs de la Nouvelle-France. L'effet fondateur canadien-français a fait couler beaucoup d'encre, notamment dans le contexte des études des maladies mendéliennes. Dans cette deuxième partie de l'introduction, je présente l'histoire du peuplement du Québec, en mettant l'emphase sur la majorité canadienne-française et sur les sources documentaires principales permettant de reconstruire leur arbre généalogique. De plus, je dresse une synthèse des connaissances du patrimoine génétique des Canadiens français du Québec, incluant leurs maladies héréditaires, leur diversité et leur structure génétique.

1. Histoire du peuplement du Québec

La période précoloniale

Durant les millénaires précédant le peuplement européen, plusieurs nations autochtones se partageaient le territoire du Québec. Suivant la fonte du glacier laurentidien qui recouvrait entièrement le Québec, les premières bandes autochtones seraient venues au sud de l'estuaire du Saint-Laurent et sur les rives de la mer de Champlain qui occupait les basses terres de la Vallée du Saint-Laurent (Courville 1996). Les études archéologiques ont révélé des preuves d'occupation autochtone dans le Sud du Québec au Lac Mégantic entre 12 500 et 12 200 ans avant aujourd'hui, vers la fin de la dernière période glaciaire (Chapdelaine 2007). Ils auraient ensuite essaimé vers le Bouclier et les Appalaches (Courville 1996). Le peuplement du nord

de l'actuel territoire du Québec serait plus tardif. Les peuples Paleo-Esquimo serait venus il y a environ 5 000 ans, aux abords de l'océan Arctique et de la Baie d'Hudson, en provenance de l'Ouest de l'Arctique (Courville 1996; Helgason et al. 2006).

Les plus anciennes traces archéologiques des Européens au Nord-Est de l'Amérique sont retrouvées à l'Anse aux Meadows, à l'extrémité nord de l'Île de Terre-Neuve sur les berges du détroit de Belle Isle, qui a été occupé autour de l'an mil par les Norois, ou Vikings, en provenance de l'Islande et de la Scandinavie (Wallace Linderoth 1990). Plusieurs sagas, dont celle d'Erik le Rouge, stipulent que les Norois auraient occupé une région appelée le *Vinland* (McGhee 1984; Wallace Linderoth 1990). Bien que sa localisation géographique demeure incertaine, le *Vinland* serait situé quelque part au sud de Terre-Neuve, sur les côtes du Golfe du Saint-Laurent au Québec, au Nouveau-Brunswick ou en Nouvelle-Écosse (Kolodny 2012). La présence estivale de pêcheurs d'origine basque est aussi attestée dès le 16^{e} siècle sur les rives du golfe du Saint-Laurent et coïnciderait avec la présence des nations autochtones Mi'kmaq dans cette région (Loewen and Delmas 2012). N'ayant pas formé de peuplement permanent, ces premiers européens venus au Nouveau Monde durant la période précoloniale n'aurait pas légué d'héritage génétique chez les Canadiens français d'aujourd'hui, sauf peut-être indirectement par les autochtones avec lesquels ils se sont possiblement métissés, ce qui reste toutefois à évaluer[12]. En 1534, sous la gouverne de la France, Jacques Cartier et ses troupes ont exploré le golfe du Saint-Laurent

[12] Le génome des nations autochtones de l'Est du Québec n'a pas été étudié à ce jour. Chez les Inuits du Groenland, un important métissage scandinave a été montré dans les lignées paternelles du chromosome Y (Bosch E, Calafell F, Rosser ZH, Norby S, Lynnerup N, Hurles ME, Jobling MA. 2003. High level of male-biased Scandinavian admixture in Greenlandic Inuit shown by Y-chromosomal analysis. *Human genetics* **112**(4): 353-363.), alors que dans les lignées maternelles tracées par l'ADN mitochondrial, une absence complète de métissage européen a été rapportée (Saillard J, Forster P, Lynnerup N, Bandelt HJ, Norby S. 2000. mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion. *Am J Hum Genet* **67**(3): 718-726. ; Helgason A, Palsson G, Pedersen HS, Angulalik E, Gunnarsdottir ED, Yngvadottir B, Stefansson K. 2006. mtDNA variation in Inuit populations of Greenland and Canada: migration history and population structure. *Am J Phys Anthropol* **130**(1): 123-134.).

et ont débarqué en Gaspésie. Au cours de ses deux voyages suivants, en 1535-36 et 1541-42, Jacques Cartier est entré au cœur du continent nord-américain par le fleuve Saint-Laurent, s'arrêtant notamment à Québec et sur l'île de Montréal. La France a tenté de fonder une colonie dès 1542, cette fois avec l'aide du sieur de Roberval, qui aurait été située à l'emplacement actuel de la ville de Cap-Rouge près de la ville de Québec. L'entreprise coloniale s'est soldée par un échec et dès l'année suivante Roberval et ses colons sont rentrés en France (Courville 1996; Courville 2000). Durant les quelques décennies suivantes, les navigations françaises vers le Québec sont stoppées, ce qui pourrait avoir été influencé par les conflits transatlantiques entre la France et l'Espagne (Loewen and Delmas 2012).

La colonisation du Nord-Est de l'Amérique

Le peuplement européen permanent du Nord-Est de l'Amérique a débuté à l'aube du 17^{e} siècle. Sur le territoire actuel du Québec, le premier comptoir de traite des fourrures est établi en 1599 à Tadoussac, sur la rive Nord de la rivière Saguenay à l'embouchure du fleuve Saint-Laurent (Courville 1996). En 1608, sous le règne d'Henri IV en France, Samuel de Champlain a fondé la ville de Québec et initié la colonisation. Les villes de Trois-Rivières et de Montréal sont fondées respectivement en 1634 et en 1642. À la veille de la Conquête par les Britanniques, en 1755, la colonie française alors appelée *Canada* est concentrée dans la vallée du Saint-Laurent (Courville 1996; Courville 2000). À cette époque, les trois quarts des territoires colonisés de l'Amérique du Nord sont détenus par la monarchie française et la Nouvelle-France s'étendait du golfe du Saint-Laurent jusqu'en Louisiane, incluant l'aire entourant les Grands Lacs et la vallée du Mississipi (Figure 3). Les autres colonies d'Amérique du Nord étaient partagées entre l'empire hispanique au Mexique et au sud des Etats-Unis et l'empire britannique qui détenait Terre-Neuve, les treize colonies aux Etats-Unis ainsi que l'Acadie (Figure 3). Une colonie française à l'origine, l'Acadie a été fondée en 1604 sur la côte Atlantique. De 1632 à 1650, elle a accueilli environ 50 familles, qui sont

considérées comme la souche fondatrice principale des Acadiens (Roy 1975; Houdaille 1980). Estimée à 440 habitants au premier recensement de 1671 et à près de 3 000 habitants au moment de la prise de possession britannique en 1713, la population acadienne a ensuite connu un fort accroissement naturel. En 1755, lorsque les Britanniques ont entrepris le Grand Dérangement, elle totalisait environ 13 000 habitants (Roy 1975; Houdaille 1980). Entre 1755 et 1785, les Acadiens furent déportés dans les colonies anglaises, en Grande-Bretagne et en France. Certains réussirent à échapper à la déportation en trouvant refuge en Nouvelle-France alors que d'autres se retrouvèrent en France (Dickinson 1994). L'Acadie constitue aujourd'hui une partie du territoire des provinces du Nouveau-Brunswick, de la Nouvelle-Écosse et de l'Île du Prince Édouard, créées après 1760 sous le Régime britannique (McInnis 2000). Durant la période de la Nouvelle-France, soit de 1610 à 1760, les treize colonies anglaises d'Amérique qui donnèrent naissance aux États-Unis ont reçu plus de 300 000 immigrants des Îles Britanniques (Gemery 2000) et plusieurs dizaines de milliers d'esclaves d'origine africaine (Walsh 2000). En 1760, les colonies anglaises d'Amérique comptent 1,3 millions d'habitants (Gemery 2000). Bien que les Américains d'origine britannique et les Afro-Américains des États-Unis soient issus d'un peuplement fondateur au même titre que les Canadiens français, leurs très grands effectifs de pionniers n'évoquent pas *a priori* un scénario d'effet fondateur.

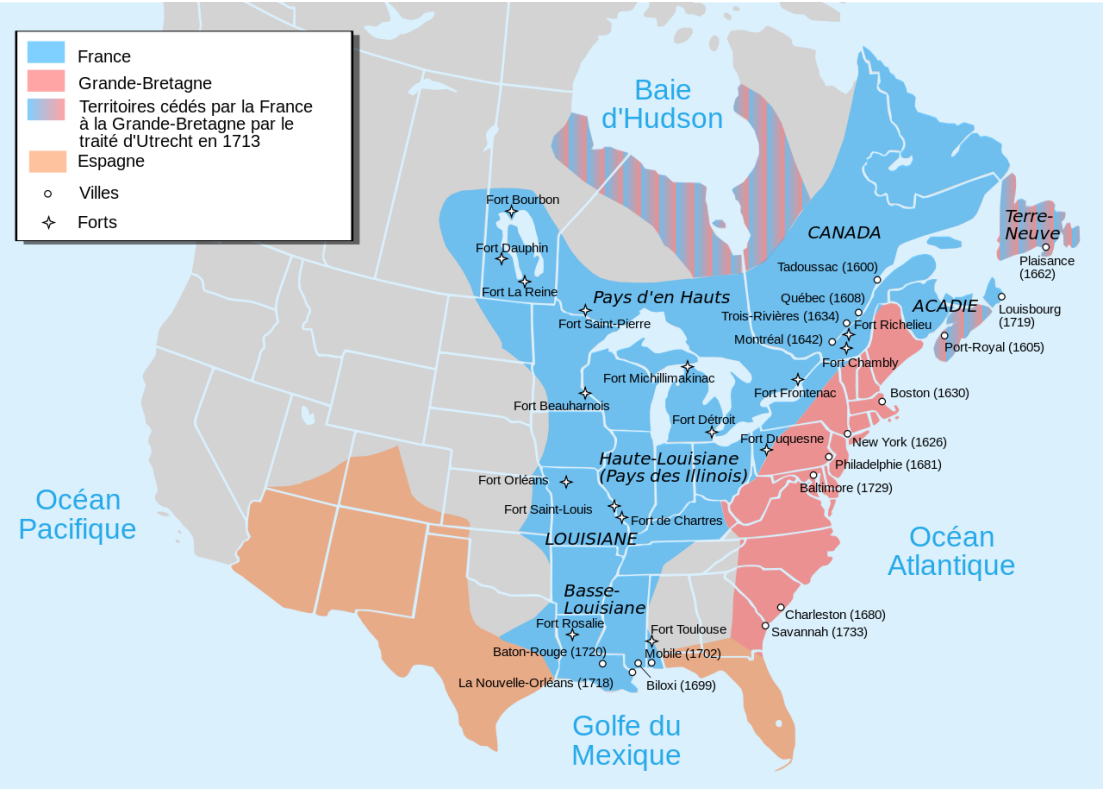

Figure 3. Carte de la Nouvelle-France vers 1750 et des autres colonies européennes.

En 1750, la Nouvelle-France (en bleu) comprend les colonies appelées Canada et Louisiane. En plus des treize colonies anglaises d'Amérique (en rouge), l'empire britannique possède les anciennes colonies françaises de Terre-Neuve et de l'Acadie (en bleu et rouge). L'empire hispanique (en orange) détient alors des territoires au sud des actuels États-Unis d'Amérique.

Source : fr.wikipedia.org/wiki/Colonisation_française_des_Amériques

La période de la Nouvelle-France

Les pionniers de la Nouvelle-France sont à l'origine de la population canadienne-française. Entre 1608 et 1760, sous le Régime Français, 8 570 pionniers ont immigré et ont eu au moins un enfant marié dans la colonie

(Brais et al. 2007). Néanmoins, durant la même période, au moins trois fois plus d'immigrants ont embarqué sur les bateaux pour la Nouvelle-France (Boleda 1990). Parmi les 27 000 embarqués, 2 000 ont péri sur les bateaux, pour cause de maladie et autres périls de la mer (Charbonneau et al. 2000). Parmi les 25 000 immigrants, la moitié n'est restée que quelques années dans la colonie et est repartie pour la France, tandis que parmi les 12 500 restants, un tiers n'ont pas eu d'enfants mariés en Nouvelle-France (Boleda 1990; Desjardins 2008). À une époque où 20 à 25 millions d'habitants peuplent la France, 8570 fondateurs donnent naissance à la population canadienne-française (Charbonneau et al. 2000; Desjardins 2008). Les fondateurs étaient près de trois fois plus nombreux que les fondatrices et ceci s'explique, du moins en partie, par les besoins du commerce de fourrure et du travail de défrichage agricole (Charbonneau et al. 1987). L'âge moyen des pionniers du 17^{e} siècle démontre que ceux-ci était majoritairement de jeunes adultes, 25 ans chez les hommes et 22 ans chez les femmes (Charbonneau et al. 2000). Les pionniers sont arrivés en majorité seuls, plutôt qu'en famille (Charbonneau et al. 1987; Guillemette and Légaré 1989). La majorité des 8 570 fondateurs de la Nouvelle-France étaient originaires de la France (Desjardins 2008). Bien que l'immigration a eu lieu en continu tout au long de la période de la Nouvelle-France, deux vagues d'entrées ont été plus importantes. La première a eu lieu entre 1663 et 1673, lorsque le roi Louis XIV envoya les 450 soldats de Carignan et un contingent de 800 femmes à marier, nommées les *Filles du Roi*, pour contrer le déséquilibre des sexes dans le marché nuptial de la colonie (Charbonneau et al. 1987). La seconde a eu lieu au moment de la Guerre de sept ans entre la France et l'Angleterre qui a pris fin en 1760, alors qu'une partie des 4 500 militaires français s'est installée dans la colonie (Charbonneau et al. 2000) et que 2 000 à 4 000 Acadiens ont trouvé refuge au Québec suite aux campagnes de déportation britanniques (Dickinson 1994; Bergeron et al. 2008).

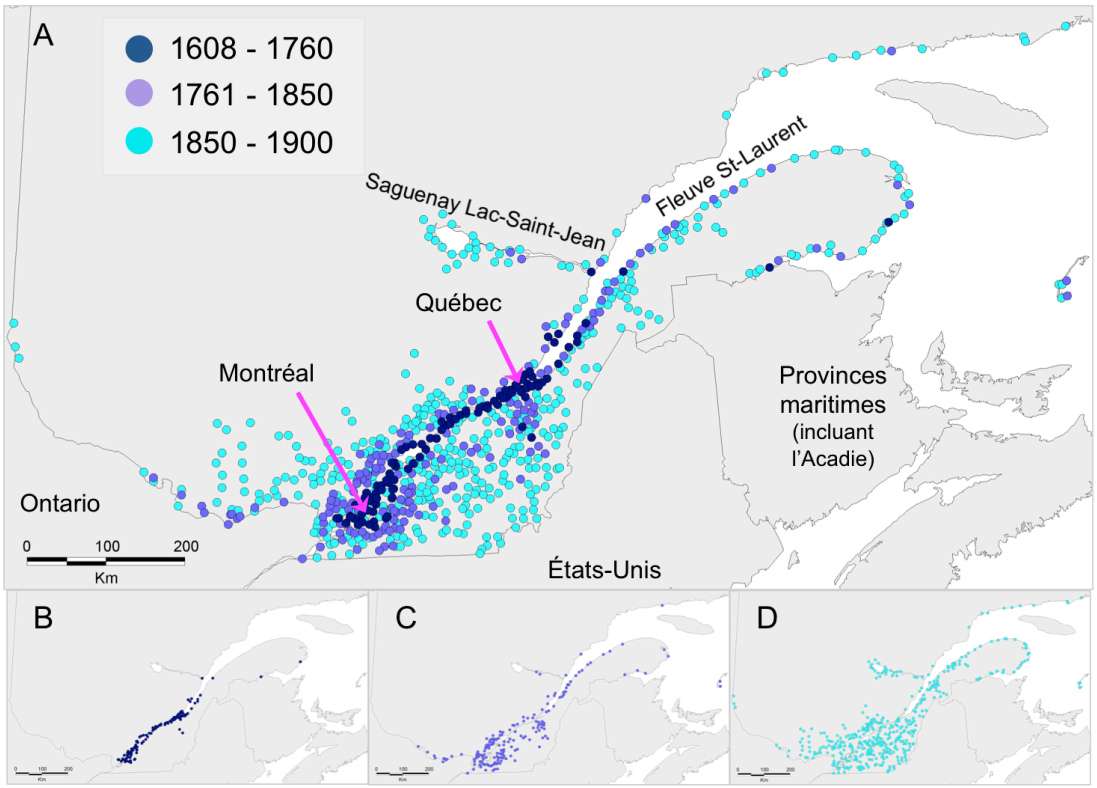

Figure 4. Progression de l'occupation du territoire du Québec.

(A) Carte de l'ouverture des localités à trois périodes. **(B)** 1608-1760. **(C)** 1760-1850. **(D)** 1850-1900. Cartes redessinée avec les données provenant de Brais et al. 2007.

La période du Régime britannique

La conquête britannique de la Nouvelle-France a eu lieu en 1760 et a été ratifiée par le Traité de Versailles en 1763. La population comptait alors 70 000 habitants. Celle-ci était répartie parmi un peu plus de cent paroisses concentrées dans la vallée du Saint-Laurent, principalement entre les villes de Québec et de Montréal (Figures 4A et 4B; Charbonneau et al. 2000; Brais et al. 2007). La colonie a d'abord été rebaptisée province de Québec puis Bas-Canada. Suite à l'établissement des Acadiens et des militaires français dans la colonie, l'immigration d'origine française a pratiquement cessée.

Entre la Conquête et 1850, à l'exception notable d'un contingent de plus ou moins un millier d'immigrants allemands établis au Québec (Wilhelmy 1984; Tremblay 2010), les immigrants outre Atlantique provenaient essentiellement des Îles Britanniques (McInnis 2000). Des milliers de Britanniques, d'Écossais et d'Irlandais ont immigré au Québec et se sont surtout installés dans les zones urbaines (McInnis 2000; Tremblay et al. 2009). Des Loyalistes à la Couronne britannique se sont aussi installés au Canada après l'indépendance des États-Unis (McInnis 2000). Bien que son ampleur soit peu connue, le métissage des Canadiens français avec ces nouveaux arrivants aurait été limité, principalement pour des considérations religieuses. La croissance de la population canadienne-française a donc été largement soutenue par un taux d'accroissement naturel élevé (Henripin and Péron 1973; Ouellet 1983; McInnis 2000; Brais et al. 2007).

Avec l'augmentation de la densité de la population dans la vallée du Saint-Laurent, la colonisation de nouvelles régions a été amorcée au début du 19^{e} siècle (McInnis 2000; Brais et al. 2007). Au nord-est de la ville de Québec, les régions du Saguenay-Lac-Saint-Jean et de la Côte-Nord sont ouvertes au peuplement (Figure 4C). Au centre sud de la province, dans les Appalaches, ce sont les régions des Cantons de l'Est et de la Beauce qui sont colonisées (Figure 4C). À l'ouest de Montréal, quelques paroisses sont fondées au sud de la région de l'Outaouais (Figure 4C). En 1851, la population du Québec atteint 900 000 habitants, incluant 700 000 Canadiens français et 200 000 Canadiens anglais formant deux sociétés distinctes à la fois du point de vue de la langue et de la religion (Brais et al. 2007).

Dans la seconde moitié du 19^{e} siècle, la forte croissance démographique des Canadiens français est accompagnée par trois mouvements migratoires principaux. Premièrement, une expansion territoriale de la population canadienne-française a eu lieu dans les régions éloignées des grands centres (Figure 4D), notamment stimulée par l'essor de l'industrie forestière (Courville 1996). Le front du peuplement s'est avancé dans les territoires plus

éloignés et de nouvelles régions ont alors été colonisées, notamment les Laurentides et l'Abitibi-Témiscamingue au nord et à l'ouest de Montréal (Figure 4D). Deuxièmement, l'industrialisation a entrainé un exode rural et la croissance des centres urbains (Courville 1996). Troisièmement, une émigration massive des Canadiens français à l'extérieur du Québec a pris place. On estime à 900 000 le nombre de personnes nées au Québec ayant émigré aux États-Unis de 1840 à 1930 (Lavoie 1972). L'émigration hors du Québec s'est d'abord dirigée vers les villes industrielles des États de la Nouvelle-Angleterre, pour se tourner ensuite vers les nouvelles aires de colonisation du Midwest Américain et de l'Ontario[13] (Brais et al. 2007). En 1900, la population canadienne-française du Québec se chiffrait à 1,35 millions sur un total de 1,6 millions d'habitants (Brais et al. 2007).

Le 20e siècle et la situation démographique actuelle

Au cours du 20e siècle, la diversité ethnique du Québec s'est grandement accrue, grâce à une immigration d'origines plus diverses. Avec la baisse de fécondité, la part de l'immigration dans l'accroissement total de la population est devenue de plus en plus importante depuis 1960 (Piché 2003; Piché 2005). Aujourd'hui, sur un total de 8,1 millions de Québécois, on estime à environ 6 millions le nombre de Canadiens français définis comme les descendants des pionniers de la Nouvelle-France. Ceux-ci représentent la grande majorité des 80% de francophones déclarés au recensement de 2011, alors que les Québécois de langue maternelle anglaise totalisent 8% de la population et les allophones, 12% (stat.gouv.qc.ca). Les anglophones et allophones résident à 82% dans la ville de Montréal qui comprend 3,1 millions d'habitants, de sorte que les autres régions du Québec sont fortement francophones (stat.gouv.qc.ca). La population québécoise est concentrée

[13] Les Canadiens Français auraient plusieurs millions de descendants à l'extérieur du Québec, ailleurs au Canada et aux États-Unis. Comme près de la moitié des Canadiens Français au 19e siècle ont émigré, on peut supposer qu'ils ont eu au moins six millions de descendants hors Québec, en assumant qu'ils ont eu le même nombre de descendants qu'à l'intérieur du Québec.

dans le sud du territoire, alors que le nord du Québec est occupé principalement par les communautés autochtones.

2. L'arbre généalogique du Québec

Les sources documentaires sur la population d'origine canadienne-française du Québec sont d'une qualité et d'une richesse remarquables. Les recensements nominatifs, dont certains ont été numérisés, sont disponibles depuis le premier relevé de 1666 (Charbonneau et al. 2000). Pour la reconstruction de l'arbre généalogique de la population, la source la plus importante provient de l'Église catholique, qui a tenu les registres des baptêmes, mariages et sépultures dans l'ensemble du territoire, et ce, depuis les débuts de la Nouvelle-France[14]. Ces actes ont été bien conservés jusqu'à ce jour et les pertes concernent surtout les premières décennies du 17e siècle, avant la résolution de 1678 de garder un dupliquât de tous les actes (Charbonneau et al. 2000). Aujourd'hui, les actes des paroisses catholiques du Québec sont disponibles au sein de registres de population informatisés incluant des initiatives commerciales, populaires et scientifiques de différentes envergures.

Le fichier de population BALSAC est le plus grand fichier de population informatisé du Québec et l'un des plus grands fichiers dans le monde (Glasson et al. 2008). Il couvre l'ensemble du territoire et contient à ce jour 3 millions d'actes informatisés, principalement des mariages, se rapportant à 5 millions d'individus. Le fichier BALSAC contient les actes de mariage pour

[14] Les prêtres catholiques avaient adopté la coutume d'inscrire dans ces actes une foule d'informations. Par exemple, dans un acte de mariage typique, en plus de la date, du lieu du mariage et du nom des époux, on retrouve entre autres l'âge (majeur ou mineur) et l'ethnicité, et surtout les noms des parents des conjoints, ce qui est nécessaire pour reconstruire les liens filiaux. Cette information est souvent manquante dans les actes protestants au Québec, ce qui rend plus difficile la reconstruction de leurs généalogies.

l'ensemble du Québec du début de la colonie jusqu'à 1965[15]. La saisie est en cours pour les actes de baptêmes et de sépultures. Pour la région du Saguenay-Lac-Saint-Jean, le fichier BALSAC comprend les actes de baptême, mariage et sépulture jusqu'à 1971. Les liens familiaux entre les individus ont été reconstitués grâce à une procédure de jumelage nominatif (balsac.uqac.ca; Vézina 2013). Grâce aux normes rigoureuses et aux multiples procédures de validation des données, le fichier de population BALSAC permet de reconstruire les généalogies des Canadiens français du Québec sur une période couvrant plus de 300 ans soit une dizaine de générations. Pour la région du Saguenay-Lac-Saint-Jean, c'est l'ensemble des familles qui est reconstituée à l'aide des actes de baptêmes, mariages et sépultures.

Les généalogies canadiennes-françaises offrent donc l'opportunité d'étudier les mécanismes d'évolution en jeu dans les populations fondatrices récentes. Mais pour innover sur la question de l'effet fondateur, le développement de nouvelles approches méthodologiques est requis (voir section « Évaluer l'effet fondateur dans les populations naturelles »). De plus, bien que de nombreuses études aient sondé de multiples échantillons, un portrait synthétique de ce grand arbre généalogique canadien-français est encore manquant. Néanmoins, on peut d'ores et déjà affirmer que des données généalogiques aussi exhaustives portant sur des populations de taille aussi importante sont retrouvées dans un nombre restreint d'autres populations, telles que chez les Anabaptistes des Plaines (Agarwala et al. 2003), les Islandais (Gulcher et al. 2001) et les Mormons de l'Utah (Neale et al. 2013).

[15] Les 69 000 actes de mariage antérieurs à 1800 proviennent du Registre du Québec Ancien du Programme de démographie historique de l'Université de Montréal (RPQA-PRDH).

3. Les maladies héréditaires des Canadiens français

Depuis 50 ans, les études cliniques et génétiques chez les Canadiens français mettent en lumière leur bagage particulier de maladies héréditaires. Dans cette section, je présente une vue d'ensemble des maladies mendéliennes caractéristiques des Canadiens français du Québec, incluant leur incidence et leur distribution géographique. Une synthèse des hypothèses proposées et explorées pour expliquer l'incidence accrue de ces maladies sera ensuite présentée, celles-ci incluent l'effet fondateur, la consanguinité, les mutations fondatrices et leurs origines.

Les maladies mendéliennes caractéristiques des Canadiens français

Chez les Canadiens français, on connaît au moins 30 maladies mendéliennes dont les mutations causales majeures sont identifiées et qui peuvent être qualifiées de caractéristiques à la population : elles ont soit une incidence plus élevée parmi les Canadiens français que chez d'autres populations, des phénotypes cliniques particuliers et/ou des mutations particulières. Cependant, il n'existe toujours pas de définition consensuelle, ni de registre, du bagage particulier des maladies héréditaires des Canadiens français (Méthot 2012), tel qu'il existe ailleurs, comme en Finlande (Norio 2003a). Les phénotypes cliniques des maladies mendéliennes des Canadiens français peuvent être retrouvés dans plusieurs revues de littérature (De Braekeleer 1991a; De Braekeleer 1991b; Scriver 2001; Laberge et al. 2005b; Dupré et al. 2006; Brais et al. 2007; Laberge 2007; Dupre et al. 2008; Noreau et al. 2013). Parmi les 25 maladies mendéliennes sélectionnées par Laberge et ses collaborateurs en 2005, 17 ont un mode de transmission autosomique récessif, cinq maladies sont autosomique dominante, deux sont liées au chromosome X et on retrouve une maladie mitochondriale. Une atteinte neurologique est présente chez une forte proportion de ces maladies, ce qui peut être expliqué par une forte activité médicale au Québec dans ce domaine, mais suggère par ailleurs un biais clinique en faveur de ces maladies dans nos connaissances du bagage des maladies canadiennes-

françaises. Depuis 2005, au moins huit nouvelles maladies mendéliennes ont été pour la première fois décrites chez plusieurs familles canadiennes-françaises et les mutations causales identifiées pour au moins cinq maladies auparavant caractérisées (Thiffault et al. 2006; Jarry et al. 2007; Meijer et al. 2007; Rossignol et al. 2007; Gosselin et al. 2008; Montpetit et al. 2008; Plante et al. 2008; Bernard et al. 2010; Srour et al. 2010; Tetreault et al. 2011; Levesque et al. 2012; Srour et al. 2012; Samuels et al. 2013). Ceci illustre bien que nos connaissances du bagage particulier des maladies canadiennes-françaises continuent de s'enrichir et, en contrepartie, suggère que nos connaissances de ce bagage demeurent encore partielles. En fait, les études cliniques qui visent à caractériser de nouveaux phénotypes mendéliens, ainsi que les études de génétique médicale qui cherchent à identifier leurs déterminants génétiques sont toujours des domaines de recherche actifs (Méthot 2012).

Bien que les maladies mendéliennes caractéristiques des Canadiens français soient plus communes au Québec, chacune d'entre elles touche un nombre relativement limité de familles, et sont donc plutôt rares, d'où leur étiquette récemment popularisée de maladies orphelines. Selon les estimations rapportées par Laberge et ses collaborateurs (2005), l'incidence connue pour quatre maladies prévalentes dans l'ensemble du Québec varie entre 1 : 260 et 1 : 77 284. Comme le nombre de naissances au Québec est aujourd'hui d'environ 88 000 par année (stat.gouv.qc.ca), on peut estimer grossièrement que le nombre d'enfants qui naîtront cette année avec l'une ou l'autre de ces quatre maladies varie entre 1 et 325 enfants.

Une explication de l'incidence accrue recherchée dans l'histoire généalogique

Depuis les premières observations de clusters de patients atteints d'un même phénotype médical, les chercheurs ont tenté d'expliquer cette incidence accrue par l'étude de l'histoire généalogique de ces patients (revu par Vézina 1996). L'incidence accrue d'une maladie mendélienne pourrait être le résultat de mariages entre proches consanguins, qui entrainent une augmentation de

l'homozygotie et peuvent ainsi favoriser l'expression de maladies récessives. Or, cette hypothèse a été réfutée pour toutes les maladies présentées ici où des liens de parenté entre conjoints entrainant une proche consanguinité ne sont pas rapportés pour la grande majorité des familles atteintes. Ceci concorde avec l'observation que depuis l'époque de la Nouvelle-France, les mariages entre proches apparentés (oncle-nièce/tante-neveu, cousins germains et petits-cousins) ont été très peu fréquents et donc largement évités chez les Canadiens français, vraisemblablement à cause de l'interdit religieux (Bouchard and De Braekeleer 1991b; Vézina et al. 2004). Les études généalogiques ont montré que la consanguinité de groupes de patients atteints de certaines maladies récessives est surtout de type éloignée, i.e. au delà de la 5^{ième} génération (revu par Vézina 1996). Comme le décrit bien Scriver : « the more likely explanation for certain disease prevalences and clustering in Quebec is founder effect and/or genetic drift » (Scriver 2001).

De nombreuses études généalogiques ont tenté de retracer le fondateur ou le couple de fondateurs ayant introduit une maladie mendélienne ou une mutation délétère dans la population (revu par Vézina 1996). Cet exercice s'est avéré dans certains cas non-concluant (ex. étude généalogique de la névrite héréditaire NHSA2 par Bhérer 2006) et certaines approches ont été remises en doute (Heyer and Tremblay 1995). Pourtant, les études ayant vraisemblablement réussi à identifier les fondateurs à l'origine de l'introduction de certaines maladies (ou mutations) sont des preuves très évocatrices des conséquences génétiques des peuplements fondateurs (ex. Laberge et al. 2005a; Vézina et al. 2005a).

Les mutations fondatrices et les effets fondateurs cliniques

À l'instar d'autres populations fondatrices, les Canadiens français ont contribué significativement à la découverte des mutations causales des maladies mendéliennes rares. Les mutations causales majeures sont connues pour au moins 30 maladies mendéliennes caractéristiques des

Canadiens français (revu par Scriver 2001; Laberge et al. 2005b; Dupré et al. 2006; Laberge 2007; Dupre et al. 2008; Noreau et al. 2013). La recherche des bases génétiques des maladies mendéliennes plus fréquentes au sein des populations fondatrices s'appuie typiquement sur l'hypothèse que les individus atteints d'un même phénotype mendélien partagent une mutation identique sur un haplotype relativement long. L'existence d'une telle mutation fondatrice facilite la cartographie et l'identification par des méthodes « identity-by-descent » et de partage d'haplotypes (Peltonen 1997; Sheffield et al. 1998; Arcos-Burgos and Muenke 2002; Chong et al. 2012). À cause de l'effet fondateur, une mutation délétère initialement rare peut effectivement augmenter en fréquence et devenir reconnaissable à cause de son impact observable sur le phénotype. Dans les études de génétique médicale, un « effet fondateur clinique » sera donc souvent décrit comme un cluster de patients non-apparentés atteints d'une maladie mendélienne causée par une mutation fondatrice qui est ailleurs relativement rare. Bien que nécessaire, cette observation n'est cependant pas suffisante pour confirmer l'effet fondateur. Une mutation délétère peut, par exemple, apparaître dans une petite population isolée et augmenter en fréquence par dérive génétique sans que cette augmentation soit associée à un phénomène de peuplement.

Au Québec, quelques études génétiques ont spécifiquement testé l'hypothèse de l'effet fondateur par l'analyse des haplotypes entourant les mutations causales. Par exemple, les études du rachitisme vitamino-dépendant et de la dystrophie myotonique de type I ont démontré que l'âge des haplotypes correspond à la chronologie du peuplement du Québec et/ou des régions de Charlevoix et du Saguenay-Lac-Saint-Jean, appuyant ainsi l'hypothèse de l'effet fondateur (Labuda et al. 1996; Labuda et al. 1997; Yotova et al. 2005).

Des mutations fondatrices rares, mais plus fréquentes chez les Canadiens français, ont aussi été identifiées pour des maladies communes, notamment dans le cas de formes familiales de cancer du sein associées à BRCA1 et

BRCA2 (Chappuis et al. 2001; Tonin et al. 2001; Vézina et al. 2005a). Par contre, dans le cas des polymorphismes génétiques plus communs associés aux maladies communes, les études n'ont pas rapporté de différences notables entre les Canadiens français et d'autres populations européennes.

Par ailleurs, on observe une remarquable hétérogénéité des mutations causales pour les maladies dites caractéristiques des Canadiens français. Il n'existe virtuellement aucune maladie mendélienne ayant une fréquence accrue chez les Canadiens français qui soit causée par une seule mutation. Une ou deux mutations fondatrices majeures et souvent plusieurs mutations mineures ont été identifiées, ce qui suggère de multiples introductions parmi la population (Scriver 2001; Laberge et al. 2005b; Yotova et al. 2005). En contraste, une seule mutation fondatrice explique 90% des maladies mendéliennes dont les gènes causals sont connus chez les communautés Anabaptistes des Plaines - Amish, Huttérites et Ménnonites – (Strauss and Puffenberger 2009). Ceci suggère une certaine hétérogénéité génétique des Canadiens français, même au niveau des allèles rares, et ne correspond pas à un scénario d'effet fondateur sévère suivi d'une isolation de la population. De plus, comme l'a exposé Scriver (2001), le spectre des mutations causales de maladies mendéliennes plus communes, telles que la phénylcétonurie, l'hypercholestérolémie familiale ou la déficience en lipoprotéine lipase, va aussi à l'encontre d'un isolement génétique de la population (Moreau et al. 2007).

La distribution géographique des maladies mendéliennes caractéristiques

Même si les maladies mendéliennes caractéristiques des Canadiens français peuvent être retrouvées dans l'ensemble du territoire, la majorité d'entre elles sont plus fréquentes ou même concentrées dans certaines régions, sous-régions ou localités, suggérant ainsi des effets fondateurs régionaux ou locaux. Par exemple, la tyrosinémie héréditaire de type I, une maladie métabolique qui est très rare ailleurs dans le monde (à l'exception notable d'une région de la Scandinavie (Laberge et al. 2005b)), a une incidence de

1 : 16 667 naissances vivantes dans l'ensemble du Québec et de 1 : 1 851 dans la région du Saguenay-Lac-Saint-Jean[16] (Desy et al. 2012).

La région du Saguenay-Lac-Saint-Jean, au Nord-Est du Québec, est sans aucun doute celle qui a été la plus étudiée. On y retrouve au moins 12 maladies dont l'incidence est plus élevée qu'ailleurs au Québec (à l'exception de Charlevoix) ou dans le monde (Tableau 1; De Braekeleer 1991b; De Braekeleer 1991a; Laberge et al. 2005b; Brais et al. 2007; Moreau et al. 2007). La forte incidence de certaines maladies mendéliennes dans cette région se traduit par une fréquence de porteurs élevée, estimée entre 1/5 à 1/55 pour 10 maladies récessives principales et à 1/89 et 1/530 pour deux maladies dominantes (Tableau 1). Globalement, un individu sur sept originaire du Saguenay-Lac-Saint-Jean serait porteur d'au moins une des sept maladies étudiées par De Braekeleer (De Braekeleer 1991a) et un individu sur quatre serait porteur d'une des cinq plus fréquentes maladies récessives (www.coramh.org). Ce fort risque a conduit à l'établissement d'un projet pilote de dépistage génétique populationnel en 2010[17]. Certaines maladies semblent même spécifiques à cette population régionale, puisqu'elles n'ont été rapportées nulle part ailleurs dans le monde, comme par exemple le syndrome de Leigh de type Canadien français (aussi appelé acidose lactique congénitale ou déficit en cytochrome C oxydase) et ce malgré les 10 ans écoulés depuis l'identification des mutations causales (Mootha et al. 2003; Laberge et al. 2005b). À l'opposé, certaines maladies fréquentes ailleurs ont une incidence plus faible au Saguenay-Lac-Saint-Jean

[16] Exceptionnellement, l'incidence de la tyrosinémie de type I est connue de façon très précise puisqu'elle fait l'objet d'un dépistage néonatal depuis 1970. Les estimations de l'incidence des maladies présentées ici doivent être considérées sous toute réserve. La plupart ont été calculées en faisant le rapport du nombre de patients atteints d'une maladie donnée sur le nombre total de naissances dans leur région d'origine pour la période cernée entre leurs dates de naissance. La fréquence de porteurs est généralement calculée d'après ces estimations.

[17] Le projet pilote de dépistage génétique au Saguenay-Lac-Saint-Jean concerne quatre maladies récessives : l'ataxie récessive spastique de Charlevoix-Saguenay, la tyrosinémie héréditaire de type I, l'acidose lactique congénitale et la neuropathie sensitivomotrice héréditaire avec ou sans agénésie du corps calleux (genetique.santesaglac.com).

(Labuda 1996). Par exemple, seulement quatre patients atteints de l'ataxie de Friedreich ont été rapportés depuis les années 1980, incluant deux sœurs dont les parents étaient possiblement originaires d'une autre région (Bernard Brais et Claude Prévost, communications personnelles). Ainsi, certains médecins généticiens ont affirmé (probablement sur la base de leurs observations cliniques) que l'éventail de maladies mendéliennes n'est pas plus grand au Saguenay, mais que ces maladies sont plus fréquentes et surtout mieux connues (Méthot 2012).

Tableau 1. Maladies mendéliennes plus fréquentes au Saguenay-Lac-Saint-Jean et dont les mutations causales majeures sont connues.

Maladie	# OMIM	Mode de transmission	Taux de porteur*
Ataxie spastique autosomique récessive de Charlevoix Saguenay (ARSACS)	270550	AR	1/22
Cystinose	219800	AR	1/39
Dystrophie myotonique 1 (Dystrophie myotonique de Steinert)	160900	AD	1/530
Fibrose kystique (Mucoviscidose)	219700	AR	1/15
Hémochromatose héréditaire	235200	AR	1/5
Hypercholestérémie familiale	143890	AD	1/122
Hyperchylomicronémie familiale (Déficience en lipoprotéine lipase)	238600	AR	1/67
Neuropathie sensitivomotrice héréditaire avec ou sans agénésie du corps calleux (syndrome d'Andermann)	218000	AR	1/23
Rachitisme vitamine D-dépendant de type I	264700	AR	1/27
Syndrome de Leigh de type Canadien français (acidose lactique, déficience en cytochrome C oxydase)	220111	AR	1/21
Syndrome de Zellweger	214100	AR	1/55
Tyrosinémie héréditaire de type I	276700	AR	1/22

* Les taux de porteur sont tirés de la revue de littérature par Laberge et al. 2005b, à l'exception du syndrome de Zellweger estimé par Levesque et al. 2012.

Des concentrations de maladies héréditaires sont aussi retrouvées dans d'autres régions ou sous-régions du Québec. Notamment au sud-est du Québec, au Bas Saint-Laurent et en Gaspésie où la fréquence des porteurs atteint jusqu'à 1/28 dans certaines sous-régions (Dupre et al. 2008). Par exemple, la dystrophie musculaire oculopharyngée a une fréquence plus élevée dans les comtés de Montmagny et de L'Islet (De Braekeleer 1991b), la maladie de Tay-Sachs à Rimouski et ses environs (De Braekeleer et al. 1992) et la polyneuropathie Charcot-Marie-Tooth récessive CMT4C dans la Baie-des-Chaleurs (Gosselin et al. 2008). Les maladies mendéliennes caractéristiques des Canadiens français sont mieux documentées à l'est de la ville du Québec. Des clusters de patients atteints d'une même maladie mendélienne sont aussi retrouvés dans les comtés ceinturant la ville de Québec, tel que l'ataxie cérébelleuse autosomique récessive de type I en Beauce (Gros-Louis et al. 2007) et plus à l'ouest, tel que la névrite héréditaire et sensitive de type II (NHSA2) dans Lanaudière (Roddier et al. 2005).

La structure de la population reflétée par le biais clinique

La distribution géographique des maladies mendéliennes caractéristiques des Canadiens français est donc non uniforme sur le territoire du Québec. Les concentrations régionales ou locales de certaines maladies suggèrent des effets fondateurs cliniques aux échelles régionale et sous-régionale. Plutôt que d'évoquer un effet fondateur panquébécois, la répartition des maladies mendéliennes apparaît donc comme une mosaïque d'effets fondateurs, semblable à ce qui est connu en Finlande (Kere 2001; Moreau et al. 2007). Ceci suggère une stratification géographique de la population canadienne-française du Québec. En somme, comme le proposent Moreau et ses collaborateurs (Moreau et al. 2007) : « cette régionalisation de l'effet fondateur, vue sous l'angle des maladies héréditaires, permet donc de se questionner sur l'homogénéité de la constitution génétique des Québécois d'ascendance française. »

4. La diversité génétique et la structure de la population

Au Québec, tout comme pour toutes les populations du monde, une des motivations principales à cataloguer la diversité génétique et à caractériser ses patrons de variation - incluant, entre autres, le clustering des génomes individuels qui démontre la structure de la population - est d'optimiser les études en épidémiologie génétique. En effet, la connaissance des patrons de diversité génétique intra- et inter-populationnels est essentielle pour le design et pour l'analyse des études cherchant les bases génétiques de la santé. En particulier, les *a priori* d'homogénéité génétique peuvent être trompeurs pour les études ciblant les bases génétiques des maladies. Par exemple, les études de cartographie génétique des maladies mendéliennes peuvent être biaisées si l'on cherche une seule mutation fondatrice causale partagée entre tous les patients. De plus, dans les études d'association populationnelles, la structure d'une population peut causer des résultats faux positifs et masquer les vrais résultats (Cardon and Palmer 2003; Marchini et al. 2004).

Dans cette section, je présente un bilan des connaissances sur les patrons de diversité génétique des Canadiens français du Québec. Ce discours s'articule autour du débat quant à la diversité génétique des Canadiens français : est-ce que ceux-ci forment une population génétiquement homogène ou hétérogène? L'idée répandue de l'homogénéité génétique des Canadiens français du Québec et les principaux fondements de ce paradigme seront d'abord présentés. Ensuite, les éléments d'hétérogénéité de la population qui ont été observés dans les études généalogiques seront introduits, ainsi que les conclusions des quelques études génétiques qui ont explicitement estimé la diversité génétique et/ou analysé la structure de la population.

Le paradigme de l'homogénéité génétique des Canadiens français

Depuis l'avènement des études de génétique au Québec, il est commun d'entendre que les Canadiens français forment un peuple homogène

génétiquement. Les éléments principaux sur lesquels se fonde cette idée répandue sont résumés dans cette citation, qui est un exemple parmi tant d'autres :

« This population [the Quebec founder population (QFP)] descended in genetic isolation from several thousand founders who emigrated from France in the 17th century. The demographic history of the QFP, which is characterized by a population bottleneck, rapid population expansion, and little admixture, makes it a valuable resource for use in genetic studies. The population has been well characterized as having reduced genetic heterogeneity for Mendelian diseases. » (Raelson et al. 2007)

Parmi les éléments factuels qui suggèrent l'homogénéité génétique de la population, on retrouve premièrement des faits historiques concernant le contexte démographique du peuplement (pour une discussion détaillée voir Bouchard and De Braekeleer 1990; Bouchard and De Braekeleer 1991b). On relate notamment un nombre « relativement petit » de fondateurs principalement d'origine française, un arrêt quasi-total de l'immigration française après la Conquête et un accroissement naturel fort, pesant plus lourd que l'immigration dans la croissance de la population. Bien que ces éléments ne mesurent en rien la diversité du pool de fondateurs ni celui de la population contemporaine, ils illustrent efficacement que les Canadiens français forment une population fondatrice récente et peuvent donner une impression d'isolement génétique de la population. Ces impressions sont davantage renforcées en comparaison avec les migrations très importantes qui sont à l'origine de la population des États-Unis. Deuxièmement, l'idée de l'homogénéité des Canadiens français a été vraisemblablement suggérée par les nombreuses observations d'une prévalence accrue de certaines maladies mendéliennes chez certaines sous-populations (De Braekeleer 1990; Bouchard and De Braekeleer 1991b; Bouchard 2004; Moreau et al. 2007). La fréquence accrue d'une ou de quelques mutations causant une maladie mendélienne, démontre bien une plus grande homozygotie au locus causal.

De plus, bien que cette observation ne permette pas à elle seule d'inférer la perte de diversité à un locus indépendant, ni chez d'autres sous-populations, celle-ci suggère la possibilité d'un effet fondateur.

Ainsi, le sous-texte du paradigme de l'homogénéité implique l'effet fondateur et ses conséquences théoriques de perte de diversité. En fait, les populations fondatrices récentes, issues d'un nombre limité de fondateurs, sont souvent présumées être génétiquement plus homogènes que les grandes populations exogames, dû à la perte de diversité causée par l'effet fondateur. Cette supposition est légitime. Celle-ci découle des résultats classiques des modèles de réduction de taille sévère suivie de dizaines de générations d'isolement et de dérive génétique (c.f. section « Prédictions théoriques de l'effet fondateur »).

En somme, dans l'état actuel de nos connaissances, l'idée d'homogénéité génétique des Canadiens français se fonde davantage sur des suppositions et des interprétations que sur des estimations réelles de la diversité génétique. Cette idée s'insère, dans un contexte plus large, dans le paradigme de l'homogénéité sociale et culturelle de la population qui est beaucoup plus ancien, et qui serait véhiculé dans l'historiographie québécoise au moins depuis le début du 20^{e} siècle, notamment dans les œuvres de Lionel Groulx (Bouchard and De Braekeleer 1990). C'est pourquoi, depuis les années 1990, on parle du paradigme de l'homogénéité génétique (De Braekeleer 1990; Bouchard and De Braekeleer 1991b; Bouchard 2004; Moreau et al. 2007).

Des éléments d'hétérogénéité observés dans l'arbre généalogique de la population

Des études démographiques et généalogiques ont depuis longtemps abordé la question de la diversité génétique et de la structure de la population canadienne-française du Québec. Évidemment, celles-ci ne mesurent pas directement la diversité génétique, mais apportent des observations qui

peuvent être interprétées comme suggérant une plus grande homogénéité, ou tel que souligné ci-dessous, une plus grande hétérogénéité génétique.

La diversité des patronymes, qui représente dans une certaine mesure la diversité génétique transmise de père en fils par le chromosome Y[18], varie substantiellement entre les populations régionales du Québec, ce qui suggère une stratification des lignées paternelles (Bouchard et al. 1985; Bouchard et al. 1987; Bouchard and De Braekeleer 1991b; Gagnon 2001).

Des études de la consanguinité ont montré que le degré d'homozygotie des Canadiens français dû aux unions entre proches consanguins est demeuré faible depuis la Nouvelle-France, hormis pour quelques sous-populations locales potentiellement plus isolées (Laberge 1967; Molloy 1990; Bouchard and De Braekeleer 1991b; Mayer and Boisvert 1994; Gagnon et al. 1998). En effet, les estimations de l'homozygotie des Canadiens français due aux unions consanguines jusqu'au 3^{e} degré (oncle-nièce/tante-neveu, cousins germains et petits cousins), qui représentent 2% ou moins des mariages contractés (Freire-Maia 1968), sont tout à fait similaires à celles rapportées chez de grandes populations européennes telles que la France, l'Italie et la Suisse; beaucoup plus faibles que certaines populations d'Amérique Latine, telles que le Brésil et le Vénézuela; et un peu plus élevées que la population d'origine européenne des États-Unis (Laberge 1967; Freire-Maia 1968; Lebel 1983; Bouchard and De Braekeleer 1991b). De plus, dans toutes les régions du Québec, les coefficients de consanguinité moyens demeurent relativement faibles jusqu'à la 5^{e} génération (Vézina et al. 2004), ce qui représente généralement la limite supérieure de notre connaissance des liens généalogiques et peut être considéré comme la frontière entre la consanguinité « connue » ou proche, de la consanguinité éloignée (ou cryptique). Ces observations suggèrent que globalement, la perte de diversité

[18] La correspondance entre les patronymes et les haplotypes du chromosome Y doit néanmoins être spécifiquement analysée dans chaque population, puisqu'on s'attend à ce que certains haplotypes soient partagés par plusieurs patronymes.

due à la consanguinité proche est limitée chez les Canadiens français et l'endogamie n'est pas un facteur qui les différencie des populations européennes, telle que la France. En revanche, les coefficients moyens de consanguinité éloignée, qui incluent le partage des ancêtres depuis la fondation de la population, sont importants pour certaines populations régionales contemporaines : pour les régions du nord-est québécois (Charlevoix, Saguenay-Lac-Saint-Jean et Côte-Nord) et pour les Îles-de-la-Madeleine (Vézina et al. 2004). Ceci suggère d'une part qu'une homogénéité génétique accrue peut être attendue pour ces populations régionales, et d'autre part, que certains facteurs démographiques doivent avoir favorisé ce plus grand partage d'ancêtres.

Les coefficients moyens de consanguinité varient non seulement entre les individus, mais aussi entre les localités et les régions du Québec (Laberge 1967; Freire-Maia 1968; Vézina et al. 2004). De la même façon, les coefficients moyens d'apparentement régionaux (calculés entre toutes les paires d'individus d'une même région) varient significativement d'une région à l'autre (Gagnon et al. 1998; Vézina et al. 2004). Dans l'ensemble de la population, la majorité des liens d'apparentement remontent au delà de la 6^{ième} génération et l'apparentement éloigné, ou cryptique, est important, avec 98% des paires d'individus ayant au moins un ancêtre commun (Tremblay et al. 2008). Vézina et ses collaborateurs (2004) ont fait ressortir un gradient ouest-est de diversité basé sur les coefficients de consanguinité et d'apparentement. En contraste, la contribution génétique des fondateurs aux populations régionales du Québec ancien (18^{e} siècle), a plutôt montré une structure tripartite organisée autour des trois ports d'entrée des immigrants à l'époque de la Nouvelle-France, soit les villes de Montréal, Trois-Rivières et Québec (Gagnon and Heyer 2001). En somme, ces observations suggèrent que la population canadienne-française du Québec est structurée à l'échelle macro-régionale. Les portraits divergents de la structure qu'elles proposent soulèvent néanmoins le besoin de nouvelles études ou du moins, de nouvelles propositions de consensus.

Les études génétiques ciblant spécifiquement la diversité des Canadiens français

Au Québec, les études de la diversité génétique des Canadiens français, basées sur des loci non liés aux maladies monogéniques, se sont d'abord appuyées sur les groupes sanguins et les protéines sériques (Magnan and Benoist 1969; De Braekeleer 1990) et plus récemment sur l'analyse de marqueurs génétiques neutres (Moreau et al. 2007; Moreau et al. 2009). Ces études ne sont pas directement comparables parce qu'elles diffèrent à la fois par leurs stratégies d'échantillonnage (différentes régions; incluant des familles ou non) et les loci génétiques étudiés. Or, elles sont arrivées à des conclusions similaires. Premièrement, contrairement à ce qu'il est communément supposé, dans son ensemble, la population du Québec n'apparaît pas moins diversifiée génétiquement qu'une population européenne. Ainsi, De Braekeleer a fait ressortir la similarité des fréquences haplotypiques des marqueurs sériques HLA-A et HLA-B, entre les échantillons tirés de trois populations régionales du Québec (Est du Québec, Montérégie et Saguenay) et de la France (De Braekeleer 1990). De plus, il a montré une diversité haplotypique accrue des Canadiens français comparativement aux isolats Huttérites et Touareg Kel Kummer (De Braekeleer 1990). La portée de cette étude est cependant limitée parce que de nombreuses études suggèrent qu'une variabilité génétique élevée est maintenue aux loci HLA par la sélection balancée (e.g. Prugnolle et al. 2005b). Plus récemment, Moreau et ses collaborateurs ont analysé la diversité génétique des lignées paternelles (haplotype composé de microsatellites du chromosome Y) et des lignées maternelles (portion hypervariable 1 de l'ADN mitochondrial) et de deux marqueurs génétiques neutres du chromosome X (DXS1238 et dys-44) (Moreau et al. 2007). Ils ont démontré que la variabilité génétique mesurée pour l'ensemble des échantillons du Québec (incluant sept populations régionales et ethniques) est tout à fait comparable aux niveaux observés en France (Moreau et al. 2007). Ces études remettent en question l'homogénéité génétique des

Canadiens français du Québec et suggèrent plutôt que leur histoire démographique (incluant le peuplement fondateur) n'a pas causé, pour l'ensemble de la population, une réduction de la variabilité génétique. Deuxièmement, les études de la diversité génétique des Canadiens français ont fait ressortir certaines différences entre les populations régionales du Québec. En particulier, la région du Saguenay-Lac-Saint-Jean se distingue au niveau des fréquences alléliques par rapport à la France pour les groupes sanguins ABO (Magnan and Benoist 1969) et pour la fréquence de certains allèles HLA-A et HLA-B (De Braekeleer 1990). Cette région montre aussi une différenciation génétique significative en comparaison avec d'autres populations régionales, pour les quatre systèmes génétiques étudiés par Moreau et ses collaborateurs (Moreau et al. 2007). Ces derniers ont aussi montré que l'échantillon de Gaspésiens qui se sont déclarés d'origine acadienne se différencie significativement pour les lignées maternelles et paternelles, alors que les Gaspésiens d'origine Loyalistes se différencient significativement seulement pour les lignées paternelles (Moreau et al. 2007; Moreau et al. 2009). Ces observations suggèrent que la population du Québec ne forme pas un ensemble homogène, mais serait plutôt structurée génétiquement, à l'échelle régionale et même sous-régionale. Les différences observées d'un locus à l'autre confirment que différents loci seront différemment affectés par l'histoire démographique. Ceci est bien illustré par la distinction entre les lignées de transmission empruntées par le chromosome Y versus l'ADN mitochondrial. Une image globale et intégrative de la diversité génétique des Canadiens français reste à définir. Celle-ci requiert non seulement l'analyse d'un maximum de populations régionales, mais aussi d'un plus grand nombre de loci neutres couvrant idéalement l'ensemble du génome.

Quelques questions ouvertes

En plus de démontrer le besoin de nouvelles études sur la diversité et la structure génomique du Québec, ces observations soulèvent des questions

importantes et non résolues au sujet de l'effet fondateur chez les Canadiens français du Québec. Notamment, comment se fait-il que les Canadiens Français, dans l'ensemble, sont aussi diversifiés qu'une population européenne? Bien que plusieurs aient proposé que le nombre de fondateurs ait été suffisamment grand pour limiter une perte de diversité (De Braekeleer 1990; Bouchard and De Braekeleer 1991b; Gagnon and Heyer 2001; Scriver 2001; Moreau et al. 2007), cette hypothèse n'a toujours pas été spécifiquement testée. De plus, l'apport des fondateurs d'origines autres que françaises peut certainement avoir contribué à enrichir la diversité génétique des Canadiens français, cependant l'ampleur du métissage est méconnue pour la plupart des régions du Québec. Par ailleurs, comment explique-t-on les concentrations régionales de certaines maladies génétiques rares? Cette régionalisation de la signature de l'effet fondateur pourrait être expliqué par les histoires démographiques des régions du Québec, tel que proposé par l'hypothèse de la régionalisation de l'effet fondateur (Gagnon and Heyer 2001; Scriver 2001; Gerbault 2006; Moreau et al. 2007). L'étude des généalogies profondes remontant jusqu'aux fondateurs de la population canadienne-française offre l'opportunité d'adresser ces questions en détail. Celle-ci promet d'apporter des éléments de réponse sur les modalités démographiques et les conséquences génétiques du peuplement fondateur initial de la Nouvelle-France et des histoires démographiques régionales des Canadiens français.

III. PROBLÉMATIQUE ET PLAN DE THÈSE

À la lecture des diverses études présentées dans ce chapitre, il apparaît clairement que l'effet fondateur a pu avoir un rôle important dans le façonnement du patrimoine génomique des populations humaines. Chaque population fondatrice est issue d'une histoire unique et son profil génétique doit donc être spécifiquement caractérisé. Cependant, de nouvelles approches méthodologiques sont requises pour évaluer, au sein des populations actuelles, la nature et l'étendue des conséquences génétiques des événements de fondation et ainsi mieux comprendre leur rôle et leur impact au cours de l'évolution humaine.

L'histoire du peuplement du Québec et de ses régions a sans équivoque façonné son patrimoine génétique, mais comment? Depuis les années 1960, de nombreuses études ont abordé cette question et on pourrait croire que celle-ci est résolue. Pourtant, le bilan des connaissances présenté en introduction nous laisse plutôt sur une controverse et surtout, sur de nombreuses questions qui demeurent ouvertes. Alors que plusieurs ont tôt fait de spéculer sur l'homogénéité génétique des Canadiens français et que cette idée s'est par la suite répandue au Québec et à l'international, des preuves généalogiques et génétiques suggèrent plutôt le contraire. Il peut donc sembler étonnant que si peu d'études aient réellement évalué, à l'aide de données expérimentales, la diversité et la structure génétique des Canadiens français. À ce sujet, nos connaissances sont encore plus partielles pour les populations régionales du Québec, notamment parce que le tableau brossé par les effets fondateurs cliniques est incomplet et surtout concentré dans les régions de l'Est du Québec.

L'objectif principal de cette thèse est de parvenir à une meilleure compréhension des patrons de diversité génétique de la population canadienne-française du Québec et de l'utiliser comme modèle afin de mieux comprendre les conséquences génétiques des événements fondateurs et de l'histoire démographique récente. Mes travaux examinent l'hypothèse que les histoires démographiques des populations régionales ont différemment affecté leur génome et dans certains cas engendré des effets fondateurs régionaux. J'adopte une approche méthodologique originale puisque j'utilise l'information contenue dans l'arbre généalogique reliant les membres contemporains aux fondateurs de la population. De plus, je m'appuie sur des résultats expérimentaux décrivant la diversité génomique d'un échantillon d'individus dont les généalogies sont également reconstruites.

Les chapitres II et III sont complémentaires. J'y étudie les patrons de diversité génétique de la population canadienne-française du Québec. Quelle est l'ampleur de la diversité génétique du Québec et de ses régions? Comment est-ce que le peuplement fondateur et les histoires démographiques régionales expliquent cette diversité? Au chapitre II, j'utilise un échantillon généalogique représentatif de l'ensemble du Québec, composé de 2 221 individus contemporains. Je caractérise l'immigration fondatrice du Québec et sa contribution aux populations régionales. J'estime l'ampleur du métissage et de la diversité génétique issue de la contribution des immigrants fondateurs. Au chapitre III, je présente une étude qui évalue empiriquement les patrons de diversité génomique de sept sous-populations du Québec et les situe par rapport à quatre échantillons de HapMap (CEU, YRI, CHB, JPT) et l'échantillon français du HGDP. Dans ces deux chapitres, je caractérise aussi la structure de la population canadienne-française du Québec, aux échelles régionale et subrégionale, à l'aide de données génomiques et généalogiques.

L'article présenté au chapitre IV a pour objectif d'évaluer les conséquences génétiques de l'histoire démographique des Canadiens français du Québec.

Comment les modalités de peuplement et les histoires démographiques régionales ont-elles affecté le spectre de fréquences alléliques, incluant les variants rares et communs? Est-ce que l'incidence relativement élevée de certaines maladies mendéliennes observé chez certaines populations régionales peut être expliquée par leur histoire démographique? Pour répondre à ces questions, je réalise des expériences de simulation conditionnelles à la structure généalogique de la population. À l'aide de l'échantillon généalogique de l'ensemble du Québec, j'évalue la nature et l'ampleur des changements de fréquences alléliques suivant le peuplement fondateur initial et l'expansion démographique régionale.

Au chapitre V, je présente une étude qui investigue les processus démographiques gouvernant les expansions territoriales humaines. Quels sont les paramètres de la reproduction de la population au front de la vague d'expansion? Quelles sont leurs conséquences génétiques? Cette étude caractérise la dynamique spatio-temporelle de l'expansion de Charlevoix et du Saguenay-Lac-Saint-Jean, de 1686 à 1960, à l'aide de généalogies descendantes comprenant plus d'un million d'individus.

Au chapitre VI, j'expose mes principales contributions et les limites de mes travaux. Je présente une synthèse du patrimoine génétique des Canadiens français du Québec. Je propose une résolution de la question de la diversité génétique et j'offre un nouveau portrait de la structure de la population. Enfin, je discute des retombées de nos résultats dans un contexte plus large, à la fois pour l'évolution des populations nouvelles et pour les études en épidémiologie génétique réalisées au Québec.

CHAPITRE II:

Admixed ancestry and stratification of Quebec regional populations

Claude Bhérer, Damian Labuda, Marie-Hélène Roy-Gagnon, Louis Houde, Marc Tremblay, Hélène Vézina

Référence:

Bhérer C, Labuda D, Roy-Gagnon MH, Houde L, Tremblay M, Vézina H. 2011. Admixed ancestry and stratification of Quebec regional populations. *Am J Phys Anthropol* **144**(3): 432-441.

CONTRIBUTION DES CO-AUTEURS

Pour cet article, ma contribution est la suivante:

- Design de l'étude avec HV et DL;
- Vérifications et validation des données généalogiques;
- Définition des regroupements régionaux avec HV, MT et DL;
- Développement des méthodes d'analyse de structure de population;
- Analyses statistiques;
- Rédaction du manuscrit.

La contribution des co-auteurs est la suivante: HV et MT ont fourni l'ensemble de données généalogiques tiré du fichier de population BALSAC. LH a supervisé le développement de la librairie de fonction GENLIB utilisée pour cette étude. MHRG a fourni un soutien méthodologique. HV et DL ont supervisé le design de l'étude. HV, DL et MT ont supervisé la définition des ensembles géographiques régionaux. HV et DL on supervisé et contribué à la rédaction du manuscrit. MHRG et MT ont révisé le manuscrit.

ACKNOWLEDGMENTS

We are grateful to Ève-Marie Lavoie from the Interdisciplinary Research Group on Demography and Genetic Epidemiology (Chicoutimi) and Michèle Jomphe from the BALSAC project (Chicoutimi) for technical assistance and Laurent Richard from the Historical Geography Laboratory at Laval University (Québec) for cartography work. We also thank Julie Hussin and two anonymous reviewers for their comments on an earlier version of the manuscript.

ABSTRACT

Population stratification results from unequal, nonrandom genetic contribution of ancestors and should be reflected in the underlying genealogies. In Quebec, the distribution of Mendelian diseases points to local founder effects suggesting stratification of the contemporary French Canadian gene pool. Here we characterize the population structure through the analysis of the genetic contribution of 7,798 immigrant founders identified in the genealogies of 2,221 subjects partitioned in eight regions. In all but one region, about 90% of gene pools were contributed by early French founders. In the eastern region where this contribution was 76%, we observed higher contributions of Acadians, British and American Loyalists. To detect population stratification from genealogical data, we propose an approach based on principal component analysis (PCA) of immigrant founders' genetic contributions. This analysis was compared to a multidimensional scaling of pairwise kinship coefficients. Both methods showed evidence of a distinct identity of the northeastern and eastern regions and stratification of the regional populations correlated with geographical location along the St-Lawrence River. In addition, we observed a West-East decreasing gradient of diversity. Analysis of PC-correlated founders illustrates the differential impact of early versus latter founders consistent with specific regional genetic patterns. These results highlight the importance of considering the geographic origin of samples in the design of genetic epidemiology studies conducted in Quebec. Moreover, our results demonstrate that the study of deep ascending genealogies can accurately reveal population structure.

INTRODUCTION

A population is stratified when relatedness is not uniform across subgroups of this population as a result of unequal, nonrandom genetic contribution of distinct ancestors. Differential contributions occur through demographic processes such as migration, founder effect, isolation and endogamy that lead to preferential mating and impact on the genetic structure of the population. Ancestors' genetic contributions can be traced in genealogical records in populations where data allowing for genealogical reconstructions are available. Otherwise, the ancestral connections of a population are typically inferred from its genetic diversity patterns. Population stratification is well recognized as a confounding factor in genetic association studies as it can lead to spurious associations (Cardon and Palmer 2003; Marchini et al. 2004). Although methods are now available to detect and correct for population stratification from genome-wide data, it is preferable for researchers to be aware of potential stratification even before designing their studies. Here, we propose a new approach to analyze population structure from extensive genealogical data, which relies on the differential genetic contribution of the founders and does not require genotype data.

European colonization of the province of Quebec began four centuries ago with the foundation of Quebec City in 1608. Over the span of 150 years of French rule, approximately 8,500 settlers, mostly of French origin, established themselves in "Nouvelle-France" (Charbonneau et al. 1993; Charbonneau et al. 2000). At the time of the British Conquest in 1760, the population, who numbered 70,000, inhabited mainly the shores of the Saint-Lawrence River and its principal tributary rivers. Following the Conquest, between 2000 and 4000 Acadians, descendants of French pioneers from Acadia (located in sectors of present-day Nova Scotia, New Brunswick and Prince-Edward Island), settled in Quebec after the British deportation campaign (Dickinson 1994; Bergeron et al. 2008). A group of American Loyalists also came to

Quebec after the war of Independence of the United States. In the last part of the 18th century until the end of the 19th century, the French Canadian population expanded rapidly, sustained essentially by a high fertility rate. Territories peripheral to initial settlement were colonized. During that period, immigrants came mainly from the British Isles (McInnis 2000). In the 20th century, the immigrants to Quebec came from much more diversified locations (Piché 2003). Today, the Quebec population numbers 7.8 million residents, of which 80% are French speaking, 8% are English speaking and 12% are allophone (www.stat.gouv.qc.ca). Eighty-two percent of the English speakers and allophones of Quebec reside in the metropolitan region of Montreal (www.statcan.gc.ca). The majority of French speakers can trace back their ancestry to the 8,500 pioneers of Nouvelle-France. Here, we focus on this portion of the population and refer to them as French Canadians.

The small number of founders – relative, for instance, to the 360,000 immigrants that left the British Isles to people the English colonies (Brais et al. 2007) - most likely contributed to the belief that the French Canadians from Quebec form a homogeneous population. However, in the past 20 years, genetic studies conducted in Quebec have demonstrated that the overall diversity of the French Canadian gene pool is not reduced compared to that of their parental European populations. For instance, mitochondrial and Y-chromosome gene diversity is nearly equal in samples from Quebec and France (Moreau et al. 2007). Genome-wide association studies for common disease did not notice substantial differences in the genetic heterogeneity of French Canadians compared to European populations. On the other hand, the patchy distribution of mutations underlying Mendelian diseases points to local founder effects (Scriver 2001; Laberge et al. 2005b), which suggests that the contemporary French Canadian population, rather than being a single randomly interbreeding entity, is stratified into genetically distinct subpopulations. In addition, genealogical studies of kinship and consanguinity in the contemporary population (Vézina et al. 2004) and of founders' genetic contribution to a cohort of couples married between 1780 and 1800 (Gagnon

and Heyer 2001) indicate some level of stratification of the Quebec gene pool.

In this study, we characterize the population structure and provide insights into the genetic diversity of the contemporary French Canadian population of Quebec using extensive genealogical data. We analyze the genetic contribution of 7,798 immigrant founders identified in the ascending genealogies of a sample of 2,221 subjects selected in all Quebec populations. We investigate the immigrant founders' characteristics and differential contribution to assess the level of diversity of the population and to address how demographic history has shaped this structure. Our study rests on the hypothesis that regional settlement histories that followed the initial founder effect in the 17^{th} century have led to some level of genetic differentiation across regional populations and that this phenomenon can be examined through the analysis of genealogical features of the population.

MATERIAL AND METHODS

Data

Sample

A sample of 2,221 subjects married in Quebec between 1945 and 1965 was drawn randomly from the BALSAC-RETRO genealogical file, which comprises 240,000 marriages of mostly Catholic confession (Bouchard and Vézina 2009). The sampling process was stratified according to the regional distribution of the population reported in the 1956 Canadian census. At the time, the Catholic population represented 88% of the 4,628,378 inhabitants of Quebec (Henripin and Péron 1973). The period of 1945-65 was chosen to maximize the sample size (based on data availability) while remaining as close as possible to the present. Married individuals are expected to have contributed to the contemporary population. Out of the 2,237 initially selected, the 2,221 individuals retained have genealogies extending back at least two generations and are unrelated at the 1^{st} and 2^{nd} degree (kinship coefficient < 0.125). They were partitioned into eight regional groupings, referred to as regions, using current geographical limits of administrative regions. The grouping was based on demographic and historical criteria in order to capture the progression of settlement (Fig. 1).

Genealogical reconstruction

Ascending genealogies of the 2,221 subjects were reconstructed using the BALSAC population database (www.uqac.ca/balsac/) and the Early Quebec Population register (www.genealogie.umontreal.ca). Complementary sources such as marriage repositories and family dictionaries were also consulted. Genealogies were reconstructed as far as the sources would allow, essentially up to the first European settlers entering the colony (mean generation depth = 9.3). Less than 0.1% of distinct ancestral links end

prematurely because of adoption or lack of information (mean generation depth = 7.9). The genealogies contain over 5 millions ancestral links connecting 153,447 distinct ancestors (see Supplementary Table 1 for descriptive statistics of the genealogies).

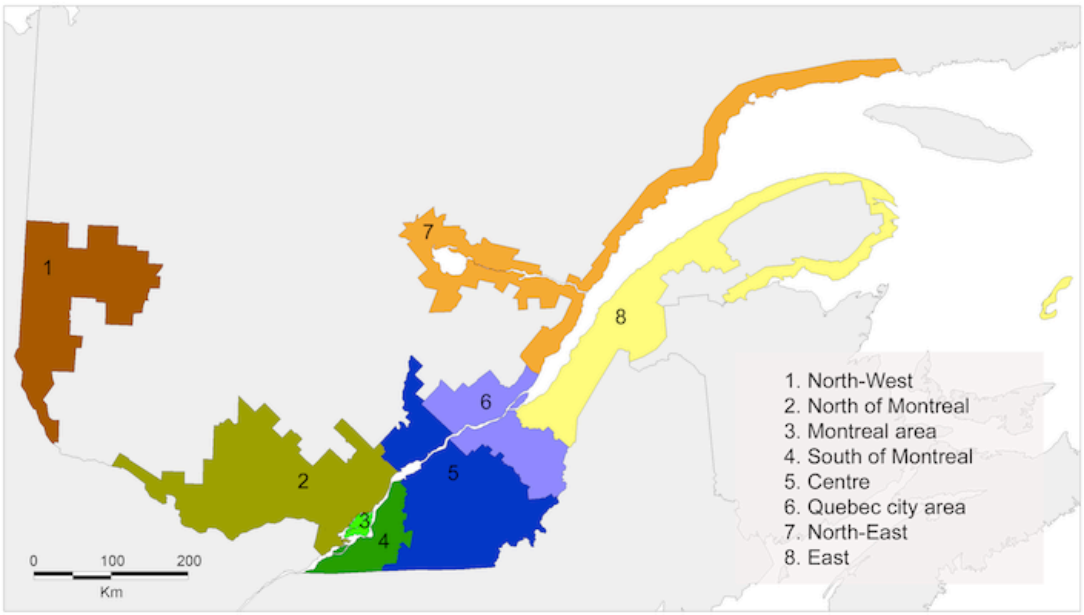

Figure 1. Quebec regional population samples.

The Quebec territory was partitioned in eight regional groupings based on geography and settlement history. The distribution of the 2,221 individuals sampled in these groupings is representative of the population repartition in 1956.

Analysis

Identification of immigrant founders

Genealogical founders are defined as individuals with no parental information available. In our genealogies, they can be the actual immigrant founders, their parents, or other native individuals with no parental information such as adoptees. The BALSAC-RETRO database records information on immigrant

status, thus allowing the identification of the individuals who were first to settle in Quebec among the ascending lineages. These individuals, defined as immigrant founders, were identified in 99.8% of the lineages. We included Amerindians among the immigrant founders as, from our standpoint, they introduced genetic diversity in the French Canadian gene pool. The remaining 0.2% of lineages was not considered in the present study. The geographic origin of immigrant founders was obtained either directly from their marriage records, from census data or indirectly by the place of origin/marriage of their parents. Place of origin was determined for 97% of all immigrant founders. We used the date of their first marriage in Quebec as a proxy of the time of arrival/settlement of immigrant founders. When this time could not be determined (in 647 instances, most likely because immigrants married before establishing in Quebec), we approximated the time of marriage by subtracting 30 years from the mean year of marriage of their children. Thirty years correspond to the average parent-children generation interval in the French Canadian population (Tremblay and Vézina 2000).

Definition of ancestors' layers

We defined an ancestors' layer as a group of ancestors present in our genealogies who married within a given period of 30 years. Genealogies of each regional sample were sliced into layers of ancestors married ± 15 years around the following pivotal years 1660, 1700, 1760, 1800, 1850 and 1900. When both parents and children were found in a layer, only the parents were retained.

Founders' genetic contribution

Genetic contribution to subjects

As parents transmit half of their autosomal genome to each child, the probability that any subject received an allele from any given founder can be calculated by summing transmission probabilities over all genealogical paths

connecting a founder to a subject (Roberts, 1968). We computed the genetic contribution of each founder to each subject ($GC_{f,s}$) as:

$$GC_{f,s} = \sum_{i=1}^{p} \left(\frac{1}{2}\right)^{g_i} \quad (1)$$

where *f* is one of the n_f founders, *s* is one of the n_s subjects, *p* is the number of genealogical paths between *f* and *s* and g_i is the number of generations separating *f* from *s* through a genealogical path *i*. Genetic contribution calculations were performed with the S-Plus®8 function library GenLib (www.uqac.ca/grig/).

Relative genetic contribution to groups of descendants

We calculated the relative genetic contribution of each founder to each regional sample (*rGC*) as:

$$rGC_f = \frac{\sum_{j=1}^{n_r} GC_{f,j}}{n_r} \quad (2)$$

where $GC_{f,j}$ is, as described in Eq. (1), the genetic contribution of a founder *f* to the j^{th} subject and n_r is the number of subjects in a regional sample. This measure describes the probability that a randomly chosen allele in a group of descendant comes from a given founder. It can be interpreted as the proportion of a group's gene pool expected to derive from a given founder. We also calculated *rGC* of each founder to each ancestors' layer of a given regional sample using Eq. (2), by replacing n_s by n_a, i.e. the number of ancestors in a given layer.

Homogeneity index

We calculated the "homogeneity index" (*HI*) for each regional sample and each ancestors' layer following Gagnon and Heyer (2001):

$$HI = \sum_{k=1}^{n_f} (rGC_k)^2 \qquad (3)$$

where n_f is the number of founders contributing to a given group of descendants and rGC_k is the relative genetic contribution of the k^{th} founder to that group (see Eq. 2). It represents the probability that two randomly chosen alleles in one group's gene pool come from the same immigrant founder. This statistic is directly related to the variance in founders' genetic contribution which determines the genetic diversity of the population.

Founders' uniform contribution number (FUN)

The FUN was calculated for each region as the reciprocal of the homogeneity index. This statistic corresponds to the number of equally contributing founders expected to produce the same genetic diversity as the actual founders in the population under study (Lacy 1989; Gagnon and Heyer 2001). If all founders contributed equally to a descendants' gene pool, the ratio of the FUN to the actual number of founders (n_f) equals to one. Unequal genetic contribution leads to smaller FUN/n_f ratio.

Population structure

To explore contemporary population structure observed through genealogical links, we used three different graphical methods. First, we applied a new approach that relies on a principal component analysis (PCA) of the genetic contribution of founders to subjects (GC-PCA) using Matlab. Specifically, we considered the matrix of genetic contribution of founders to subjects (see Eq. (1)) that has n_s rows and n_f columns. We retained the first two principal components (PCs) that explain the most variance as determined by the scree test (Supplementary Fig. S1). We tested for significant differentiation of the regional samples on the first two PCs through ANOVA using the R statistical package.

Second, we performed a PCA of founders' incidence as proposed by Calboli et al. (2008) (Calboli-PCA). Specifically, we used the founder to subject incidence matrix that has 1 in row *i* and column *j* if *i* is descendant of founder *j* and 0 otherwise. Using this matrix, we also applied a *K*-means clustering to identify *K*=2 clusters of subjects having a maximum number of common founders to replicate the analysis of (Calboli et al. 2008).

Third, multidimensional scaling analysis (MDS) of the pairwise kinship coefficient matrix was computed, using one minus kinship coefficients as a measure of distance. Kinship coefficients were calculated using Karigl recursive algorithm (Karigl 1981) as implemented in the GenLib function library (www.uqac.ca/grig/).

RESULTS

Time of arrival and origins of the founders

In the genealogies of the 2,221 subjects sampled, we identified a total of 7,798 immigrant founders (Table 1 and Fig. 2A). Among these founders, there were 2.6 times as many males as females (Fig. 2B), consistent with the skewed male-to-female ratio among the first settlers of Nouvelle-France (Charbonneau et al. 1993). Seventy-two percent of the immigrant founders settled during the French Regime (1608-1760) and 24% came in two major waves of immigration, the first between 1663 and 1673 and the second a hundred years later between 1755 and 1765 (Fig. 2B, Supplementary Fig. S2 and Table S2). The first wave corresponds to the arrival of French women, the so-called “Filles du Roy” who were sent from France to encourage stable family-based settlement in Nouvelle-France (Charbonneau et al. 1993). The second wave coincides with the British Conquest. It included Acadians escaping deportation by the British from their original settlements in Acadia as well as French soldiers who stayed in Quebec once the war ended (Charbonneau et al. 2000).

Sixty-eight percent of the immigrant founders came from France. French founders represent the vast majority of Europeans who settled before the British Conquest (Fig. 2C - Supplementary Table S3). Only 3% of the founders married before 1700 did not originate directly from France. The proportion of such founders increased to 35% in the period from 1700 to 1760. Acadians represent 14% of all the immigrant founders (Fig. 2C - Supplementary Table S3). The remaining founders of known origin came from Great-Britain (4%), Germany (2%), Ireland (3%), other European countries (1%) and other American locations (besides Acadia) (4%) (Supplementary Table S3). Amerindian origin was documented for one per cent of founders. Overall, the period of arrival and origins of the immigrant founders appearing in the genealogical ascendance of our contemporary sample well reflected

those of the pioneer immigrants that settled in Quebec prior to the British Conquest (Charbonneau et al. 2000).

Table 1. Distribution of subjects and founders per region.

	Region	Number of subjects	Number of founders
1	North-West	87	3,724
2	North of Montreal	242	4,343
3	Montreal area	722	6,317
4	South of Montreal	178	4,384
5	Centre	348	4,790
6	Quebec City area	272	3,676
7	North-East	157	2,628
8	East	215	3,188
	Whole Quebec	2,221	7,798

Partitioning of founders among regions

Not every founder contributed descendants to all regions of Quebec. One fifth of the founders (*n*=1,612; 20.7%) were common, that is, they contributed to all eight regions of Quebec (Fig. 2A). More than half of the founders (*n*=4,330; 55.5%) contributed to 2-7 regions and more than one fifth (*n*=1,856; 23.8%) were specific to only one region (Fig. 2A). All founders common to all eight regions married during the French rule and 97.2% of them before 1700. By contrast, 70% of the specific founders arrived after the installment of the British rule in 1760. We observed a negative correlation between founders' arrival time and the number of regions where they have descendants (Pearson's r = -0.6, *p-value* < 2.2×10^{-16}).

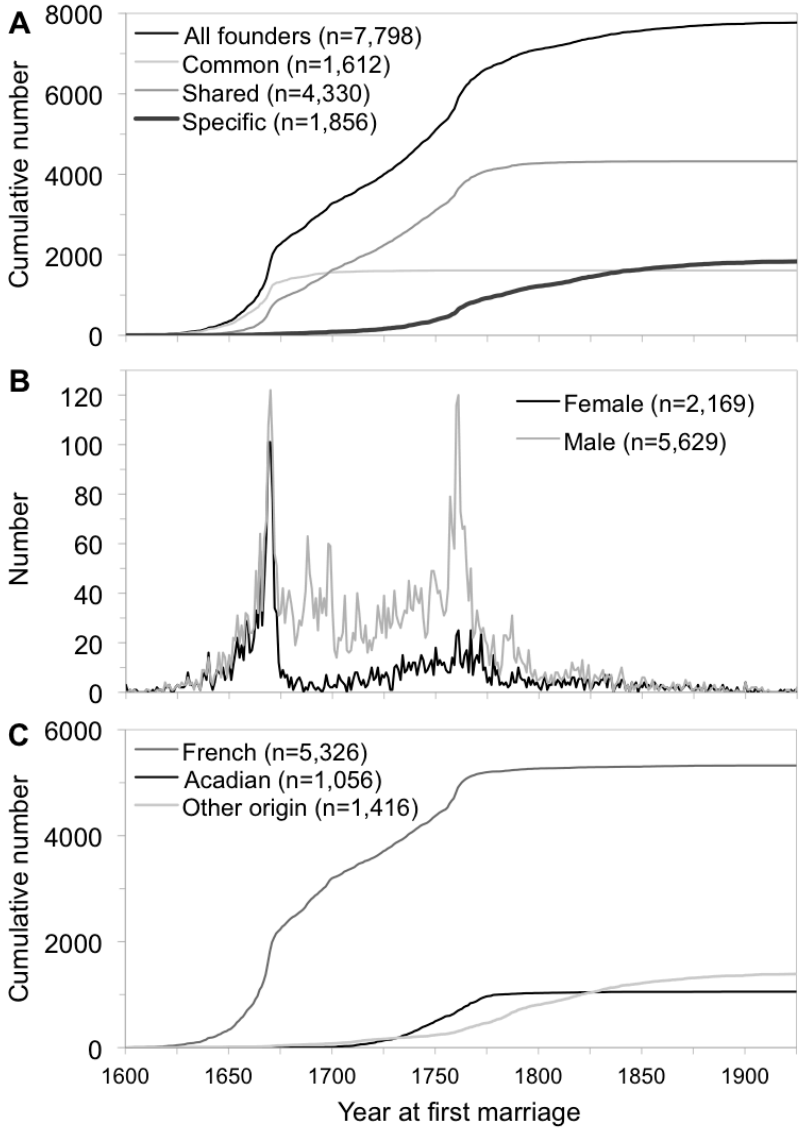

Figure 2. Time of arrival of the 7,798 immigrant founders.

We used the year at first marriage as an estimation of the time of arrival. **(A)** Cumulative distribution of the immigrant founders according to year at first marriage and regional representation. Specific founders contributed to one region (*n*=1,856), shared founders contributed to 2–7 regions (*n*=4,330), while common founders contributed to all eight regions (*n*=1,612). **(B)** Distribution of immigrant founders according to year at first marriage. **(C)** Cumulative distribution of the immigrant founders according to year at first marriage and origin.

Table 2. Genetic contribution (%) of the founders according to their origin.

		Origin							
	Region	French	British	German	Irish	Other European	Acadian	Amerindian	Other American
1	North-West	90.4	1.2	0.2	0.1	0.8	5.2	0.2	1.0
2	North of Montreal	90.0	1.4	0.6	1.4	1.0	2.6	0.2	2.2
3	Montreal area	87.5	2.0	0.4	0.8	1.5	4.1	0.3	2.2
4	South of Montreal	89.9	1.7	0.6	1.0	0.6	3.5	0.1	1.8
5	Centre	89.2	1.0	0.3	0.9	0.8	6.0	0.1	0.9
6	Quebec City area	93.8	1.2	0.2	0.4	1.0	2.3	0.2	0.6
7	North-East	90.3	2.3	0.4	0.4	1.0	3.4	0.3	1.3
8	East	76.4	3.4	0.5	1.8	0.9	11.1	0.3	3.5
	Whole Quebec	89.1	1.8	0.4	0.9	1.2	3.8	0.2	1.6
	% of total number of founders	68.3	4.1	1.8	2.7	1.2	13.5	1.2	3.8

Note : Immigrant founders of unknown origin (*n*=252) explained a minor proportion of the total (3.2%) and of the regional gene pools (0.2-2.3%).

Mosaic origins of Quebec regional populations

Nearly 90% of the regional gene pools were contributed by French founders (Table 2) and 2-6% by Acadians, who are second in numerical importance. Other groups of immigrant founders each contributed 2% or less. These proportions were similar across regions except for the East where the genetic contribution of French founders was reduced (76%) to the advantage of Acadian (11%), British (3.4%) and American founders (3.5%). Despite the elevated contribution of French founders, the subjects from our sample were nearly all admixed: on average, each genealogies contained immigrant founders from 6.75 distinct origins. While virtually all genealogies (99.1%) had at least one founder originating from France (Table 3), founders of other origins also appeared in a large proportion of the genealogies: British in 93% of genealogies, Acadian in 79% and notably, Amerindian in 47%.

Table 3. Proportion of the genealogies (%) in which appears at least one founder of a given origin.

		Origin							
	Region	French	British	German	Irish	Other European	Acadian	Amerindian	Other American
1	North-West	98.9	98.9	12.6	16.1	96.6	87.4	58.6	50.6
2	North of Montreal	99.6	91.7	31.8	41.3	96.3	63.2	48.8	75.6
3	Montreal area	98.8	90.4	18.8	28.4	92.7	73.8	44.9	69.1
4	South of Montreal	99.4	91.0	23.6	29.2	91.6	80.9	48.9	72.5
5	Centre	99.4	90.2	13.8	8.6	96.6	87.6	51.1	52.3
6	Quebec City area	100.0	97.1	12.5	7.0	98.5	75.7	61.0	29.8
7	North-East	99.4	99.4	15.3	7.0	97.5	79.0	35.7	62.4
8	East	98.1	93.0	11.6	19.5	91.6	94.4	31.2	64.2
	Whole Quebec	99.1	92.6	17.9	21.3	94.7	78.5	47.1	61.0
Number of founders		5,326	317	143	214	97	1,056	95	298

East-West gradient of diversity

Within each regional sample, founders did not contribute equally. Uneven contribution of the founders is shown in Figure 3 where the cumulative proportion of the genetic contribution of founders is plotted against the cumulative proportion of contributing founders. If all founders contributed equally to a gene pool, a linear dependence would be expected, as shown by a straight line in Figure 3. In contrast, the observed dependence was not linear: a small fraction of founders explains a large proportion of the gene pool while a greater fraction contributes less. For instance, 11% of the Montreal region founders explained 50% of its gene pool, while the remaining 89% founders explain the other half. The uneven contribution of founders was more pronounced in the eastern regions of Quebec (Quebec City area, North-East and East). In the North-East, half of the gene pool was explained by 3% of the founders (n=86).

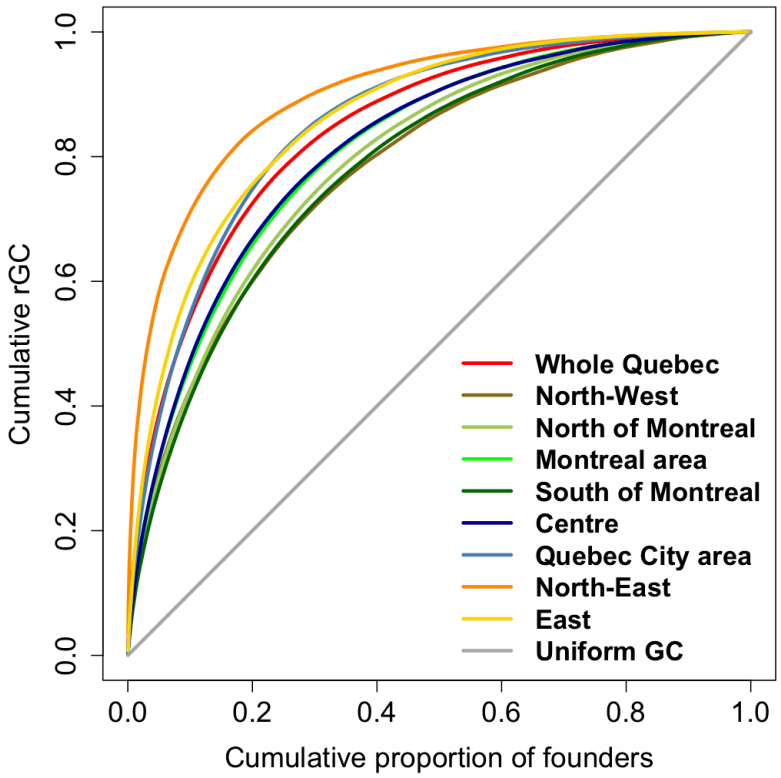

Figure 3. Cumulative distribution of the founders' relative genetic contribution (*rGC*).

Founders are plotted in decreasing order of genetic contribution. Under the hypothesis of uniform genetic contribution (Uniform *GC*), all founders would have contributed equally to a given gene pool, a linear dependence would be observed between the proportion of founders and the proportion of total genetic contribution explained by these founders.

The lowest homogeneity index was observed in the Montreal area region and the highest in the North-East (Table 4). In the Montreal region, the FUN calculation indicates that 1,787 equally contributing founders would provide the same level of diversity as the actual 6,317 founders. In contrast, in the North-East, only 160 equally contributing founders are required to provide the same level of genetic diversity as the actual 2,628 founders. The normalized

FUN/n_f ratio for the North-East of 6% shows that six equally contributing founders are equivalent to 100 actual founders. We found a higher FUN/n_f index - about 30% - in all other regions, except for the East (18.2%) and Quebec City area (21.3%), indicating a greater homogeneity of the eastern regions of Quebec. Overall, these results show a West-East decreasing gradient of genetic diversity. The FUN/n_f ratio was measured at different time points in the genealogies of each regional sample to evaluate the progression over time of the concentration of founders' genetic contribution. We observed that the FUN/n_f ratio diminishes in the first two centuries of Quebec history and stabilizes to its contemporary level for all regions in 1800 except for the North-East which does so in 1850 (Supplementary Fig. S3).

Table 4. Founders' uniform contribution number (FUN) and its ratio to the actual number of founders in each sample (FUN/n_f).

	Region	HI (x 10^4)	FUN (=1/HI)	Number of founders (n_f)	FUN/n_f (%)
1	North-West	8.5	1,177.1	3,724	31.6
2	North of Montreal	6.9	1,452.9	4,343	33.5
3	Montreal area	5.6	1,786.6	6,317	28.3
4	South of Montreal	6.2	1,604.5	4,384	36.6
5	Centre	7.3	1,375.5	4,790	28.7
6	Quebec City area	12.8	783.1	3,676	21.3
7	North-East	62.3	160.4	2,628	6.1
8	East	17.2	580.2	3,188	18.2
	Whole Quebec	6.8	1,461.0	7,798	18.7

Differential contribution of early and late founders

Figures 4A and 4B illustrate the progression of the founders' genetic contribution to the ancestor layers according to their period of arrival, for the Montreal area and North-East region (Supplementary Fig. S4 shows all regions). A greater part of the regional gene pools was contributed by the early compared to the late founders as expected in a simple model of population in expansion where every individual has the same number of

descendants and a constant number of new migrants arrive every discrete generation (for a graphic illustration of the model see Supplementary Fig. S5). All regional populations of Quebec (except for the North of Montreal) had a minimum of a third of their gene pools descending from the earliest founders, defined as those founders who married before the first major immigration wave of 1660 (Supplementary Table S4). These founders were mostly French, represented 9.3% of all founders (Supplementary Table S2) and 93% of them were found in at least six out of eight regions (75% common to all regions). For the North-East and Quebec City ancestors, the contribution of the earliest founders was even higher, increasing over time up to 50% and 44% respectively. This increase is not compatible with a simple model of population expansion (Supplementary Fig. S5). Even if we assume that the rate of migration tends to zero, the contribution of the earliest founders is expected to stabilize and not to increase. For the North-East and Quebec City area, it suggests that the earliest founders had on average a higher number of descendants than expected under the assumption of uniform reproductive success of all founders. This points to a higher reproductive success of the earliest founders and their descendants throughout the period.

A substantial fraction of regional gene pools was also explained by founders married between 1660 and 1700 (Figs. 4A-B and Supplementary Fig. S4). These founders represented 32% of all founders (Supplementary Table S2), were mostly French (97%) and had a lower regional representation than the earliest founders with 74% appearing in at least six regions (40% common to all eight regions). However, the fraction of the gene pool explained by the 1660-1700 founders differed across regions as it decreased to the profit either of the earliest founders (in the Quebec City area and North-East) and/or of latecomers, arrived after 1700 (in all regions). Notably, the East region displayed the highest contribution of late founders (Supplementary Fig. S4) who are more specific and have more diversified origins (Table 2). *A priori*, this may suggest a higher genetic diversity in the East region, but the low FUN/n_f ratio observed in this region points to a lower genetic diversity (Table

4). The FUN/n_f ratio is calculated within regions and does not take into account the origins and specificity of the founders; therefore we cannot exclude the possibility that the late founders brought new genetic variation in the East.

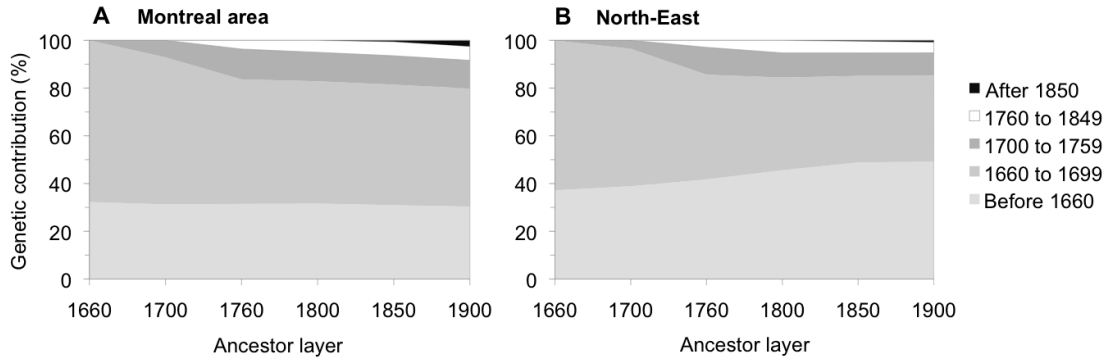

Figure 4. Progression over time of the genetic contribution of immigrant founders.

The founders' relative genetic contribution (%) to the ancestor layers is plotted according to their period of arrival for **(A)** Montreal area and **(B)** North-East regions. The ancestors in each layer were married 615 years around the selected years.

Stratification of Quebec regional populations

To detect population structure using ascending genealogies, we proposed a new approach based on PCA of immigrant founders' genetic contributions to the 2,221 subjects (GC-PCA). This analysis was compared to a MDS of pairwise kinship coefficients. Both methods showed graphical evidence for stratification of the contemporary French Canadian population (Fig. 5). North-East subjects significantly clustered together on the first axis (ANOVA *p-values* < $1x10^{-30}$ for all pairwise comparisons of regions on PC1) and East subjects did so on the second axis albeit to a lesser extent (ANOVA *p-value* < $1x10^{-11}$ for all pairwise comparisons of regions on PC2) (Fig. 5 and

Supplementary Fig. S6). Thus, patterns of founders' genetic contribution and kinship in the North-East and East regional populations appear to be distinct from the rest of Quebec. Except for the North-West and North-East regions, subjects tended to be distributed on a West-East gradient along the second PC according to their region of marriage, reflecting their geographical location along the St-Lawrence River (Fig. 5 and Supplementary Fig. S6). This is supported by a R^2 of 0.15 (*p-value* < 2.2×10^{-16}) for the linear regression of PC2 on the subjects' region of marriage (recoded 1 to 8 from west to east). The Calboli-PCA and the K-means clustering of founder / subject incidence matrix also positioned subjects along the West-East axis but provided less information on population structure than the GC-PCA (Supplementary Fig. S7).

Comparison between GC-PCA and MDS of kinship coefficients showed that both methods are highly correlated (PC1 and MDS1: Pearson's $r = 0.99$; PC2 and MDS2: Pearson's $r = 0.98$ - Supplementary Fig. S8). This was expected since for two given subjects *i* and *j*, the probability of sharing an allele identical by descent from a common founder equals the sum of products of founders' genetic contribution to subject *i* and to *j* divided by two, over all founders n_f. This probability equals the kinship coefficient when the common ancestors of two subjects are not inbred and when the founders appear at the same level of generation so that they have a uniform genetic contribution to the subjects. Hence, this explains why the GC-PCA and the MDS of kinship coefficients gave very similar results. The differences between the two methods thus reflect inbreeding among common ancestors and inequality in founders' genetic contribution.

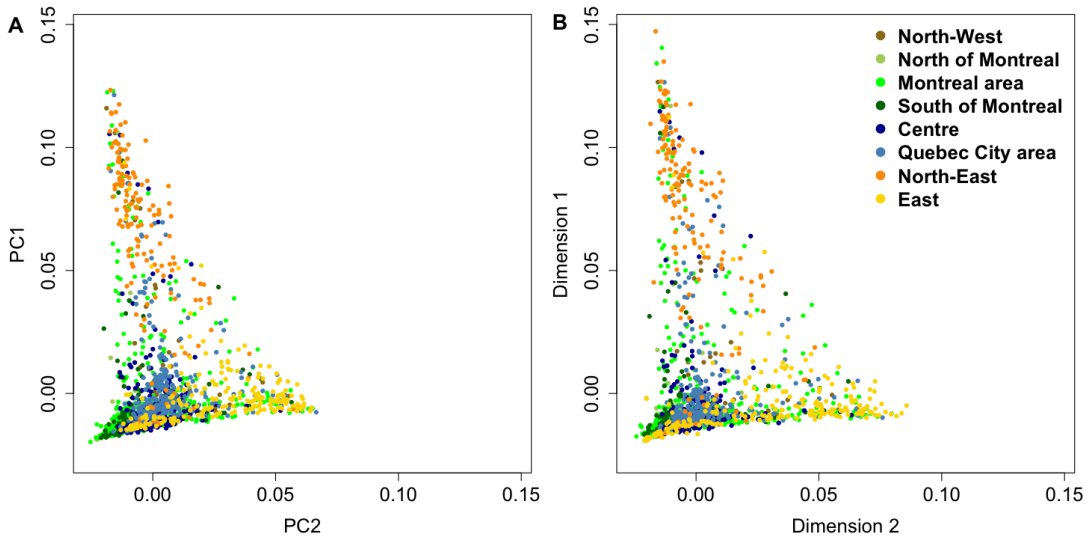

Figure 5. Structure of Quebec regional populations.

Subjects are plotted according to **(A)** the two first components of the GC-PCA (axes were rotated with the varimax method) and **(B)** the two first dimensions of the MDS based on pairwise kinship coefficients (i.e., 1-kinship). PC1 explains 6.1% of the variation and PC2 2.4%. Four outliers were excluded from the analyses (*n*=2,217).

Figure 6 shows the correlation between each founder genetic contribution and the top two principal components of GC-PCA. Founders displaying a correlation coefficient with PC1 greater than 0.2 arrived for the most part before 1700 and were common to all eight regional groupings (Fig. 6A). Moreover, founders with a high correlation to PC1 were also those who had the highest genetic contribution the North-East sample (Supplementary Fig. S9). These results suggest that the distinction between the North-East and the other regions observed on PC1 is explained by the higher contribution of early founders found in many regions rather than to the contribution of specific founders to the North-East. Figure 6B indicates that the PC2 was also mostly influenced by early founders but on the whole the correlation was not as strong. Founders who had a greater genetic contribution to the East region

tended to have a greater coefficient of correlation (Supplementary Fig. S10). Both PC1 and PC2 were negatively correlated to founders with a higher contribution to regions located West of the Quebec City area (Supplementary Fig. S10). Altogether, this analysis of PC-correlated founders recapitulated the results of our descriptive analysis of the immigrant founders' genetic contribution and clearly illustrated the differential impact of early versus late founders in shaping the contemporary structure of the Quebec population.

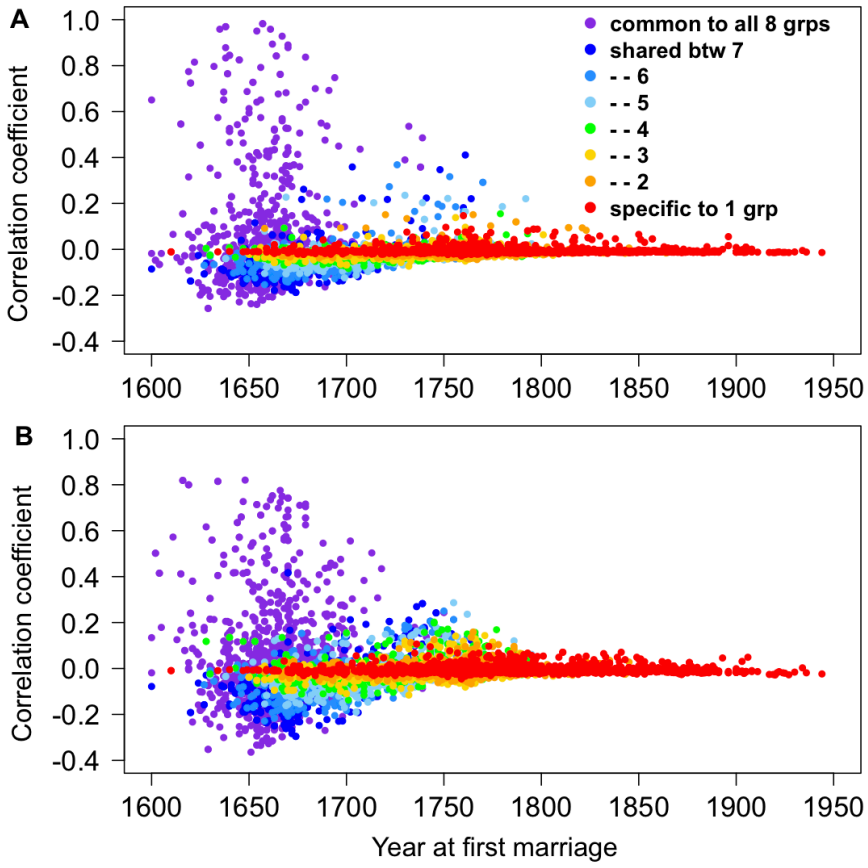

Figure 6. Correlation between the founders' genetic contribution.

Correlation between founders' GC and **(A)** PC1 and **(B)** PC2 according to year at first mariage and regional representation. Each point corresponds to a founder and is colored according to the number of regions to which they contribute.

DISCUSSION

We analyzed the population structure of French Canadian from Quebec using extensive genealogical ascendance of 2,221 subjects sampled over all the territory and partitioned in eight regional groupings. We provided evidence for stratification of the French Canadian populations at the regional level and showed that this structure was correlated with their geographical location along the St-Lawrence River. In addition, we found a West-East decreasing gradient of diversity among regional populations, consistent with a previous genealogical study of kinship and consanguinity (Vézina et al. 2004). Regional populations shared a common pool of diversity contributed by the earliest founders, mostly French, but received a differential and more specific input of the latecomers, of more diverse origins. In particular, our results contrasted the regions located West of the Quebec City area from the North-East and East regions, which both show patterns of ancestry supporting their distinct genetic identity. Taken together, our results demonstrate that regional gene pools of Quebec cannot be considered homogeneous and underline the specificity of each region that can be understood in light of its settlement and subsequent demographic history.

The North-East displayed the highest homogeneity and appeared as a distinct cluster from the rest of Quebec. The southernmost part of that region, Charlevoix, was colonized at the end of the 17th century by a small number of descendants of the pioneers coming mainly from the Quebec City area (Jetté et al. 1991). In the middle of the 19th century, settlement started in the North of the region (Côte-Nord and Saguenay-Lac-St-Jean). The Saguenay-Lac-St-Jean population was founded by pioneers coming mostly (but not exclusively) from the Charlevoix region and grew rapidly due to a particularly high fertility rate in a context of relative isolation (Roy et al. 1988; Lavoie et al. 2005). Many mutations underlying Mendelian diseases, elsewhere very rare, have reached an elevated frequency in that region, thus pointing to a strong

regional founder effect (Labuda et al. 1996; Scriver 2001; Laberge et al. 2005b; Yotova et al. 2005). In the ascending genealogies of patients affected by five of these diseases, it was previously shown that the 17th century founders had a high contribution, explaining nearly 80% of the cohorts' gene pool (Heyer 1995). The high contributors are also the most likely to have introduced the diseases' mutations. Since three of these five diseases are specific, although not exclusive, to the Charlevoix and Saguenay-Lac-St-Jean regions, we can hypothesize that the 17th century founders with high contributions were the ones that also contributed to the genetic differentiation of the region. Here, we demonstrated that the genetic differentiation of the North-East is not explained by the input of specific founders, but by the higher contribution of a subset of earliest founders. These founders were, for the most part, common to all regions but had a higher reproductive success in the North-East. This is likely to be the result of the founder effect *per se* whereby the successive settlements of the three regions comprised in the North-East were biased toward descendants of the earliest settlers. This could be simply due to random sampling, but cultural explanations have also been proposed such as kin-structured migrations (Jetté et al. 1991) and social transmission of reproductive behavior (Austerlitz and Heyer 1998).

The East region was found to be the second most homogeneous sample and was characterized by a higher contribution of later founders from Acadia, Great Britain and other parts of America. French Canadian settlement of the East, starting in Côte-du-Sud nearby Quebec City, progressively reached the Gaspe Peninsula in the 19th century, which was already occupied by descendants of the deported Acadians and of British immigrants, including Loyalists to the British Crown who took refuge on this territory after the War of Independence of the United States at the end of the 18th century (Desjardins et al., 1999). Significant genetic differentiation among self-declared descendants of the major groups of founders of the Gaspe region was found based on analysis of genomewide diversity (Roy-Gagnon et al. 2011) and parental lineages (Moreau et al. 2009). The diverse ancestry of the East

region, ranging from deep-rooted families in Côte-du-Sud to diverse ethno-cultural groups in the Gaspe Peninsula, was reflected in our population structure analysis, as subjects from that region are relatively dispersed even if they do form a significant cluster.

In our genealogical sample, we identified 5,623 immigrant founders who married before the British Conquest. They represent two thirds of the 8,570 settlers previously reported to have at least one married children in Nouvelle-France (Brais et al. 2007). The time of arrival and origins of these founders are found to be consistent with the composition of the pioneer immigration under the French rule (Charbonneau et al. 2000). For reasons of data availability, our sample comprised subjects married between 1945 and 1965. In the following decades, major migration movements linked to urbanization processes have taken place. In a recent study on the genealogical structure of the Lanaudiere region located in the periphery of Montreal, we showed that these movements were linked to a reduction in genetic differentiation and diversification of ancestry (Bhérer et al. 2008). We therefore expect that the population structure found in this study underwent some changes. However, this effect should be more pronounced in the Montreal and Quebec City areas which are the two major poles of internal migrations.

Our results confirm the genetic importance of the French immigrants in the contemporary French Canadian gene pool: they are the most numerous founders and have the highest contribution to all regional samples. However, this study also puts in the balance arguments in favor of the heterogeneity of the pool of founders. Following the British Conquest, immigration diversified and had a variable impact on the regional populations (Bergeron et al., 2008; Tremblay et al., 2009). Moreover, French immigrants came from all regions of France (Vézina et al. 2005b) and principally landed as single member of their family (Guillemette and Légaré 1989; Charbonneau et al. 1993). Contemporary regions of France have been shown to be genetically heterogeneous (Dubut et al. 2004; Richard et al. 2007). Assuming that this

was also the case during the 17th and 18th centuries, the amalgamation of founders from different French regions together with immigrants from other European and American countries is expected to have inflated the genetic diversity introduced in Quebec.

In our sample, almost all subjects had mixed origins, including French and non-French. This indicates that admixture events have shaped the genome of nearly all French Canadians, like other post-Colombian populations of the New World. Notably, half of the subjects had at least one reported Amerindian founder in their ancestry. However, while the mestizo populations of Latin America, for example, have a mean estimated proportion of Amerindian ancestry of more than 20% (Wang et al. 2008), we found that 0.2% of the French Canadian gene pool was of Amerindian origin. Though this estimate is a lower bound because historical sources do not always allow identification of Amerindian ancestors, to our knowledge this is the first estimation of the Amerindian genetic contribution to the contemporary French Canadian gene pool based on a large genealogical sample. Overall, our study also puts forward that the French Canadians descend from a pool of founders relatively large and of diverse origins. This might be sufficient to explain why the French Canadian population of Quebec has maintained comparable levels of genetic diversity as European populations (e.g. De Braekeleer 1990; Moreau et al. 2007).

In this paper, we showed that analysis of ascending genealogies allows population structure to be assessed without genotyping. A new approach was proposed to detect and visualize population stratification by PCA of founders' genetic contribution and was validated by its high correlation with a commonly used approach, namely the MDS of pairwise kinship coefficients. In a parallel work, we showed that the latter approach accurately mirrors the genetic structure in French Canadian pedigrees inferred from genomewide SNP data (Roy-Gagnon et al. 2011). By induction, GC-PCA can therefore be used to reflect genetic structure. Such PCA may not be of practical use if the only

intent is to describe genetic structure of a population since MDS of pairwise kinship coefficients is more efficient and can readily be used. However, it has the advantage of pointing out which founders have the strongest weight in creating the structure. We illustrated that the PC-correlated founders, by analogy with PC-correlated markers, can be used to inform how the demographic history of a population shapes its genetic structure. Our approach could be used in historical studies to identify and further describe the highly PCA-correlated founders. Extensive genealogies are found in a number of human populations, such as the Icelanders (Helgason et al. 2005) and the Utah residents (Cannon Albright 2008). Deep pedigrees are also found in many other species with important economic incidence such as dogs (Calboli et al. 2008) and cattle. Our approach could also be of interest for the development of breeding strategies in captive populations aiming to maintain genetic diversity.

Regarding the Quebec population, we demonstrated that genetic epidemiology studies must take into account the characteristics of regional populations to optimize their study design. For instance, in order to minimize genetic heterogeneity, research aiming to identify rare variants implicated in complex traits might be better suited in the North-East or the East of Quebec. Stratification of Quebec regional populations highlights the need to detect and correct for genetic structure in genetic association studies, especially when sampling without regard to the geographic origin of individuals. Moreover, since population structure in Quebec is linked to geography, optimal design of studies should include the place of birth of parents and grandparents to guide the selection of individuals to sample and to genotype.

In the past decade, much effort has been devoted to analyze population structure from genotype data. In this study, we show that the analysis of deep ascending genealogies can effectively reveal population structure and therefore be a useful tool to explain the consequences of historical

demographic processes in structuring of genetic variation and to develop more powerful research design in population association studies.

SUPPLEMENTARY MATERIAL

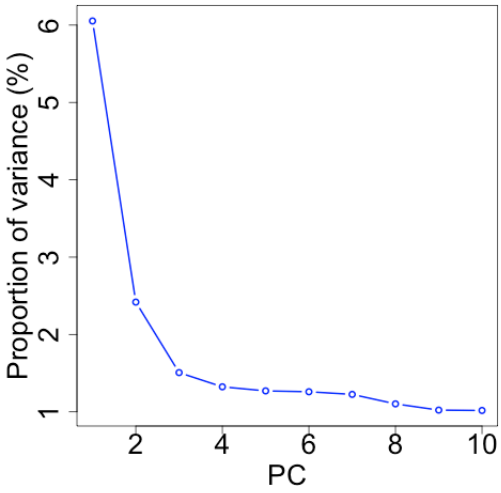

Figure S1. Proportion of variance (%) explained by the first ten PCs of the GC-PCA.

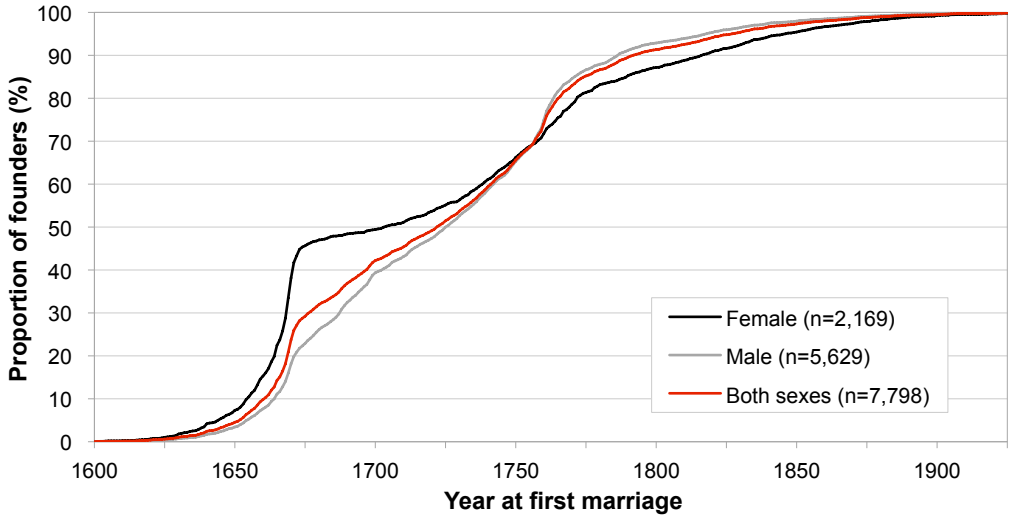

Figure S2. Cumulative distribution (%) of immigrant founders according to year at their first marriage.

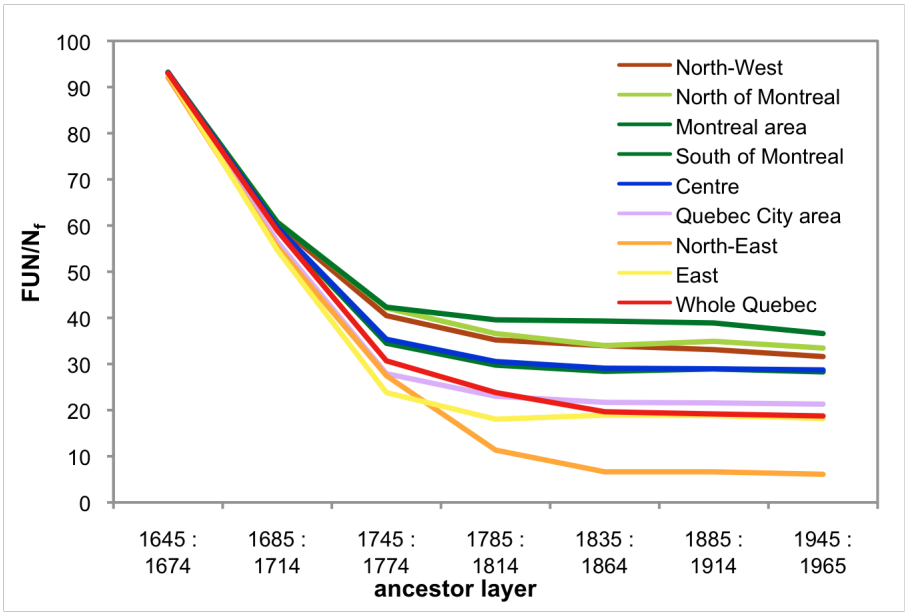

Figure S3. FUN/n_f ratio of the ancestor layers for each regional sample.

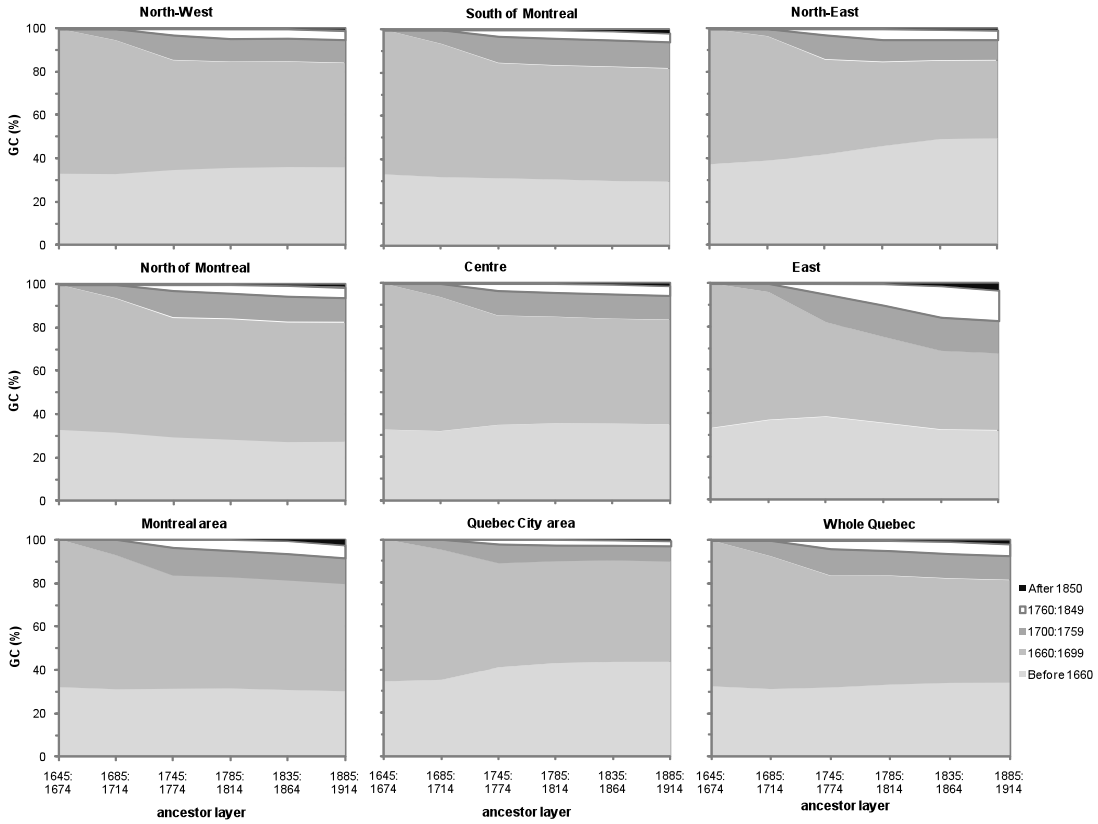

Figure S4. Immigrant founders' relative genetic contribution to the ancestor layers according to their period of arrival.

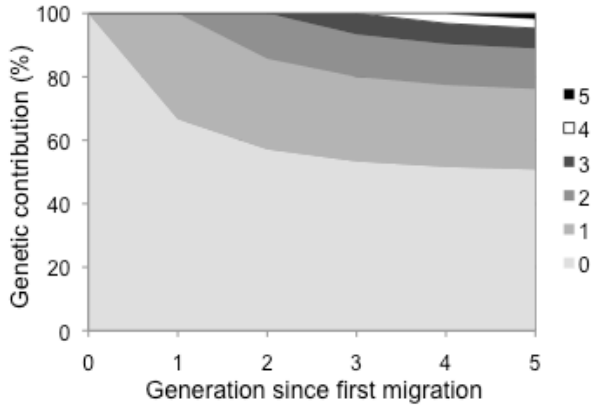

Figure S5. Expected genetic contribution of successive generations of migrants to the genetic pool of a founder population in expansion.

We assumed that every individual has the same probability to reproduce with a growth rate of r =2 descendants per individual per generation. This is close

to the growth rate of 2.13 observed in Quebec during the 18th century (Austerlitz and Heyer, 1999). In addition, we assumed a constant number of new migrants of n_m=100 individuals per generation with a sex-ratio equals to one.

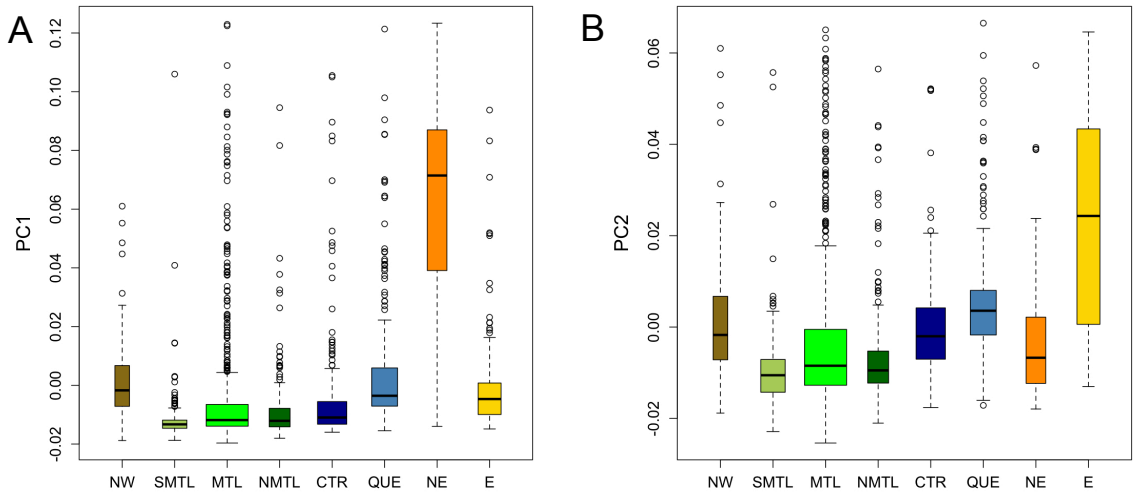

Figure S6. Boxplot of the top two axis of variation of GC-PCA.

Boxes width is proportional to the number of subjects per sample. NW: North-West; NMTL: North of Montreal; MTL: Montreal area; SMTL: South of Montreal; CTR: Centre; QUE: Quebec City area; NE: North-East; E: E

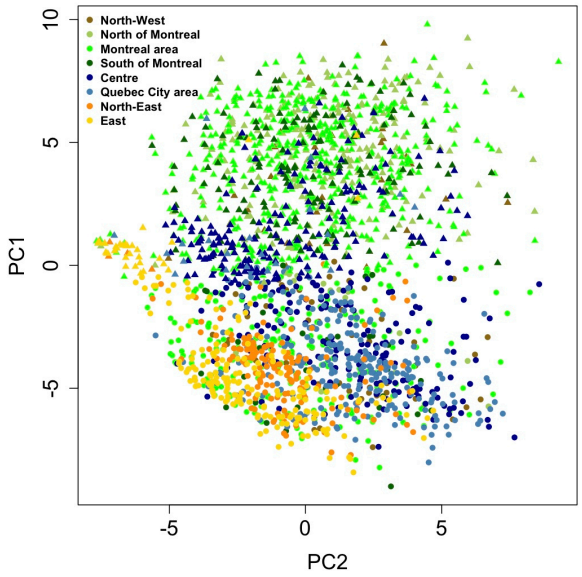

Figure S7. Top two axis of Calboli-PCA.

Each point represents a subject and is colored according to region of marriage. Circles and triangles differentiate the two clusters identified by the two-means clustering analysis. PC1 explains 4.3% of the variation and PC2 2.9%.

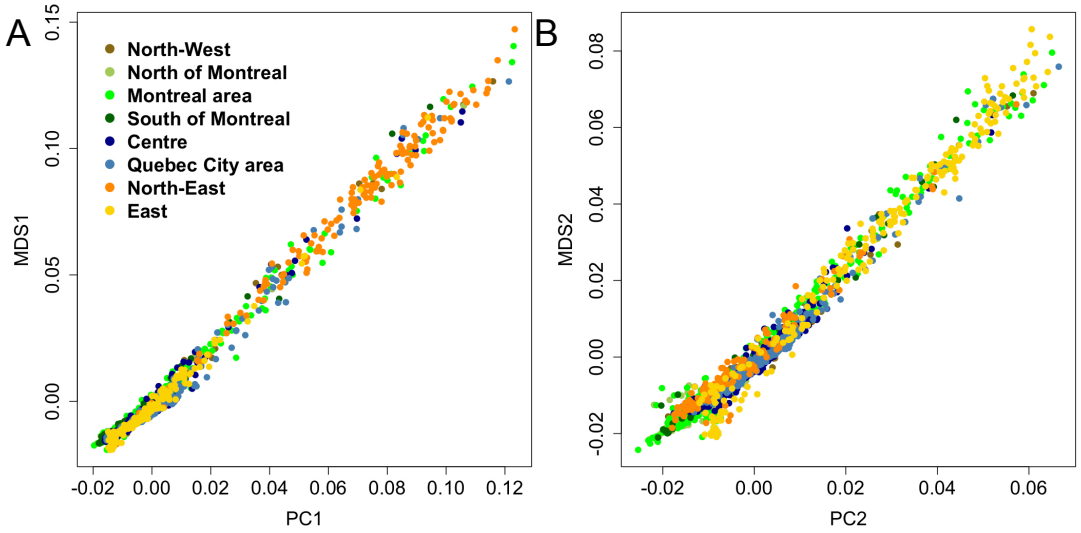

Figure S8. Correlation between GC-PCA and MDS of pairwise kinship coefficients.

(A) shows the correlation between the first principal component (PC1) and the first dimension (MDS1) (Pearson's r = 0.99, p-value < 2.2×10^{-16}). Likewise, **(B)** shows the correlation between PC2 and MDS2 (Pearson's r = 0.98, p-value < 2.2×10^{-16}).

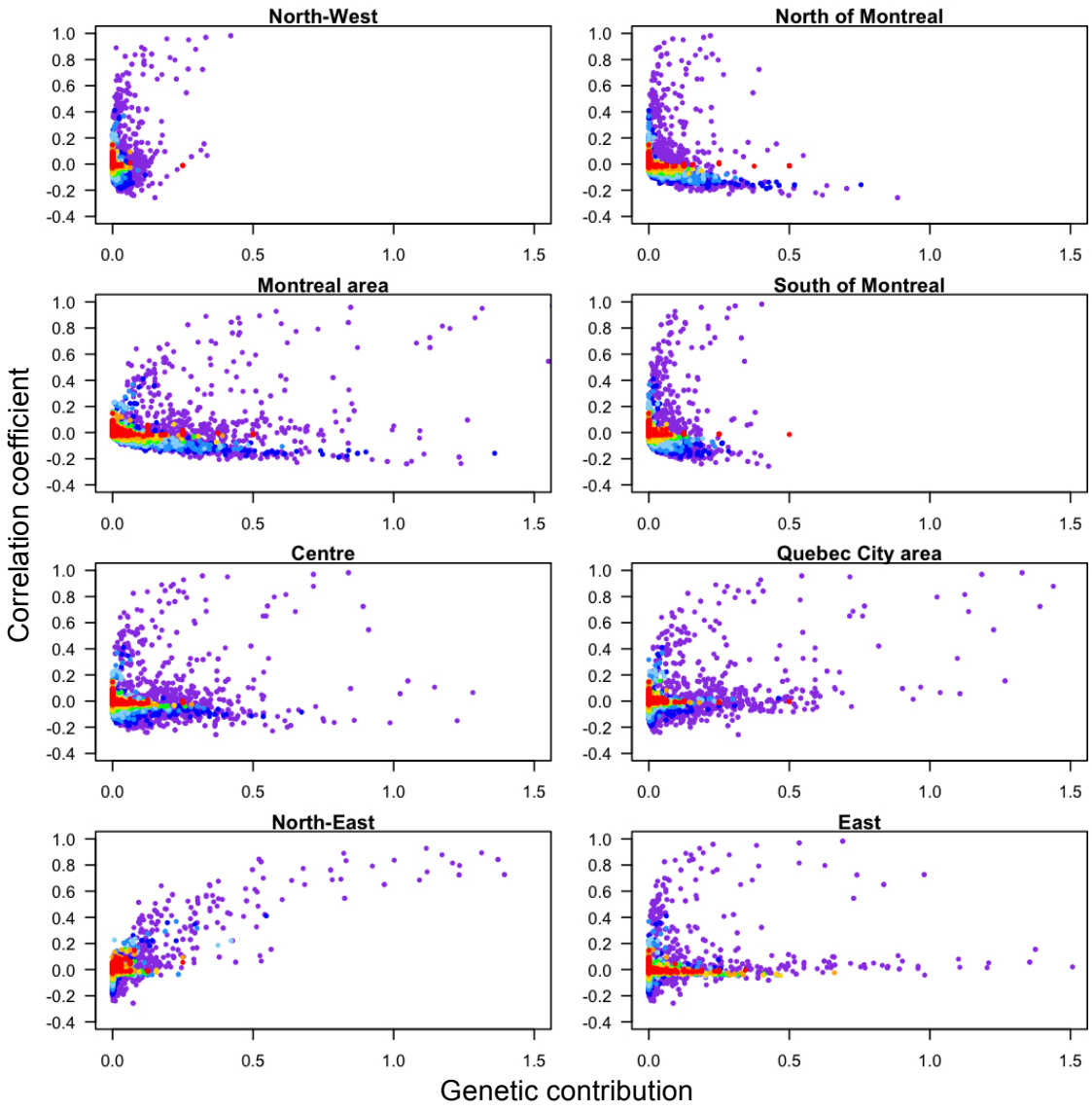

Figure S9. Correlation between founders' genetic contribution and the first component of GC-PCA according to their genetic contribution to each regional sample and their regional representation.

Each point corresponds to a founder and is colored according to the number of regional samples where they appear as in Fig.6. Note that 16 founders who had a genetic contribution of over 1.5 in the Montreal area do not appear in the plot, as well as 8 in the North-East and 1 in the East region.

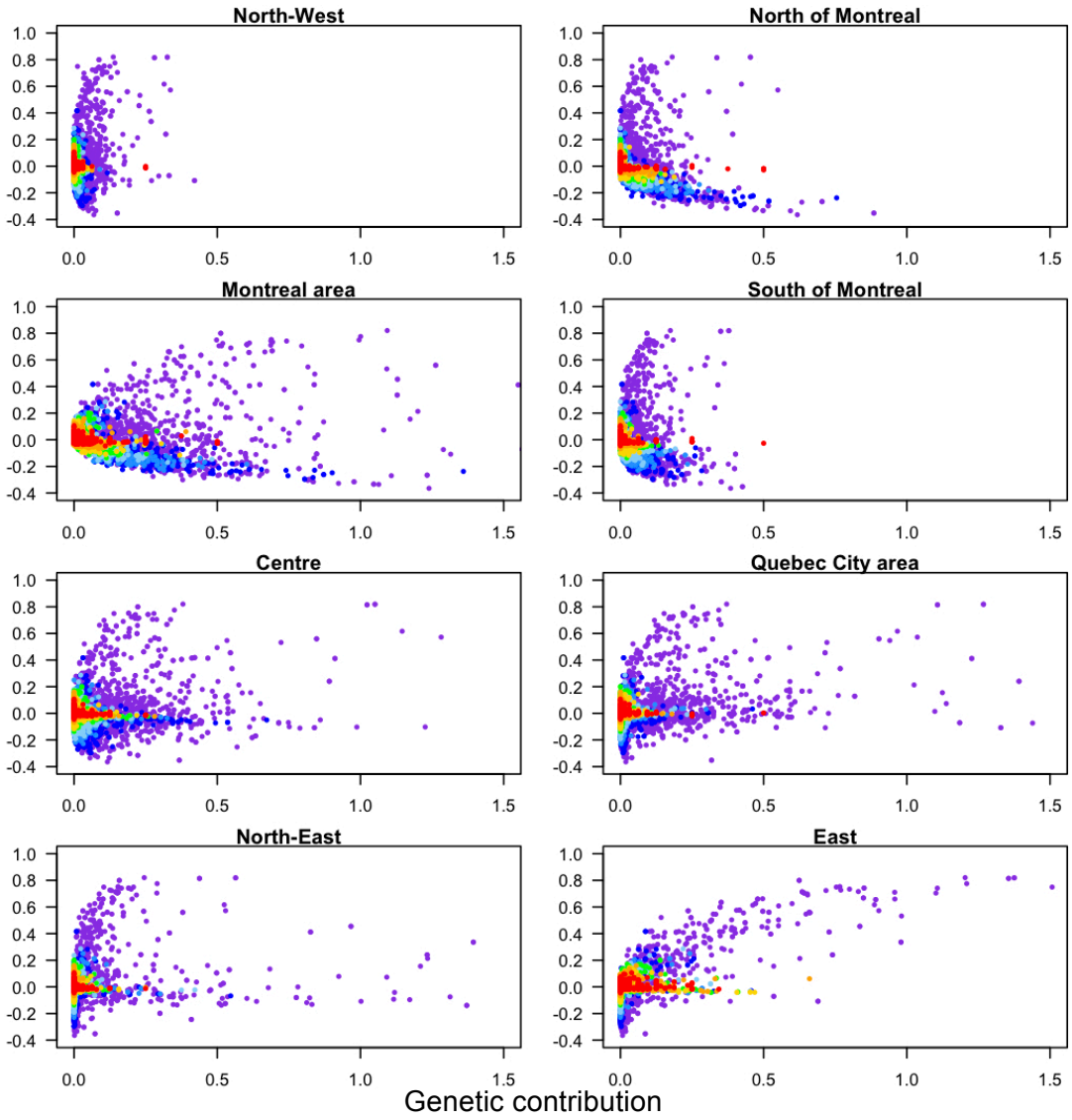

Figure S10. Correlation between the founders' genetic contribution and the second component of GC-PCA according to their genetic contribution to each regional sample and their regional representation.

Each point corresponds to a founder and is colored according to the number of regional samples where they appear as in Fig.6. Note that 16 founders who had a genetic contribution of over 1.5 in the Montreal area do not appear in the plot, as well as 8 in the North-East and 1 in the East region.

Table S1. Descriptive statistics of the genealogies.

	Region	Number of			Mean kinship coefficient (x 10^4)	Mean genealogical depth (σ)
		subjects	genealogical links	distinct ancestors		
1	North-West	87	213,048	29,557	6.64	9.6 (1.0)
2	North of Montreal	242	537,398	43,709	5.63	9.2 (1.1)
3	Montreal area	722	1,517,192	87,706	4.02	9.1 (1.3)
4	South of Montreal	178	399,338	42,114	4.41	9.2 (1.2)
5	Centre	348	784,738	54,478	5.95	9.4 (1.0)
6	Quebec City area	272	670,968	40,605	10.63	9.6 (0.8)
7	North-East	157	456,906	22,478	61.70	9.8 (1.0)
8	East	215	455,028	27,193	15.52	9.0 (1.4)
	Whole Quebec	2,221	5,034,616	153,447	5.28	9.3 (1.2)

Table S2. Distribution (%) of immigrant founders according to the period of their first marriage.

Period of marriage	Female		Male		Both sexes		Male ratio
	n	%	*n*	%	*n*	%	
Before 1660	316	14.6	406	7.2	722	9.3	1.3
[1660, 1700[	751	34.6	1774	31.5	2525	32.4	2.4
[1700, 1760[	462	21.3	1917	34.1	2379	30.5	4.1
[1760, 1800[	353	16.3	1121	19.9	1474	18.9	3.2
[1800, 1900[	260	12.0	377	6.7	637	8.2	1.5
After 1900	16	0.7	23	0.4	39	0.5	1.4
Unknown	11	0.5	11	0.2	22	0.3	1.0
Total	2169	0	5629	0	7798	0	2.6

Table S3. Distribution of the immigrant founders in the eight regions according to their origin

Region	French	British	German	Irish	Other European	Acadian	Amerindian	Other American	Unknown	All origins
North-West	87.2	1.0	0.2	0.3	0.7	8.3	0.4	1.3	0.6	100.0
North of Montreal	86.4	0.8	0.9	1.0	0.9	6.5	0.6	2.0	0.9	100.0
Montreal area	75.3	2.8	1.2	1.3	1.2	12.0	1.1	3.0	2.1	100.0
South of Montreal	86.2	0.9	0.7	0.7	0.7	7.8	0.6	1.5	0.8	100.0
Centre	80.9	1.3	0.8	0.7	1.0	12.2	0.5	1.7	1.0	100.0
Quebec City area	86.7	1.2	0.6	0.7	0.9	7.3	0.6	1.0	0.9	100.0
North-East	86.2	1.5	0.3	0.5	0.8	7.7	0.5	1.4	1.1	100.0
East	77.7	3.1	0.5	1.5	0.7	10.4	0.7	3.1	2.4	100.0
Whole Quebec	68.3	4.1	1.8	2.7	1.2	13.5	1.2	3.8	3.2	100.0
Number of founders	5,326	317	143	214	97	1,056	95	298	252	7,798

Table S4. Relative genetic contribution of founders to regional samples according to the period of first marriage

	North-West	**North of Montreal**	**Montreal area**	**South of Montreal**	**Centre**	**Quebec City area**	**North-East**	**East**	**Whole Quebec**
Before 1660	37.0	27.5	31.2	30.5	36.1	43.8	50.0	32.1	34.7
[1660, 1700[	47.6	53.5	48.5	50.3	47.3	46.4	36.0	35.4	46.6
[1700, 1760[	10.6	11.2	11.3	12.0	10.8	6.5	8.2	15.2	10.8
[1760,1850[	3.9	5.2	5.9	4.5	4.7	2.5	5.1	14.7	5.8
After 1850	0.9	2.6	2.8	2.0	1.0	0.9	0.7	2.6	1.9
Unknown	0.0	0.0	0.2	0.6	0.0	0.0	0.0	0.0	0.1
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

CHAPITRE III:

Genomic and genealogical investigation of the French Canadian founder population structure

Marie-Hélène Roy-Gagnon, Claudia Moreau, Claude Bhérer, Pascal St-Onge, Daniel Sinnett, Catherine Laprise, Hélène Vézina, Damian Labuda

Référence:

Roy-Gagnon MH, Moreau C, Bhérer C, St-Onge P, Sinnett D, Laprise C, Vezina H, Labuda D. 2011. Genomic and genealogical investigation of the French Canadian founder population structure. *Human genetics* **129**(5): 521-531.

CONTRIBUTION DES CO-AUTEURS

Pour cet article, ma contribution est la suivante:

- participation au design de l'étude ;
- collecte des données sur le terrain (i.e. recrutement de participants dans les régions de Lanaudière, Montréal, Outaouais et Québec);
- analyses pour la sélection des sujets génotypés (complétude, apparentement, consanguinité, ancrage géographique);
- développement d'une nouvelle méthode d'analyse généalogique de la structure de population;
- analyses statistiques des données généalogiques (statistiques descriptives, apparentement, consanguinité et structure);
- design des analyses génomiques sur les "Runs of Homozygosity";
- analyses comparatives entre les estimateurs d'homozygotie basées sur les données génomiques et généalogiques;
- révision du manuscript.

La contribution de mes co-auteurs est la suivante: MHRG est l'auteure principale de cette étude. Elle a dirigé le design de l'étude avec DL; supervisé les analyses bioinformatiques et statistiques des données génomiques et généalogiques, notamment les méthodes d'analyse de structure et rédigé le manuscrit. CM a collecté les données sur le terrain; supervisé la sélection des sujets et l'envoi des échantillons pour le génotypage; réalisé les analyses bioinformatiques et statistiques des données génomiques avec PSO; et réalisé les versions finales des figures de l'article. PSO a réalisé des analyses bio-informatique et statistiques des données génomiques. DS a contribué à la collecte des données en Gaspésie. CL a fourni les échantillons du Saguenay Lac-St-Jean. HV et DL ont initié et supervisé l'ensemble de l'étude, de la collecte des données aux analyses statistiques. CM, PSO, DS, CL, HV et DL ont révisé le manuscrit.

ACKNOWLEDGMENTS

We are grateful to all participants who generously shared their DNA and information required to reconstruct their genealogies and to Laurent Richard from the Historical Geography Laboratory at Laval University (QC, Canada) for cartography work. Support of the *Réseau de Médecine Génétique Appliquée* (RMGA) of the *Fonds de la Recherche en Santé du Québec* (FRSQ) as well as of the Canadian Institutes of Health Research (CIHR; to DL and HV) is gratefully acknowledged. CB was supported by a studentship from the *Fondation de l'Hôpital Sainte-Justine* and the *Fondation des Étoiles* and is now scholar from the FRSQ.

ABSTRACT

Characterizing the genetic structure of worldwide populations is important for understanding human history and is essential to the design and analysis of genetic epidemiological studies. In this study, we examined genetic structure and distant relatedness and their effect on the extent of linkage disequilibrium (LD) and homozygosity in the founder population of Quebec (Canada). In the French Canadian founder population, such analysis can be performed using both genomic and genealogical data. We investigated genetic differences, extent of LD, and homozygosity in 140 individuals from seven sub-populations of Quebec characterized by different demographic histories reflecting complex founder events. Genetic findings from genome-wide single nucleotide polymorphism data were correlated with genealogical information on each of these sub-populations. Our genomic data showed significant population structure and relatedness present in the contemporary Quebec population, also reflected in LD and homozygosity levels. Our extended genealogical data corroborated these findings and indicated that this structure is consistent with the settlement patterns involving several founder events. This provides an independent and complementary validation of genomic-based studies of population structure. Combined genomic and genealogical data in the Quebec founder population provide insights into the effects of the interplay of two important sources of bias in genetic epidemiological studies, unrecognized genetic structure and cryptic relatedness.

INTRODUCTION

Recently, there has been a strong interest in characterizing the genetic structure of worldwide populations using genome-wide Single Nucleotide Polymorphism (SNP) panels. These studies revealed fine structure in European (Seldin et al. 2006; Heath et al. 2008), Japanese (Yamaguchi-Kabata et al. 2008) and Indian populations (Reich et al. 2009), among others, as well as Finnish (Jakkula et al. 2008) and Icelandic (Price et al. 2009) founder populations. Apart from their important contribution to our understanding of human history, studies characterizing the structure of human populations are essential to the sound design and analysis of genetic epidemiological studies. Indeed, false positive associations can be found at higher rates when phenomena such as population stratification or cryptic relatedness are observed. Population stratification occurs when the overall study population consists of genetically distinct subgroups. When undetected or unaccounted for, it can cause false positive results or mask true results in population-based association studies (Freedman et al. 2004; Marchini et al. 2004). Cryptic relatedness refers to the presence of unknown (or unaccounted for) biological relationships among participants of a study and can also lead to false positive results (Devlin and Roeder 1999; Voight and Pritchard 2005). The extent to which results are affected by these phenomena depends of course on the specific characteristics of the population under study (levels of population structure and kinship, marker allele and disease frequencies, etc.). Here we examined genetic structure and distant relatedness in the founder population of Quebec (Canada) and their effect on the extent of linkage disequilibrium (LD) and homozygosity. With the accessibility and high quality of genealogical data documenting distinct settlement and migration patterns over four centuries, the Quebec founder population provides a unique model to investigate these phenomena using both genomic and genealogical data.

About 80% of the Quebec population (7.8 million) is French speaking. Most of them, hereafter called French Canadians, descend from ~8,500 settlers who came mostly from France over the span of a century and a half starting in 1608 (Charbonneau et al. 2000). Following the British Conquest of 1759, French immigration practically ceased and the French Canadian population expanded rapidly in a context of relative isolation caused by linguistic and religious barriers. Throughout the 19th and 20th century, immigrants of various origins mixed into the French Canadian population with a very limited genetic impact (Vézina et al. 2005b; Bhérer et al. 2011). Population growth led to the colonization of new regions of Quebec, including remote and isolated regions, favoring population subdivision. In this study we focus on the two main cities of Quebec, Montreal and Quebec City, and also on three peripheral regions located in the eastern part of the province: the Gaspesian Peninsula (Gaspesia), Saguenay–Lac-St-Jean and the western part of the North Shore (Figure 1).

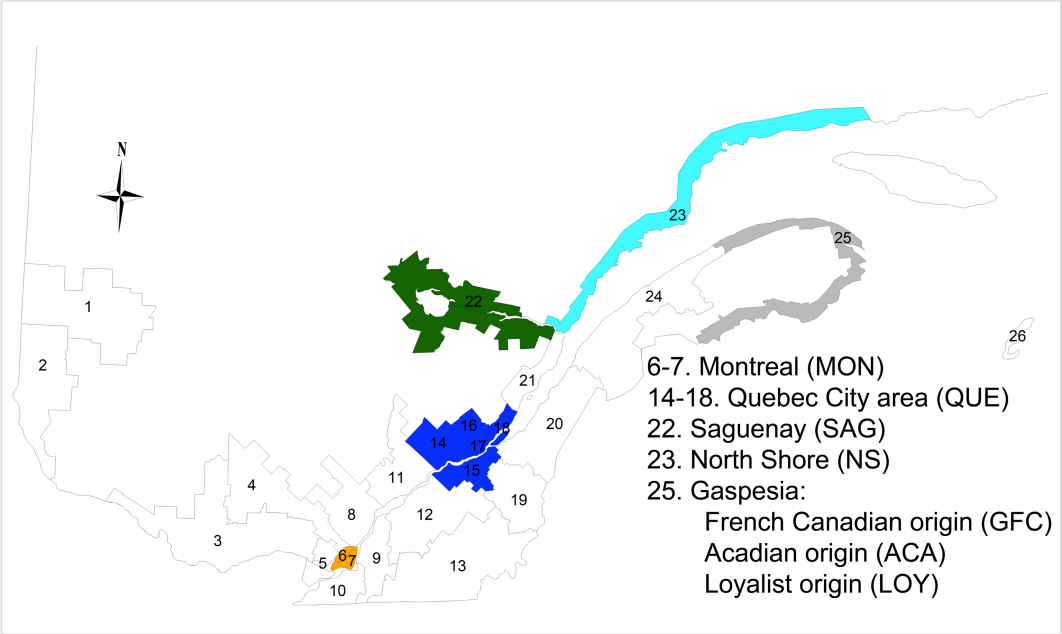

Figure 1. Map of Quebec regions.

The regions where participants were recruited for each regional or ethno-cultural sample are indicated.

Permanent European settlement began in Gaspesia during the second half of the 18th century with the arrival of Acadians, descendants of French pioneers in Acadia (located in sectors of present-day Nova Scotia, New Brunswick and Prince-Edward Island) who escaped deportation by the British (Desjardins et al. 1999). They were soon joined by English-speaking United Empire Loyalists who chose to remain under British rule after the American Declaration of Independence. During the 19th century, many French Canadians from the lower part of the St. Lawrence valley were attracted to Gaspesia for its developing fishing, naval, and lumber industries (Desjardins et al. 1999). These three ethno-cultural populations (French Canadians, Acadians, and Loyalists) married mostly among themselves (Desjardins et al. 1999). The settlement of Saguenay started in the 1840s with French Canadians coming from the neighboring region of Charlevoix and subsequently from other regions of the St. Lawrence Valley. From 5,000 inhabitants in 1850, the Saguenay population is now 273,000 mostly due to a high birth rate (Pouyez and Lavoie 1983; Institut de la statistique du Québec 2010). The western part of the North Shore was mostly colonized by French Canadians from the regions of Charlevoix and Bas-St-Laurent between 1840 and 1920 (Frenette 1996; Institut de la statistique du Québec 2010). We also included in our study French Canadians from Montreal and Quebec City. The French-Canadian population of these two cities is composed of descendants of the first European settlers but also of the migrants from rural regions who moved to urban areas in the context of urbanization and industrialization processes in the 19th and 20th century.

An important advantage of the Quebec population for genetic research is the availability of major population registers, such as the BALSAC population register and the Early Quebec Population Register. The information contained in these databases comes primarily from vital statistics (births, marriages, deaths). As of September 2010, the BALSAC population register contains about 2.9 million records which have been computerized and linked to cover the whole province for the 19th and 20th centuries (mostly marriage records)

(Bouchard and Vézina 2009). The Early Quebec Population Register contains all records from the beginning of settlement (1608) to 1800 for a total of 700,000 records (Desjardins 1998). Using these population registers, it is possible to reconstruct ascending genealogies of subjects from the present-day population going back over four centuries.

Using these genealogical data and genome-wide genotypic data, our goal was to gain insights about the effects of complex founder events on the genetic characteristics of the resulting population. We thus investigated genetic differences, in terms of allele frequencies and sharing, extent of LD, and homozygosity, in seven sub-populations of Quebec (Figure 1) characterized by different demographic histories: French Canadians, Acadians, and Loyalists from Gaspesia as well as French Canadians from Saguenay, North Shore, Quebec City, and Montreal. Genetic findings were correlated to genealogical information for each of these sub-populations. We also situated our Quebec samples among the International HapMap Consortium samples (International HapMap et al. 2007) and the French samples of the Human Genome Diversity Panel (HGDP French) (Cann et al. 2002). Our genomic data showed significant population structure and relatedness present in the contemporary Quebec population. Our extended genealogical data corroborated these findings and indicated that this sample structure reflects the settlement patterns, providing a validation of genomic-based studies of population structure.

MATERIALS AND METHODS

Study populations and data collection

We recruited individuals from seven sub-populations. All participants provided informed consent and the study was approved by the CHU Sainte-Justine Ethics Committee. Only individuals more distantly related than 1st cousins were retained in the genotyped sample. For the Gaspesian Peninsula sub-populations, we obtained peripheral blood samples from volunteers who described their ethnic affiliation as French Canadian, Acadian, or Loyalist (Moreau et al. 2009). DNA was extracted using the Puregene DNA Purification kit (Gentra). For the North Shore, Montreal, and Quebec City sub-populations, saliva samples from volunteers were obtained using the Oragene DNA kit (DNA Genotek). For Saguenay, we selected unaffected individuals (one per family) from an ongoing family study of the genetics of asthma (Poon et al. 2004). Families were recruited for this study through affected probands from the Saguenay region who had all four grandparents of French Canadian origin. In an effort to exclude recent migrants to the different regions of Quebec, whenever possible we selected participants with at least one parent born in the region before 1960 or who were themselves born in the region before 1960. Genealogies were reconstructed as far back as possible using the BALSAC population register and the Early Quebec Population Register. Additional sources, such as marriage repositories and family directories were also consulted as needed. Using these genealogies, we confirmed that individuals in our study were not closely related. Apart from three outlier pairs of Acadian individuals with kinship coefficients between 0.03125 (equivalent to first cousins once removed) and 0.05575 (less related than first cousins), only 0.5% of pairs had kinship coefficients between 0.015625 (2nd cousins) and 0.03125. All other pairs had kinship coefficients lower than 0.0155.

For comparison purposes, we downloaded data from two open access sources: the International HapMap project (International HapMap et al. 2007) and the Human Genome Diversity Panel (HGDP) (Cann et al. 2002). We used release 27 of the HapMap data (II+III) and retained the founders of the HapMap samples: 119 CEU, 120 YRI, 90 CHB, and 91 JPT. We downloaded genotypic data on the 29 French samples from HGDP (not including the French samples of Basque origin). Genomic positions are according to NCBI build 36.

Genotyping and quality control

One hundred and forty-three individuals were genotyped on Illumina HumanHap650Y arrays at the McGill University and Genome Quebec Innovation Center according to the recommended protocols. We performed quality control for the entire Quebec sample. Quality control filters were applied at the individual and SNP levels using the PLINK software v1.05 (Purcell et al. 2007). We retained individuals with at least 90% genotypes among all SNPs, yielding a final sample size of 140 people: 20 Gaspesian French Canadians, 20 Acadians, 20 Loyalists, 22 from Saguenay–Lac-St-Jean, 20 from the North Shore, 16 from the Quebec City area, and 22 from Montreal (Table S1, Figure 1). At the SNP level, we retained SNPs with at least 90% genotypes among all individuals and we only analyzed common SNPs (MAF $\geq$ 5%) located on the autosomes and in Hardy Weinberg equilibrium [exact test (Wigginton et al. 2005), p-value > 0.001], yielding 540,078 SNPs. The same quality control criteria were applied separately to the HapMap CEU and HGDP French data (after retaining only SNPs overlapping with those on the Illumina HumanHap650Y array), yielding 539,101 and 542,155 SNPs, respectively.

Statistical analysis

Genomic data

For each SNP, we considered the ancestral allele to be that allele present in the chimpanzee if available or if unavailable in the orangutan or macaque (UCSC, February 2009 assembly). SNPs for which the ancestral allele could not be identified were assigned the HapMap CEU major allele (three SNPs; results were not affected by the exclusion of these SNPs, data not shown). Using the number of SNPs retained after the quality control filters noted above, we estimated the ancestral allele frequencies and pairwise r^2 and D' (up to 15 Mb) in each population from the maximum-likelihood estimates of the two-SNP haplotype frequencies using the expectation-maximization (EM) algorithm implemented in Haploview (Barrett et al. 2005). To avoid bias in the comparison of LD levels due to differences in sample sizes, we randomly selected 16 individuals per population to estimate LD levels in order to obtain equal sample sizes. Sensitivity analysis using different sub-samples of 16 individuals yielded similar results (data not shown). Principal components analysis (PCA) on the genotypic data was performed using the EIGENSOFT software version 2.0 (Patterson et al. 2006) with default parameters. To remove the effect of LD on the PCA, we used the PLINK software to select SNPs in approximate linkage equilibrium (pairwise $r^2 < 0.2$ in sliding windows of size 50 shifting every five SNPs), yielding 66,378 SNPs in the Quebec population, 58,627 SNPs when merged with the HapMap CEU and HGDP French populations, and 35,712 SNPs when merged with all HapMap populations and HGDP French. In addition to a formal test for population structure (based on a Tracy-Widom statistic), EIGENSOFT provides classical analysis of variance (ANOVA) tests of differences in mean values for each principal component across sub-populations as well as estimates of genomic inflation factors (Devlin and Roeder 1999) when case-control status is specified. Fst statistics (Reynolds et al. 1983; Slatkin 1995) and associated *p*-values based on 110 permutations were obtained using the Arlequin software

version 3.11 (Excoffier et al. 2005) on the subset of SNPs in low LD. The number and length of ROHs were investigated using the PLINK software using all SNPs satisfying quality control filters. We considered segments of 1 Mb or longer with 100 consecutive homozygous SNPs (at least 1 SNP per 50 kb) as extended ROHs. These were identified by sliding windows of size 5000 kb containing a minimum of 50 SNPs. A maximum of one heterozygote and of 5 missing genotypes were allowed within each window. A genomic estimate of inbreeding was obtained from these ROHs for each individual by taking the genomic length of all ROHs of at least 2.5 Mb divided by the total genomic length scanned by the sliding windows (McQuillan et al. 2008).

Genealogical data

Completeness of the genealogical data, kinship and inbreeding coefficients were calculated using the S-Plus® 8.0 function library GenLib. Kinship and inbreeding were calculated using Karigl recursive algorithm (Karigl 1981) on all available information for each genealogical lineage (mean depth of lineages: nine generations, maximum depth: 17 generations). Multidimensional scaling (MDS) was performed on the matrix of kinship distance between individuals, i.e., distance = 1-kinship estimate. PCA was also performed on a matrix of individual ancestors' origins. The geographic or ethno-cultural origins of ancestors were defined as the region of marriage of their parents within Quebec (among the 26 regions illustrated on Figure 1 and listed in Table S2, or unknown) for the non-immigrant ancestors or as their place of origin (France, Acadia, Great Britain, United States and Canada, or other) for the immigrant founders. In each individual genealogy the proportion of ancestors from each origin was calculated, yielding a matrix with 140 rows (number of individuals) and 32 columns (number of origins) for the PCA analysis. MDS and PCA of genealogical data were performed using S-Plus® 8.0. The R statistical environment version 2.7.2 was used for additional

programming and graphing. We also performed a multivariate regression analysis of a distance matrix (Zapala and Schork 2006) to test the association between the geographical and ethno-cultural origins of ancestors and variation in dissimilarities among individuals with respect to genetic sharing using the web application provided by the authors. The distance matrix (multivariate response) entry for each pair of individuals was equal to one minus the proportion of alleles shared identical-by-state (IBS), calculated with the PLINK software. The independent variables were the proportions of ancestors from each origin (including the 32 origins defined above). We used a permutation-based test to assess significance (Zapala and Schork 2006).

RESULTS

Genomic view of Quebec genetic structure

Using genotypic data on 140 individuals from Quebec (16 to 22 from each sub-population), we first examined allelic frequency differences between Quebec and the two reference populations (HapMap CEU and HGDP French) for the common autosomal SNPs (MAF≥5% in at least one population) of the Illumina HumanHap650Y array. A high correlation was observed between allele frequencies of common SNPs in Quebec and in the reference populations (Pearson's coefficient of 0.98 for HapMap CEU and 0.97 for HGDP French, Figure S1), consistent with similar distributions of common single nucleotide variations in these three populations (Figure S2). Correlation of common (MAF≥5% in at least one sub-population) SNPs allele frequencies was also high among Quebec regional and ethno-cultural populations (Pearson's coefficient greater than 0.90, Figure S3).

Table 1. Pairwise Fst statistics for the 7 sub-populations from Quebec.

	MON	QUE	GFC	LOY	ACA	NS	SAG
MON							
QUE	0.0008						
GFC	0.0023	0.0020					
LOY	0.0023	0.0019	0.0033				
ACA	0.0059	0.0055	0.0063	0.0068			
NS	0.0032	0.0032	0.0047	0.0045	0.0080		
SAG	0.0030	0.0027	0.0041	0.0041	0.0075	0.0012	

All *p*-values = 0 except MON-QUE *p*-value = 0.153, based on 110 permutations.

We calculated Fst statistics to assess population subdivision within Quebec and between Quebec and the two reference populations. We observed Fst values of 0.0014 and 0.00078 between the Quebec sample, taken as a whole, and HapMap CEU and HGDP French, respectively. Fst between the

Montreal sample and HapMap CEU and HGDP French were 0.0020 and 0.0012, respectively. Within Quebec (Table 1), Fst values ranged from 0.001 (between Quebec City and Montreal, and Saguenay and North Shore) to 0.008 (between Acadians and North Shore, and Acadians and Saguenay). We also investigated genetic structure within Quebec using PCA of the genotypic data as implemented in the EIGENSOFT software (Patterson et al. 2006). First, we situated Quebec among the reference populations of HapMap (CEU, YRI, CHB, and JPT) and HGDP French. As expected, Quebec individuals clustered with the HapMap CEU and HGDP-French individuals (Figure S4). However, finer structure could be detected when PCA was performed for Quebec only (Figure 2A-B). This analysis identified five sub-populations. Eigenvectors 1 and 2 (Figure 2A, Table S3) distinguished between Gaspesian Acadians, Saguenay-North Shore, Montreal-Quebec City area, and Gaspesian Loyalists-French Canadians, while eigenvector 3 further separated Gaspesian Loyalists and French Canadians (Figure 2B, Table S3). As observed with PCA from other founder populations (Jakkula et al. 2008; Price et al. 2009), although the first few principal components explained a small proportion of the overall variance (1.05, 0.94, and 0.84% for the first three components), this proportion was higher than that expected by chance (Tracy-Widom statistics *p*-values<5.4E-25, Table S4). A large number of SNPs across the genome contributed to these principal components as opposed to a few highly differentiated SNPs (Figure S5).

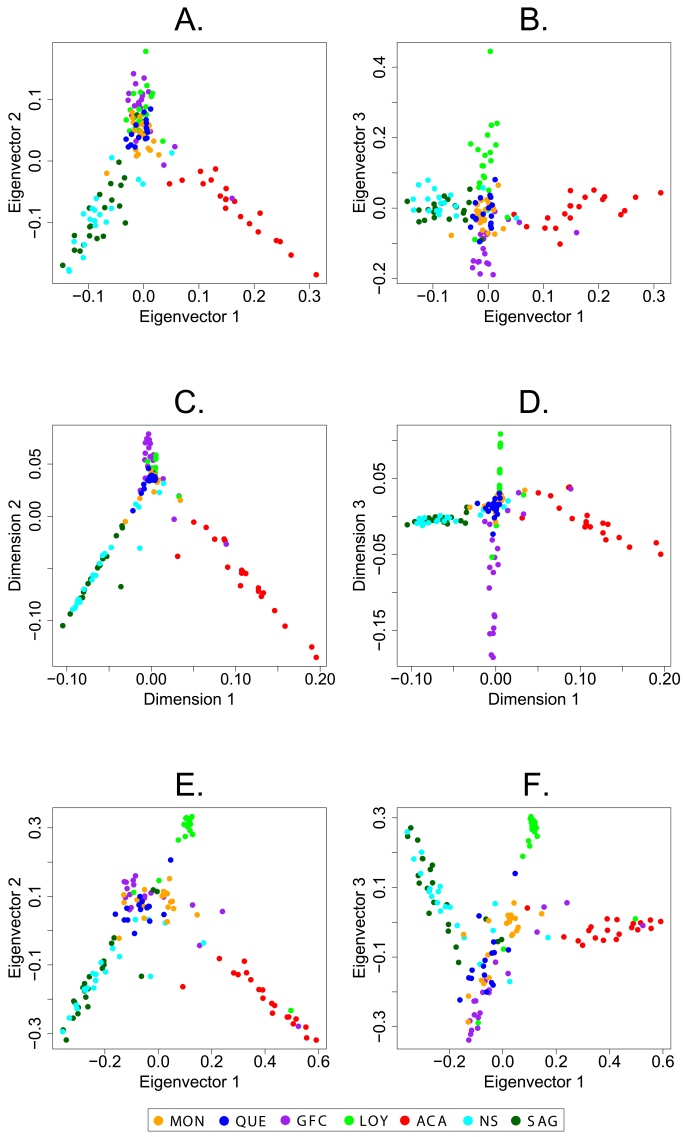

Figure 2. Quebec population structure captured by genomic and genealogic data.

(A-B) Plots of the first three eigenvectors from the PCA of the genomic data. **(C-D)** Plots of the first three dimensions from MDS with distance matrix based on genealogic estimates of kinship. **(E-F)** Plots of the first three eigenvectors from the PCA of geographical and ethno-cultural origins. MON=Montreal, QUE=Quebec City area, GFC=Gaspesian French Canadians, LOY=Loyalists, ACA=Acadians, NS=North Shore, SAG=Saguenay.

We estimated genomic inflation factors (Devlin and Roeder 1999) among pairs of Quebec regional and ethno-cultural populations using EIGENSOFT by assuming that study cases came from one population and controls from another. The genomic inflation factor measures the increase in the median association test statistic resulting from population stratification. These factors ranged from 1.1 to 1.4 (Table S5), indicating that association studies in Quebec may yield false positives if careful matching on the sub-population or adjustment is not performed. Even regional matching could fail to control for population stratification as illustrated by the region of the Gaspesian Peninsula, which includes three ethno-cultural populations identified as genetically distinct with our analyses (see also Moreau et al. 2009). Estimated genomic inflation factor among these three sub-populations was above 1.3 between Gaspesian French Canadians and Acadians and also between Loyalists and Acadians, and was 1.2 between Gaspesian French Canadians and Loyalists.

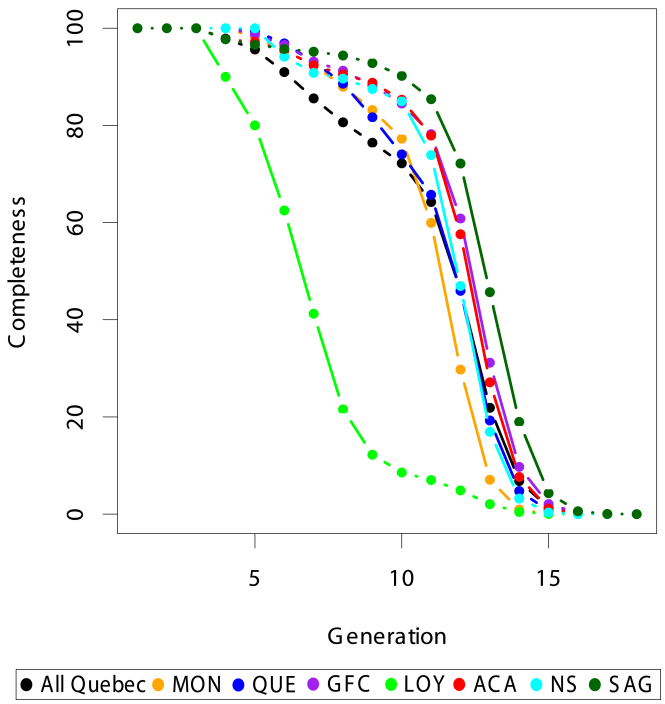

Figure 3. Completeness of the genealogic data.

MON=Montreal, QUE=Quebec City area, GFC=Gaspesian French Canadians, LOY=Loyalists, ACA=Acadians, NS=North Shore, SAG=Saguenay.

Genealogical view of Quebec genetic structure

Figure 3 shows the completeness of our genealogical data at each generation as measured by the proportion of ancestors observed in the data at a given generation divided by the expected number of ancestors (i.e., the maximum number that can be observed for a given generation). Except for the Gaspesian Loyalists sample, genealogies were over 90% complete up to the 5^{th} generation and over 80% up to the 9^{th} generation (Figure 3). The Gaspesian Loyalists had lower completeness mainly because they arrived later in Quebec and to a lesser extent because Protestant records were far less complete and well kept than Catholic records (which cover French Canadians and Acadians). At the 10^{th} generation, the majority (78%) of pairs of individuals are related. Average kinship coefficients estimated from the genealogical data overall, within sub-populations, and between sub-populations are shown in Figure 4 and S6. At the 10^{th} generation, estimated average kinship was $\leq$ 0.001 in Montreal, Quebec City, and Gaspesian Loyalists, 0.002-0.004 in Gaspesian French-Canadians, Saguenay, and North-Shore, and 0.009 in Acadians (Figure 4), indicating some population structure. Kinship also varied between sub-populations, with Loyalists showing the lowest level of relatedness with the others (Figure S6). To corroborate our genomic results of population structure, we performed multidimensional scaling (MDS) using a distance matrix based on the genealogical kinship coefficients (i.e., distance was taken as 1-kinship estimate). Figure 2C-D shows that the first three dimensions of the MDS produced results strikingly similar to those observed with the PCA on genomic data.

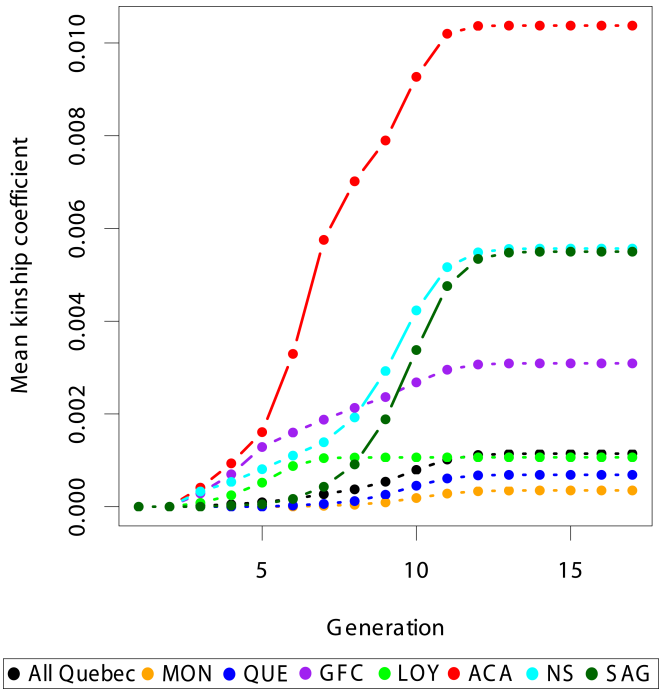

Figure 4. Average kinship estimated from genealogical data.

MON=Montreal, QUE=Quebec City area, GFC=Gaspesian French Canadians, LOY=Loyalists, ACA=Acadians, NS=North Shore, SAG=Saguenay.

To further support that the observed population structure is consistent with the founder events that took place in Quebec, we used the geographical and ethno-cultural origins of all ancestors present in the genealogies. For each individual, we calculated the proportion of ancestors from each geographic or ethno-cultural origin and performed a PCA on these data. Figure 2E-F shows plots of the first three eigenvectors from this PCA, which identified structure similar to that obtained from genotypic data and genealogical estimates of kinship. Using multivariate regression analysis of a distance matrix (Zapala and Schork 2006), we found that the geographical and ethno-cultural origins of ancestors were significantly associated with variation in genetic sharing as

estimated from the genomic data (p<0.001 for most origins, together explaining 23% of the variation in genetic sharing).

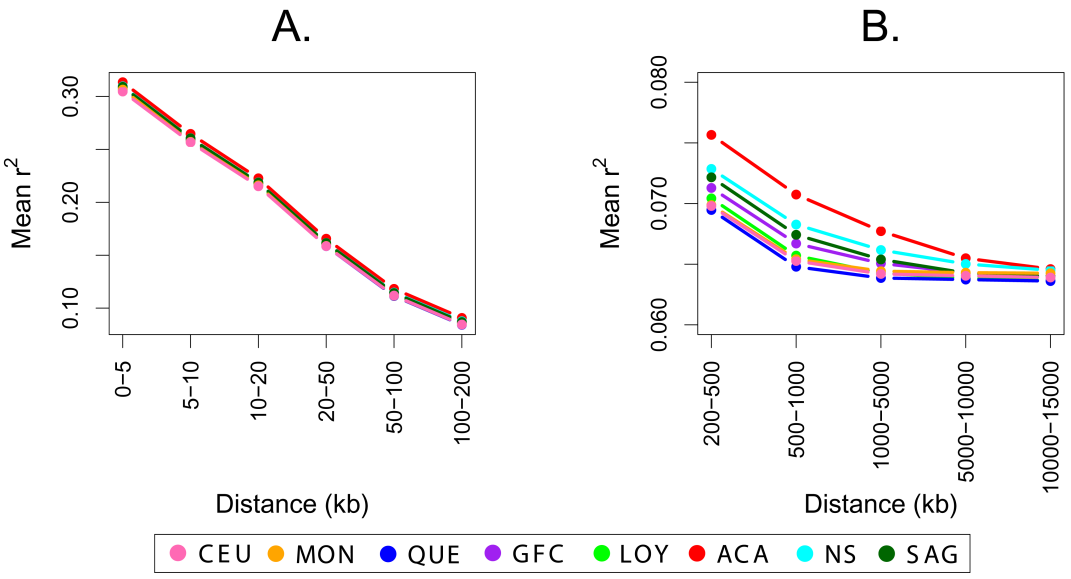

Figure 5. Average LD over the genome in Quebec and HapMap CEU.

Average r^2 estimates shown were obtained from 16 randomly selected individuals from each Quebec regional or ethno-cultural population and HapMap CEU. **(A)** SNPs located <200kb apart. **(B)** SNPs located >200 kb apart. MON=Montreal, QUE=Quebec City area, GFC=Gaspesian French Canadians, LOY=Loyalists, ACA=Acadians, NS=North Shore, SAG=Saguenay.

Linkage disequilibrium (LD) and extended runs of homozygosity (ROH)

We investigated the effects of the population structure resulting from the founder events on the extent of LD and homozygosity present in the sub-populations. Average pairwise r^2 for SNPs located within 15 Mb of each other

is shown in Figure 5. LD was slightly higher in Acadian, North Shore, and Saguenay individuals, especially for long-range LD (Figure S7), while it was similar in HapMap CEU, Montreal, and Quebec City area. Strong LD ($r^2 \geq 0.8$ or $D'=1$) was similar across populations but slightly higher in Acadians (Table S6). Lastly, we described extended ROHs within Quebec and compared it to the two reference populations. As shown in Figures 6A-C and S8, the number and length of extended ROHs was similar in HapMap CEU, HGDP French, Montreal, and Quebec City but where higher in the other regions of Quebec, where within-population relatedness estimates were also higher. Genealogical and genomic-based estimates of inbreeding were highly correlated (Pearson's correlation coefficient of 0.87, Figure 6D).

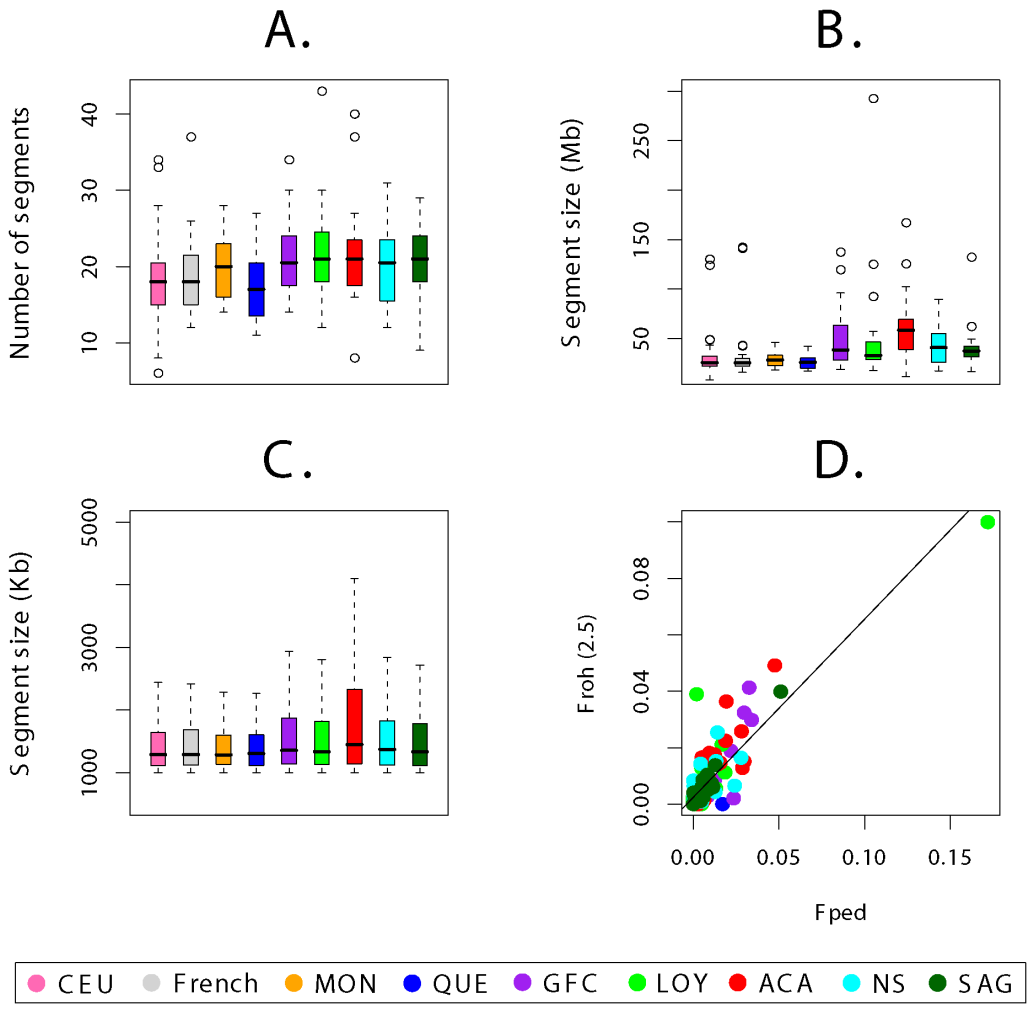

Figure 6. Distribution of extended Runs of Homozygosity (ROHs) for Quebec, HapMap CEU, and HGDP French.

(A) Number of ROHs longer than 1 Mb per individual. **(B)** Total length covered by ROHs per individual. **(C)** Length of the ROHs that are longer than 1 Mb, outliers excluded for clarity. **(D)** Correlation between genealogic inbreeding coefficient estimates (Fped) and genomic estimates based on ROHs longer than 2.5 Mb (Froh) in the Quebec population. MON=Montreal, QUE=Quebec City area, GFC=Gaspesian French Canadians, LOY=Loyalists, ACA=Acadians, NS=North Shore, SAG=Saguenay.

DISCUSSION

Using dense genotypic data and extended genealogical data, we provided evidence for population structure in the French Canadian founder population of the Quebec province of Canada. This structure is consistent with Quebec's settlement history where colonization of new regions after the initial French immigration wave led to population differentiation. Saguenay and North Shore were geographically relatively isolated regions while the three sub-populations located in the Gaspesian Peninsula were of different ethno-cultural background and did not often inter-marry. Individuals within these regional or ethno-cultural populations are more closely related among themselves than with individuals from other sub-populations. Individuals from the cities of Montreal and Quebec clustered together in the middle of the other regions and were also more similar to HapMap CEU and HGDP French. We also observed a similar population structure pattern using regional and ethno-cultural origins of genealogical ancestors. These origins were significantly associated to variation in allele sharing among individuals and further support that the structure observed is consistent with the known founder events.

We have previously shown that population structure could be identified in Quebec from a larger sample of genealogical data only (Bhérer et al. 2011). Here we found strikingly similar patterns of population structure with both genetic and genealogical data. Concordance of genetic and genealogical data was previously reported in an Italian population using a small number of microsatellites (Colonna et al. 2009). This concordance is expected in theory since realized allele sharing is captured by PCA on genomic data, while genealogical data provide expected sharing. However, in practice, the concordance of results from these two data sources will depend on the coverage and quality of genomic and genealogical data. The concordance of these two sources of data in our study also illustrates the relationship

between PC projections of individuals and the underlying genealogical history of the individuals' genomes, as described by McVean 2009, who showed that PC projections can be obtained from the average coalescent times between pairs of samples. When genealogies are known at least in part, in this case the part that is relevant to the observed population structure, the coalescent times are reflected in the kinship coefficients calculated from these genealogies, which we used to derive population structure from genealogical data.

As expected given the large number of European-descent founders of the French Canadian population, we did not find large differences in common SNPs allele frequency (MAF≥0.05) between our Quebec sample and HapMap CEU or HGDP French. We also did not find large allele frequency differences between regional or ethno-cultural populations of Quebec. However, the latter result is limited by the relatively small sample sizes of our regional and ethno-cultural populations. Nonetheless, we found that small allele frequency differences in common SNPs do contribute to differentiation between Quebec and the reference populations and among Quebec sub-populations. The differentiation between Quebec and the European populations was smaller than among Quebec sub-populations, where more substantial differentiation likely occurred because of the sub-founder effects that did not include as many founders as that of the entire Quebec population. Our Fst values were comparable to those reported for other founder populations (Jakkula et al. 2008; Price et al. 2009). The sub-population most differentiated from the others (Acadians with Fst of 0.006-0.008) also had the highest levels of LD and homozygosity, as indicated by longer extended ROHs. This is consistent with Acadians showing the lowest diversity among Gaspesian groups, observed at the level of uniparentally transmitted markers (Moreau et al. 2009). The Acadian sub-population of the Gaspesian Peninsula went through a first bottleneck with immigration from France to Acadia (now Nova Scotia and New Brunswick) in the first half of the 17th century and a second bottleneck with settlement to Quebec following their deportation. This sub-

population was more prone to genetic drift because of its small number of founders and relative isolation, with more out-migration (emigration) than immigration (Moreau et al. 2011b). The other regional sub-populations also showed higher homozygosity compared to the cities of Montreal and Quebec, consistent with the founder effects that led to these sub-populations. Indeed, fragmentation of the genetic pool of Quebec was already anticipated from genealogical data (Gagnon and Heyer 2001; Bhérer et al. 2011) as well as from regional partition of hereditary disorders (Scriver 2001) ascribed to local founder effects (Labuda et al. 1996; Yotova et al. 2005).

In agreement with a neutral distribution of allele frequency differences resulting from genetic drift and given our Fst estimates and sub-population sizes (Price et al. 2009), we obtained values of genomic inflation factor λ above 1 (ranging from 1.1 to 1.4) between pairs of Quebec sub-populations and between Montreal and the reference populations. These values suggest that association studies in Quebec could yield inflated false-positive rates and should take into account population structure, especially since genomic inflation factors may be higher with the larger sample sizes used in case-control studies (Devlin and Roeder 1999). Our results also suggest that carefully considering the possibility of both population stratification and cryptic relatedness is important in association studies performed in Quebec. We found levels of moderate to distant relatedness that would not be identified in an association study unless genealogical or high-density genetic data were collected, thus likely leading to inflated type I error rates due to cryptic relatedness. Based on our genomic and genealogical estimates of inbreeding, which ranged from 0.001 to 0.01 (genealogical estimate) on average depending on the region, association test statistics in studies of 500 cases and 500 controls could be inflated from 1.5 to 6 times (Devlin and Roeder 1999). Despite our limited sample size, our study clearly indicates the need to correct for potential biases due to genetic correlation present in samples from Quebec, but further studies are needed to assess the extent to

which both population stratification and cryptic relatedness impact genetic association studies.

Several methods exist to take into account population structure (Price et al. 2010). In Quebec, matching cases and controls on the region where sampling was performed may seem like a reasonable option if the ethno-cultural origin (for example Acadian and Loyalist) is also taken into account. However, methods that require genome-wide genotypic data, such as PCA correction (Price et al. 2006), structured association (Pritchard et al. 2000), or genomic control (Devlin and Roeder 1999) are more robust. Genomic control also has the advantage of correcting for cryptic relatedness, although it may not be the most powerful approach. Mixed models that can explicitly incorporate population structure and cryptic relatedness have been shown to outperform both PCA correction and genomic control (Kang et al. 2010). These models use high-density genotypic data to estimate the level of relatedness and control for it (Kang et al. 2010; Price et al. 2010; Zhang et al. 2010). More traditional mixed models from the classical polygenic theory (Boerwinkle et al. 1986; Ober et al. 2001; Lange et al. 2005) use the complete genealogy of the sample to take into account the genetic correlations among individuals. Genomic-based estimates of relatedness are a good proxy to genealogies, which are rarely known in human association studies. In Quebec however, the similarity of our conclusions from genomic and genealogical data suggest that mixed models could be implemented using either high-density genotypic or genealogical data. A combination of the two sources of data could also be valuable as they provide complementary information.

By corroborating genomic-based results by genealogical analysis, our study illustrates and confirms interpretations of recent genomic-based findings in founder and other populations. The founder population of Quebec also provides an interesting example of a population in which two mechanisms of population substructure are at play: population stratification and cryptic relatedness due to multiple, subsequent founder effects. Studying the

interplay of these two sources of bias for genetic association studies is important and the rich genealogical information available on the French Canadian population makes it an interesting model for these studies.

SUPPLEMENTARY MATERIAL

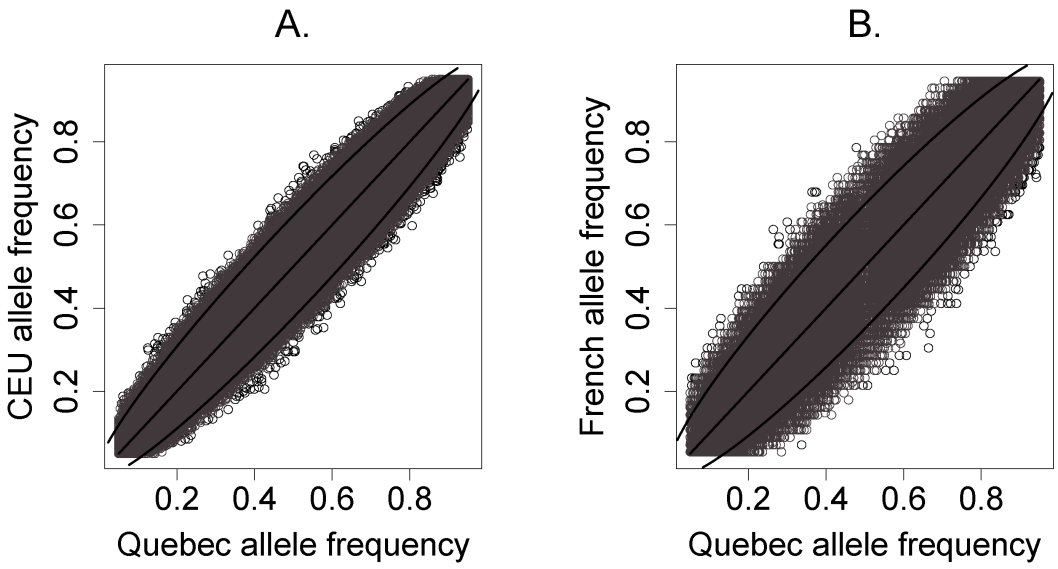

Figure S1. Comparison of ancestral allele frequency estimates between Quebec and the reference populations.

Only common SNPs with minor allele frequencies of 0.05 or more are included. Bold lines indicate no allele frequency difference between the two samples, i.e., a slope of one through the origin, with associated 95% confidence interval. **(A)** HapMap CEU. **(B)** HGDP French.

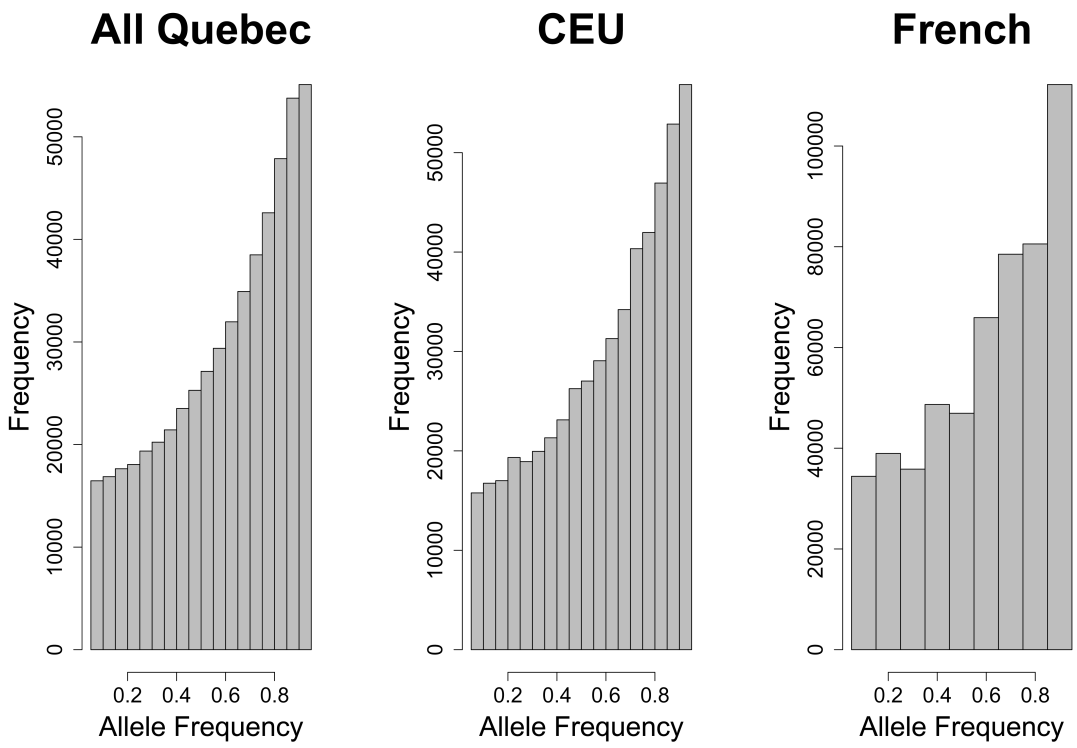

Figure S2. Distribution of ancestral allele frequency estimates in Quebec, HapMap CEU, and HGDP French.

Only common SNPs with minor allele frequencies of 0.05 or more are included.

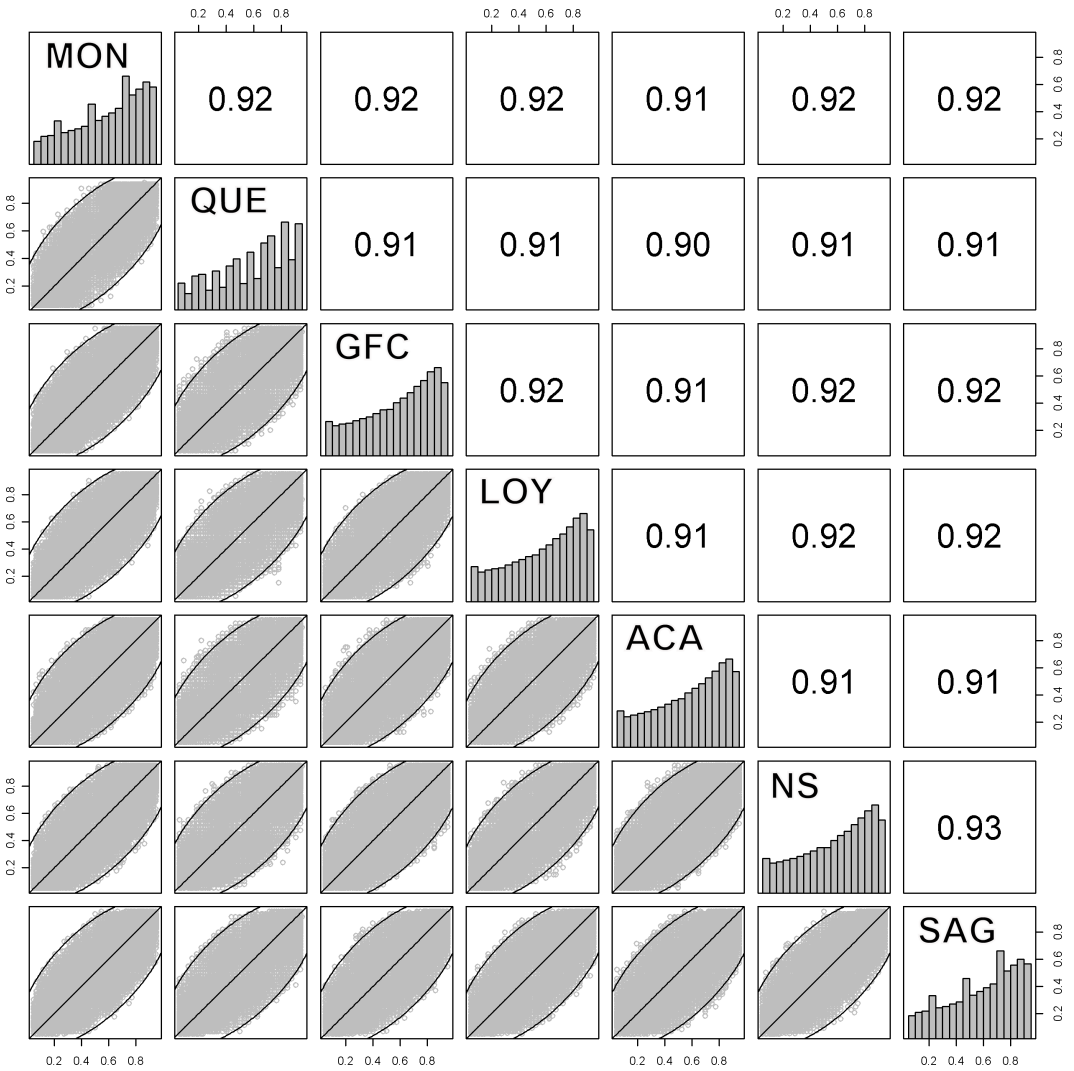

Figure S3. Comparison of ancestral allele frequency estimates among Quebec regional and ethno-cultural populations.

Only common SNPs with minor allele frequencies of 0.05 or more in at least one population are included. Diagonal shows histograms of allele frequencies. Lower diagonal shows pairwise scatter plots of allele frequencies. Bold lines indicate no allele frequency difference between the two samples, i.e., a slope of one through the origin, with associated 95% confidence interval. Upper diagonal shows Pearson's correlation coefficients. MON=Montreal, QUE=Quebec City area, GFC=Gaspesian French Canadians, LOY=Loyalists, ACA=Acadians, NS=North Shore, SAG=Saguenay.

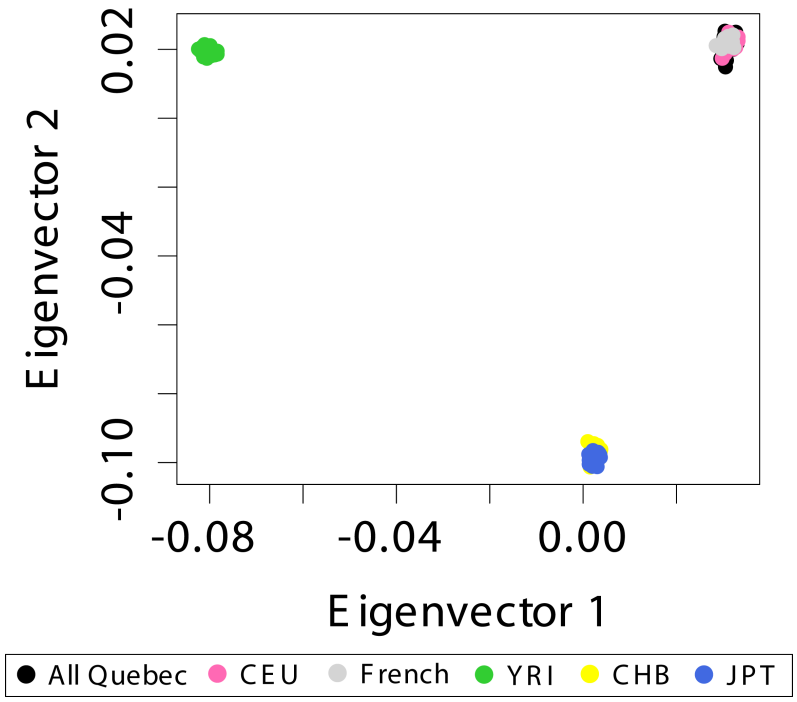

Figure S4. Plot of the first two eigenvectors from a principal components analysis of Quebec, HapMap CEU, and HGDP French populations.

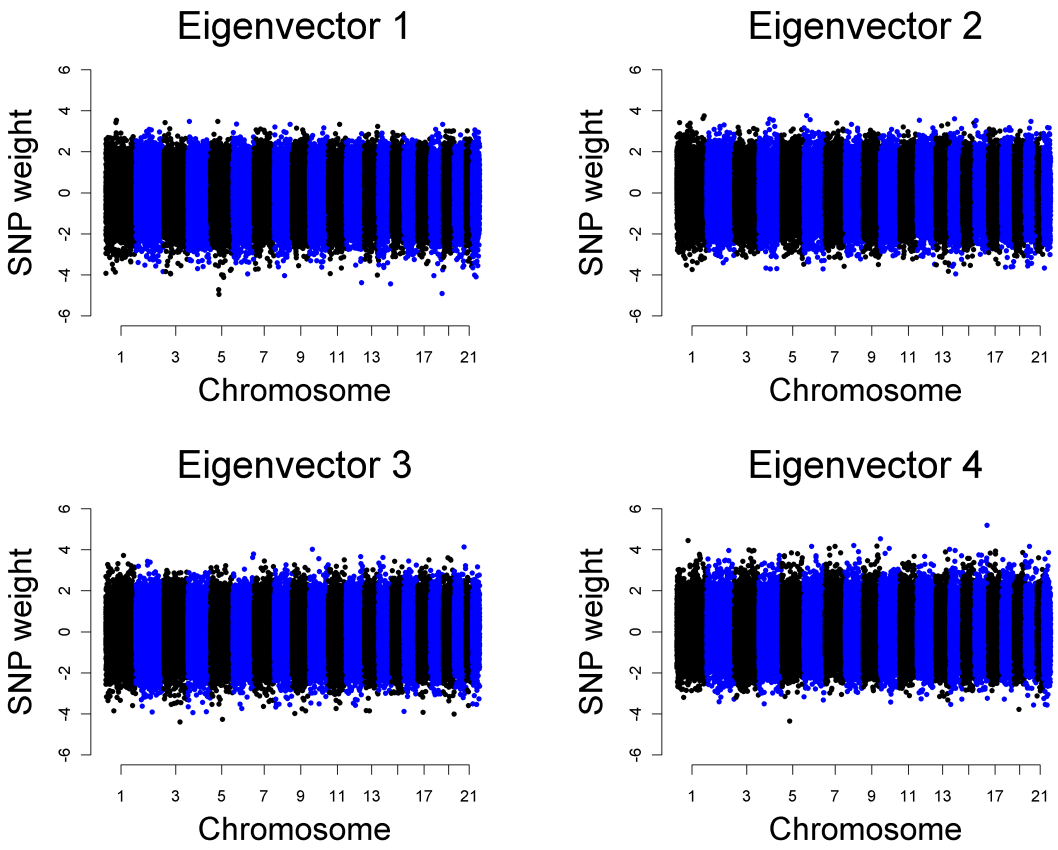

Figure S5. SNP weights for the first four principal components of the analysis including Quebec samples.

Weights are plotted for all SNPs in low linkage disequilibrium included in the principal components analysis performed with the EIGENSOFT software.

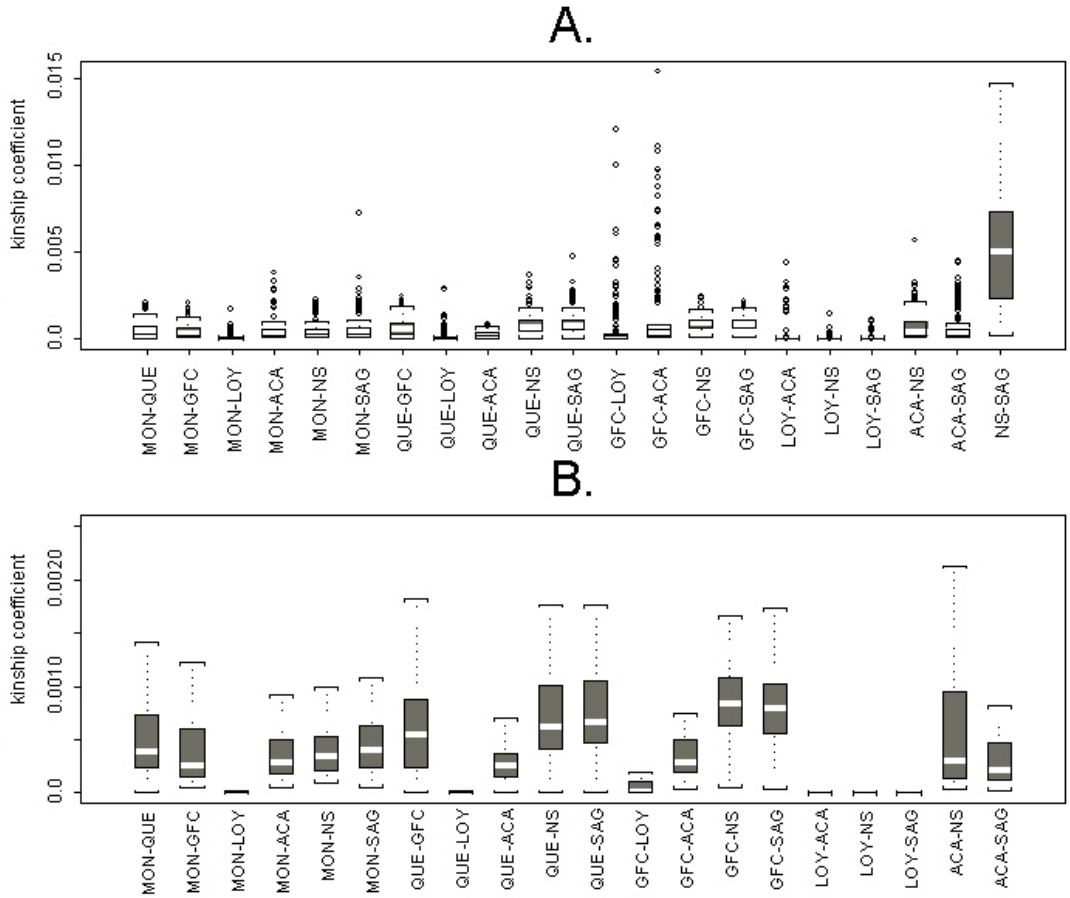

Figure S6. Kinship estimated from genealogic data.

Kinship was calculated considering all genealogical links and is shown among individuals from two sub-populations for each pair of sub-populations. Panel **(A)** shows all pairs of individuals for all pairs of populations while panel **(B)** excludes outliers and NS-SAG for clarity. MON=Montreal, QUE=Quebec City area, GFC=Gaspesian French Canadians, LOY=Loyalists, ACA=Acadians, NS=North Shore, SAG=Saguenay.

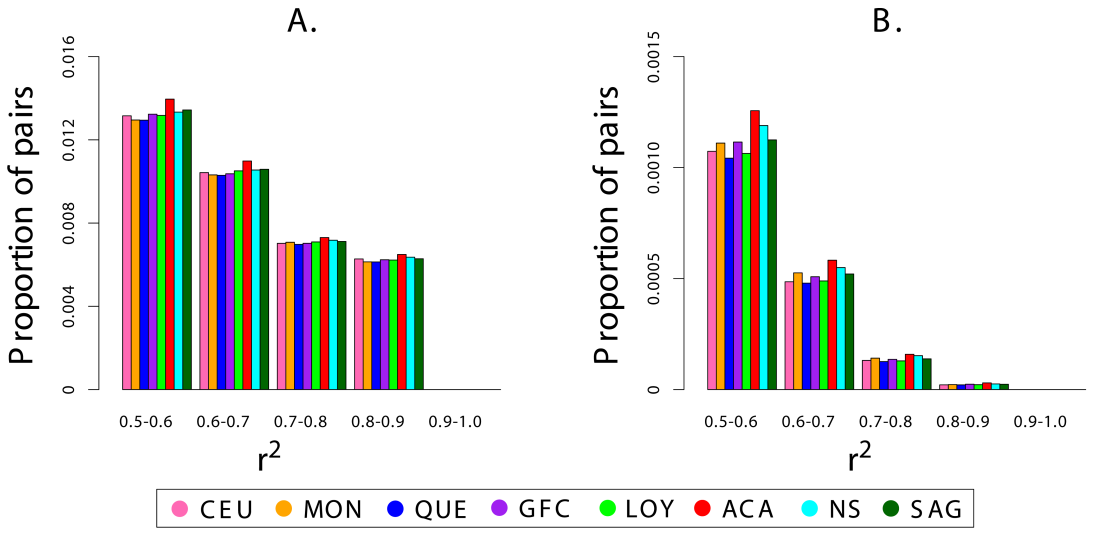

Figure S7. Proportions of pairs of SNPs at different linkage disequilibrium levels (in terms of r^2).

(A) SNPs located <200kb apart. **(B)** SNPs located >200 kb apart. MON=Montreal, QUE=Quebec City area, GFC=Gaspesian French Canadians, LOY=Loyalists, ACA=Acadians, NS=North Shore, SAG=Saguenay, CEU = HapMap CEU.

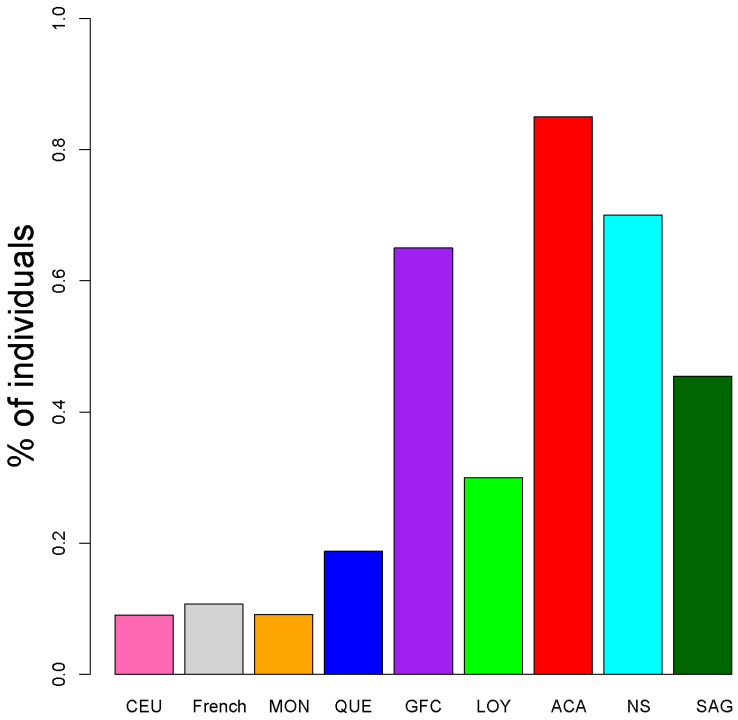

Figure S8. Percentage of individuals with at least one Run of Homozygosity (ROH) greater than 5 Mb.

MON=Montreal, QUE=Quebec City area, GFC=Gaspesian French Canadians, LOY=Loyalists, ACA=Acadians, NS=North Shore, SAG=Saguenay, CEU = HapMap CEU, French = HGDP French.

Table S1. Sample sizes and quality control details.

Sub-population	*n* sampled	*n* after quality control filters	Average genotyping rate per individual	# SNPs with HWE $p < 0.001$*
MON	22	22	0.999	75
QUE	16	16	0.994	35
GFC	22	20	0.991	80
LOY	20	20	0.993	83
ACA	21	20	0.995	67
NS	20	20	0.990	101
SAG	22	22	0.994	45

*Among common (MAF≥0.05) SNPs (in Quebec overall) with at least 90% genotypes.
MON=Montreal, QUE=Quebec City area, GFC=Gaspesian French Canadians, LOY=Loyalists, ACA=Acadians, NS=North Shore, SAG=Saguenay.

Table S2. The twenty-six regions of Quebec illustrated in Figure 1.

Number (illustrated in Figure 1)	**Region**
1	Abitibi
2	Témiscamingue
3	Outaouais
4	Laurentides
5	Rive nord-ouest de Montréal
6	Laval
7	Île de Montréal
8	Lanaudière
9	Richelieu
10	Rive-Sud de Montréal
11	Mauricie
12	Bois-Francs
13	Estrie
14	Portneuf
15	Lévis-Lotbinière
16	Environs de Québec
17	Ville de Québec
18	Côte-de-Beaupré
19	Beauce
20	Côte-du-Sud
21	Charlevoix
22	Saguenay–Lac-St-Jean
23	Côte-Nord
24	Bas-Saint-Laurent
25	Gaspésie
26	Îles de la Madeleine

Table S3. ANOVA p-values for pairwise population comparisons on the first 3 eigenvectors from EIGENSOFT principal components analysis of Quebec only.

Population 1	Population 2	Eigenvector 1	Eigenvector 2	Eigenvector 3
ACA	GFC	3.3E-11	3.7E-12	2.9E-07
ACA	NS	1.7E-15	0.73	0.07
ACA	LOY	1.4E-13	7.1E-15	2.0E-05
ACA	MON	2.1E-14	1.3E-12	0.31
ACA	QUE	8.2E-12	1.6E-11	0.65
ACA	SAG	0.0E+00	0.89	0.99
GFC	NS	3.8E-06	4.8E-11	6.8E-10
GFC	LOY	0.36	0.71	1.0E-09
GFC	MON	0.30	4.6E-03	8.1E-07
GFC	QUE	0.34	3.0E-02	8.3E-06
GFC	SAG	1.4E-07	5.0E-11	1.8E-08
NS	LOY	7.3E-07	5.2E-13	1.6E-04
NS	MON	5.2E-07	8.4E-11	2.7E-03
NS	QUE	7.9E-06	1.3E-09	3.3E-02
NS	SAG	0.96	0.86	2.9E-02
LOY	MON	0.84	4.4E-05	1.6E-06
LOY	QUE	0.79	5.9E-04	6.1E-05
LOY	SAG	3.9E-09	7.1E-13	5.6E-06
MON	QUE	0.96	0.39	0.66
MON	SAG	3.0E-09	1.8E-10	0.23
QUE	SAG	7.9E-09	2.7E-09	0.59

MON=Montreal, QUE=Quebec City area, GFC=Gaspesian French Canadians, LOY=Loyalists, ACA=Acadians, NS=North Shore, SAG=Saguenay.

Table S4. Eigenvalues and associated statistical tests from EIGENSOFT principal components analysis of Quebec only.

Number	Eigenvalue		Tracy-Widom Test		ANOVA *p*-value
	Value	% variance explained	Statistic	*p*-value	
1	1.47	1.056	94.18	1.13E-266	0
2	1.31	0.942	62.02	1.90E-143	2.22E-16
3	1.17	0.843	18.53	5.38E-25	4.44E-16
4	1.15	0.825	11.18	1.16E-12	0.19
5	1.14	0.818	8.65	3.58E-09	0.03
6	1.13	0.811	6.33	8.31E-07	0.82
7	1.11	0.799	0.70	0.07	0.62
8	1.10	0.792	-2.33	0.81	0.21
9	1.10	0.791	-2.13	0.76	0.19
10	1.10	0.789	-2.35	0.82	0.10

Table S5. Genomic inflation factors λ if cases were selected from population 1 and controls from population 2.

Population 1	Population 2	λ
MON	ACA	1.311
MON	GFC	1.130
MON	NS	1.152
MON	LOY	1.109
MON	QUE	1.159
MON	SAG	1.135
MON	HapMap CEU	1.093
MON	HGDP French	1.043
LOY	ACA	1.359
LOY	GFC	1.155
ACA	GFC	1.337
ACA	SAG	1.388

MON=Montreal, QUE=Quebec City area, GFC=Gaspesian French Canadians, LOY=Loyalists, ACA=Acadians, NS=North Shore, SAG=Saguenay.

Table S6. Percentage of pairs of SNPs, within a specified distance, that are in high linkage disequilibrium.

	CEU	MON	QUE	GFC	LOY	ACA	NS	SAG
Distance (Mb)	**$r^2 \geq 0.8$**							
0-5	13	13	13	13	13	14	13	13
5-10	8	9	9	9	9	9	9	9
10-20	6	6	6	6	6	6	6	6
20-50	3	3	3	3	3	3	3	3
50-100	1	1	1	1	1	1	1	1
100-200	0	0	0	0	0	0	0	0
200-500	0	0	0	0	0	0	0	0
500-1000	0	0	0	0	0	0	0	0
1000-5000	0	0	0	0	0	0	0	0
5000-10000	0	0	0	0	0	0	0	0
10000-15000	0	0	0	0	0	0	0	0
	$r^2 = 1$							
0-5	9	10	10	10	10	11	10	10
5-10	6	6	6	6	6	7	6	6
10-20	4	4	4	4	4	4	4	4
20-50	2	2	2	2	2	2	2	2
50-100	1	1	1	1	1	1	1	1
100-200	0	0	0	0	0	0	0	0
200-500	0	0	0	0	0	0	0	0
500-1000	0	0	0	0	0	0	0	0
1000-5000	0	0	0	0	0	0	0	0
5000-10000	0	0	0	0	0	0	0	0
10000-15000	0	0	0	0	0	0	0	0
	$D' = 1$							
0-5	81	82	81	82	82	83	82	82
5-10	72	72	72	73	73	74	73	73
10-20	64	64	64	65	65	66	65	65
20-50	54	54	54	54	54	56	55	55
50-100	46	46	46	46	46	48	47	47
100-200	41	41	41	42	41	44	43	42
200-500	38	39	39	39	39	41	40	40
500-1000	38	38	38	38	38	40	39	39
1000-5000	37	38	38	38	38	39	39	38
5000-10000	37	38	38	38	38	39	38	38
10000-15000	38	38	38	38	38	39	38	38

High linkage disequilibrium defined by $r^2 \geq 0.8$, $r^2 = 1$, or $D' = 1$.
MON=Montreal, QUE=Quebec City area, GFC=Gaspesian French Canadians, LOY=Loyalists, ACA=Acadians, NS=North Shore, SAG=Saguenay, CEU = HapMap CEU.

CHAPTER IV:

Genealogical evidence of allele frequency shuffling promoting genetic differentiation in a human founder population

Claude Bhérer, Julie G. Hussin, Marie-Hélène Roy-Gagnon, Laurent Excoffier, Hélène Vézina, Damian Labuda

Référence:

Bhérer C, Hussin JG, Roy-Gagnon MH, Excoffier L, Vézina H, Labuda D, (en préparation). Genealogical evidence of allele frequency shuffling promoting genetic differentiation in a human founder population.

CONTRIBUTION DES CO-AUTEURS

Pour cet article, ma contribution est la suivante:

- design de l'étude;
- réalisation des simulations généalogiques;
- analyse et interprétation des résultats;
- développement du programme Coalped avec LE;
- rédaction du manuscrit.

La contribution des co-auteurs est la suivante: JH a réalisé les analyses de Fst et contribué à la rédaction du manuscrit; MHRG a conseillé sur le design de l'étude dans les premières étapes de ce projet; LE a développé le programme Coalped; HV a fourni les données généalogiques et révisé le manuscrit; DL a supervisé le design de l'étude, l'analyse des résultats et il a révisé le manuscrit.

ACKNOWLEDGMENTS

We thank Youssef Idaghdour for helpful discussions. We thank BALSAC colleagues for technical assistance. We also thank Laurent Richard from CIEQ at Laval University for cartography work. LE was supported by a Swiss NSF grant No 3100A0-126074, DL and HV by the *Réseau de Médecine Génétique Appliquée* of the *Fonds de Recherche en Santé du Québec* (FRSQ), and CB was a recipient of an FRSQ studentship.

ABSTRACT

Formation of a new population by a limited number of founders may trigger a cascade of genetic changes that have been described in theory for decades. Yet, in natural habitats, the nature and extent of genetic changes resulting from founder events are difficult to grasp, especially in human populations. Here, we evaluate the genetic consequences of Quebec founding settlement and its subsequent regional expansion by simulating changes in allelic frequencies conditional on extensive genealogies. We use a sample of 2,221 contemporary French Canadians and their genealogies comprising 150,000 distinct ancestors and 8,834 distinct founders. Allele dropping simulations showed that in the whole Quebec sample, the site frequency spectrum was robust to major changes, with no deviation from equilibirum and neutral common alleles (1% or more) expected to persist in the population. Nonetheless, we observe an extensive reorganization in frequency of founders' diversity down the genealogical lineages. This reshuffling effect varied in strength across regions, leading to significant genetic differentiation. Moreover, an excess of rare alleles was observed in the Western regions, in contrast to a deficit of rare alleles observed in Eastern Quebec. A novel algorithm for coalescence simulations within the known genealogies allowed us to show a skewed distribution in the transmission probability of founders' unique alleles, with a small subset of founders having a high reproductive success. In the North-Western, North-Eastern and Eastern regions, we find that a unique allele could have increased in frequency up to 5%, thus potentially explaining the clinical founder effect of specific Mendelian disorders observed in the North-East. Our results demonstrate that regional settlement histories resulted in regional founder effects and genetic differentiation over a very short evolutionary time.

INTRODUCTION

The evolutionary history of modern humans is replete with founding events and peopling of new territories (Henn et al. 2012). Recent genomic studies have confirmed that one of the primary demographic signatures observed in human populations are ancient reduction in population size, many of which were associated with prehistoric colonization of continents (Gravel et al. 2011; Li and Durbin 2011). Historical records also document a variety of recent founder events in the last two thousand years, such as religious diasporas and European colonialism migrations that prompt creation of new populations and profoundly altered the distribution of humans throughout the globe. Colonization processes thus likely played an important role in shaping present-day patterns of human genetic variation. Yet, in humans, empirical data are lacking and most of our knowledge on the genetic consequences of founding events, i.e. the founder effect, derives from simplified population models.

In the canonical model of founder (or bottleneck) effect, a limited number of founders originating from the same source population migrate in a single pulse to establish a new population. In this setting, founders carry with them only a fraction of diversity from the source population thus leading to an inevitable loss of genetic diversity. As first demonstrated by Nei et al. (1975), this loss will be governed by the reduction in size of the population (the number of founders) and the subsequent growth rate of the population. Lower population size or growth rate will lead to a loss of more alleles, especially those that are rare. The number of alleles is expected to decline and recover more quickly then heterozygosity. Overall, changes in allele frequencies promoted by a drastic reduction in size may cause a deviation of the site frequency spectrum (i.e. the sampling distribution of allele frequency at polymorphic sites across the genome) from equilibrium expectations (Nei et al. 1975; Luikart et al. 1998; Marth et al. 2004). Neutral and even deleterious

alleles that do persist in the population may increase to high frequency. Therefore, the founder effect may be responsible for the elevated incidence of some otherwise rare monogenic diseases. Consistently, medical genetic studies conducted in founder populations since the 1960's have described the particular heritage of rare Mendelian disease.

In the absence of data from previous generations, in particular from the founding individuals, as it is the case for most human populations, the genetic consequences of founder events are typically inferred post-hoc. In humans, direct measurements of genetic changes that experienced founder populations are lacking. Studies of founder population typically infer their genetic differentiation to a given reference sample. However, such genomic inferences are challenged with the difficulty of finding an accurate reference sample of the possibly extinct source population. Genomic inference studies are also limited by a number of population genetic models assumptions that may be violated in real populations. Most importantly, the recent past is poorly approximated by Kingsman's coalescence, specifically below $log_2(N)$ generations where N is the effective population size (Wakeley et al. 2012), thus raising further challenges for genomic inference in recently founded populations.

The study of deep-rooted population genealogies offers an alternative and independent approach to estimate the genetic changes that occurred in the recent past of human populations. The genealogical tree connecting ancestors who contributed to today's population results from the complex interplay between demographic and selective processes and can be seen as the "effective" demographic history. Because it includes the segregation paths of today's genetic variation, deep-rooted genealogies may be used to simulate genetic data. To date, population genetics studies exploiting that potential remain relatively uncommon and methodologies specific to genealogical data are still underdeveloped (reviewed by Larmuseau et al. 2013).

The French Canadians of Quebec, in North-Eastern North America (Fig. 1A), are renowned as a classical example of founder population with rich demographic archives. The European colonization of Quebec started in 1608, with 8,500 founders who settled in the following 150 years under the French regime (Desjardins 2008). At the time of the British Conquest in 1760, the banks of the St-Lawrence river from Montreal to Quebec City were populated by 70,000 French Canadians (Figure S1 - Charbonneau et al. 1993; Charbonneau et al. 2000). With a continuous income of migrants but most of all a high fertility rate, the growth of the population promoted its expansion in the peripheral regions of Quebec. Today, four centuries after its foundation, the French Canadians number 6.2 million individuals. For more than 30 years, medical genetic studies of Mendelian disorders in Quebec have pointed out signals of founder effect among specific regional French Canadians populations: a markedly higher incidence of some Mendelian disorders, while some others are virtually absent (reviewed in Scriver 2001; Laberge et al. 2005b). Genealogical simulation studies have shown for a few specific regions that it is plausible that one founder deleterious gene could account for a carrier rate of 5%, a hypothesis coined the “one gene one founder” hypothesis (Heyer 1999; Tremblay et al. 2003). However, the nature and extent of resulting genetic changes are only partially tackled for most regional populations of Quebec.

In this study, we use real population genealogies of Quebec to simulate the transmission of alleles between the founders and their contemporary descendents. This approach allowed us to study the expected recent genetic changes in the whole Quebec population as well as in eight regions. We show that despite a modest impact on the site frequency spectrum, the demographic history caused an extensive reorganization of the founders’ diversity, in just a few generations, leading to significant genetic differentiation among regions. We further highlight unexpectedly high contribution of a subset of founders, allowing rise in frequency of some of their alleles such as described for Mendelian diseases in certain regions of Quebec. Doing so, we

aim to better understand the genetic consequences of the foundation and recent demographic history of the French Canadian people of Quebec.

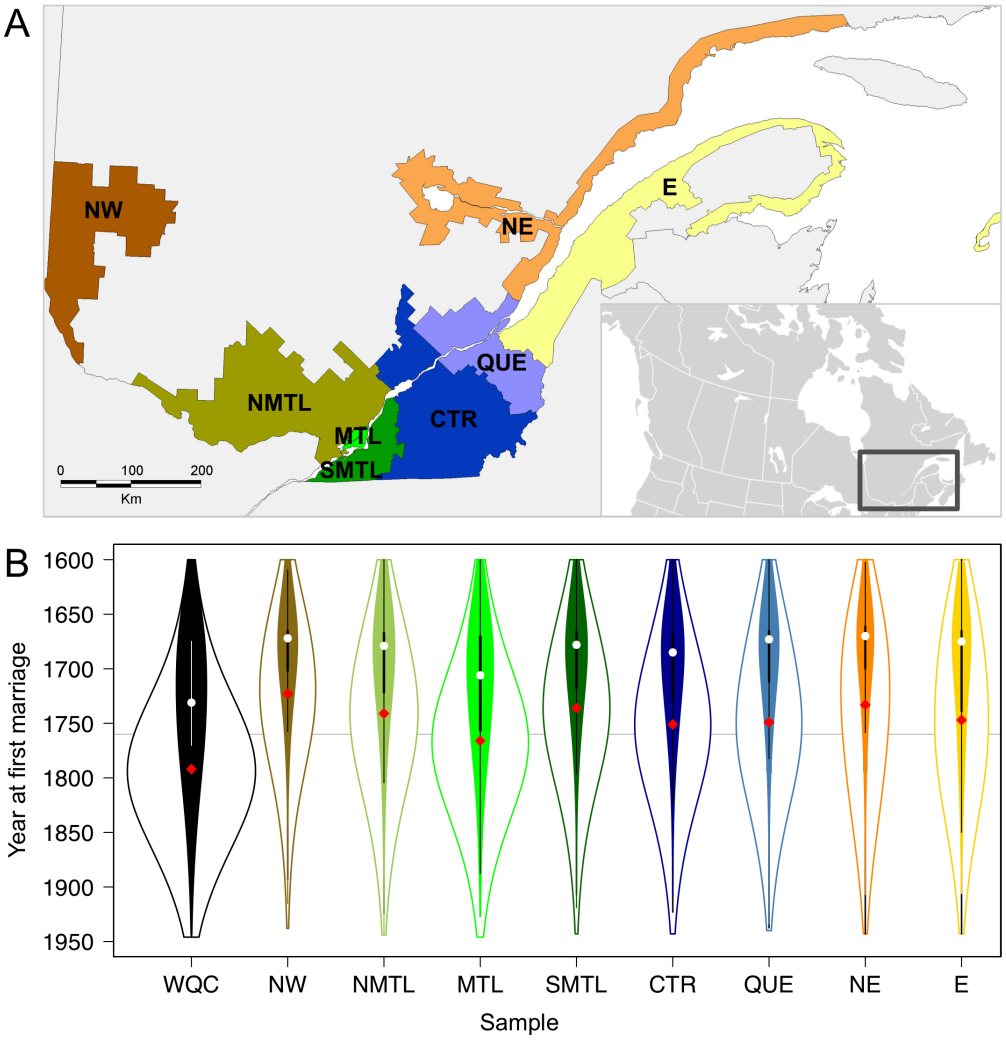

Figure 1. Map of Quebec and topology of the genealogical samples.

(A) Map of Quebec divided in eight regions. WQC: Whole Quebec, NW: North-West, NMTL: North of Montreal, MTL: Montreal City Area, SMTL: South of Montreal, CTR: Centre, QUE: Quebec City area, NE: North-East, E: East. **(B)** Topology of genealogical tree represented by violin plots of the distribution of year at first marriage of all ancestors traced in genealogies (empty violin) superimposed by violin plot of the founders (filled violin). Boxplots within founders' violins show interquartiles and white dots median of

the distribution of founders' year at first marriage, while red diamonds show ancestors' median year at first marriage. Expansion of each violin is proportional to the absolute number of ancestors and founders (Table S1). A grey line highlights the year of the British Conquest, 1760.

MATERIAL AND METHODS

Sample and Genealogical Data

The genealogical dataset is composed of a sample of n_s=2,221 individuals married in Quebec between 1945 and 1965 and their genealogies (Vézina et al. 2005b; Bhérer et al. 2011). Sampled individuals were required to have complete genealogical information for parents and grandparents and to be unrelated at the first, second or third degree of relationship with kinship coefficient below 0.125. This whole Quebec sample (WQC) provides regional representation of the Quebec population in 1956, as reported in the Canadian census (Table S1). Eight regions were defined based on historical and geographical criteria (Bhérer et al. 2011) : North-West (NW; n_s=87), North of Montreal (NMTL; n_s=242), Montreal area (MTL; n_s=722), South of Montreal (SMTL; n_s=178), Centre (CTR; n_s=348), Quebec city area (QUE; n_s=272), North-East (NE; n_s=157) and East (E; n_s=215) (Fig. 1A, Table S1).

Genealogies were reconstructed backward in time using the BALSAC population database (http://balsac.uqac.ca) and a few complementary sources (Bhérer et al. 2011). Each ascending genealogy starts with one of the 2,221 individuals in the current generation and goes back through generations of ancestors, as far as the sources would allow, essentially up to the first immigrant in Quebec or their parents. A total of 153,447 ancestors were traced back in the genealogical dataset (Bhérer et al. 2011). In addition to the genealogical connections between these individuals, we got

information on their origin, immigrant status, place and date of marriage. When the date at first marriage was unknown, we approximated it by subtracting 30 years from children's mean year at first marriage. Genealogical depth, completeness, and genetic contribution of all ancestors were calculated using the S+ package GenLib 8.4.18.

We identified the founders in the genealogical dataset using the information on the individuals' origin (Supporting methods). First, we selected as founders the last known Native American or immigrant individuals going back in time among the ascending lineages. Second, in the cases where none of these two types of founders were known in a given ascending lineage, we selected as founder the last known individual, defined as a genealogical founder (i.e. an individual who has no parental information available).

Simulations conditional on genealogies

We used the above-described genealogical data to simulate the Mendelian transmission of allele (i) forward-in-time from the founders to current generation (allele dropping) and (ii) backward-in-time from the current generation to the founders (coalescence within fixed population genealogy).

General assumptions

Neutral diallelic loci were simulated. The transmission of alleles followed Mendel's law of independent segregation, with the two parental alleles of any individual having 50:50 chances to be transmitted. We modeled Mendel's law of independent assortment by performing each simulation independently for each genetic locus, but on the same known (fixed) genealogy, assuming free recombination between loci. In this setting, each segregating site arose from a unique mutation in the founders or in their prior ancestors. We did not modeled de novo mutations in generations following the founders. The impact of false genealogical links – such as undeclared adoption and false paternity - in the Quebec genealogical dataset was considered negligible, since

BALSAC genealogies were estimated to have a false connection rate below 0.75% for both paternal and maternal lineages (Jomphe 2011), among the lowest worldwide (Anderson 2006).

Allele dropping simulations and analysis

We used the allele dropping procedure to simulate (i) the current generation site frequency spectrum and (ii) the probability distribution of allele frequency changes. Each allele dropping simulation proceeds as follow. First, assuming Hardy-Weinberg equilibrium, the founders' genotypes are randomly assigned given an arbitrarily chosen allele frequency p_f. Then, Mendelian transmission of the founders' alleles was simulated forward-in-time in the genealogical lineages. Finally, we obtained a genotype for each individual in the current generation. Our allele dropping simulation algorithm is implemented in the S+ package GenLib 8.4.18. Analysis of simulated data and drawing of figures were performed with R.

We first simulated the site frequency spectrum, defined as the distribution of the number of segregating sites S in a sample of alleles or sequences. For a sample of size n under the infinitely many sites mutation model, the expectation of S_i, the number of alleles (or mutated sites) found in exactly i copies, is given by (Watterson 1975):

$$E\left[S_i\right] = \frac{\theta}{i} \qquad \text{(eq. 1)}$$

where θ is the population mutation rate. We concentrated on the frequency of the less frequent or minor allele. The distribution of the minor allele frequency is described by the folded site frequency spectrum, where m is equal to $n/2$ or $(n-1)/2$ respectively for even or uneven sample size:

$$E\left[S_i\right] = \frac{\theta}{i} + \frac{\theta}{m-i} \qquad \text{(eq. 2)}$$

To assign the founders' genotypes, we defined an equilibrium folded site frequency spectrum in a fictive ideal population from where originated the founders. For this founders' equilibrium spectrum, we arbitrarily set founder's diversity (θ_f) to two times the total number of founders (n_f = 8,834). With this θ_f, the number of singleton alleles (i.e. an allele seen only once among the pool of all founders) equals the number of founders' alleles ($S_1 = \theta_f$ = 2 x 8,834 = 17,668). The total number of segregating sites S_{nf} simulated summed up to 179,468. Iteratively, we simulated the transmission of the minor allele of each S_{nf} segregating site with the allele dropping simulation procedure. For each the WQC sample and each of the eight regional samples, we obtained the number of alleles surviving in the current generation, which by definition is the number of segregating sites S_{ns} and we computed their folded site frequency spectrum. To calculate confidence intervals, we repeated this whole process 100 times and obtained 100 current generation site frequency spectra for each sample. For each current generation site frequency spectrum, we calculated two estimators of population mutation rate: θ_S and θ_π. Watterson's estimator based on the number of segregating sites, θ_S, is derived from eq.1 (Watterson 1975):

$$\hat{\theta}_S = \frac{S_n}{\sum_{i=1}^{n-1} \frac{1}{i}} \qquad \text{(eq. 3)}$$

The second estimator we calculated, θ_π, which represents the mean number of pairwise differences between individual sequences (Tajima 1983), was calculated as the sum of site heterozygosities :

$$\hat{\theta}_\pi = \sum_{i=1}^{n-1} S_i \bullet \frac{2i(n-i)}{n(n-1)} \qquad \text{(eq. 4)}$$

We tested deviation of current generation site frequency spectrum from equilibrium expectations by three means. First, for each sample, we computed 100 site frequency spectra to calculate the mean S_{ns} and the two-

sided confidence interval at 95% confidence level. We compared our simulated S_{ns} to equilibrium expectations of S_n (eq.1) assuming no loss of diversity (θ = 17,668) and considering sample size equal to respective current generation samples. Second, we compared this expected equilibrium frequency spectrum with the mean simulated frequency spectrum by use of a chi-square goodness-of-fit test. Third, for each sample, we calculated the mean difference between the two estimators of population diversity parameter above described: θ_π - θ_S. Discrepancy between these two estimators can involve rupture with idealized population model (Tajima 1989). Confidence intervals were built using the 100 independent simulations of site frequency spectrum.

In order to analyze the genetic differentiation among regions, we computed F_{ST} summary statistics using the allele frequency data obtained for 100 simulated site frequency spectra. For a given pair of regions, F_{ST} was calculated at each polymorphic site as the difference between the heterozygosity of the pooled samples (representing the total population) and the mean heterozygosity across regions (subpopulations) divided by the heterozygosity of the pooled samples (Nei 1973). Notably, the range of F_{ST} values, from 0 (panmictic population) to 1 (complete divergence between populations), is influenced by the frequency of the most frequent allele and the genetic diversity at the locus (Jakobsson et al. 2013). Mean F_{ST} values between each pair of regional samples were calculated by averaging over the values obtained at each simulated loci. We obtained 95% confidence intervals of pairwise mean F_{ST} by bootstrapping over samples in one frequency spectrum genotype dataset.

In a second set of analysis, we used the allele dropping procedure to simulate, for the whole Quebec sample and the regional samples, the extent of frequency changes to expect given their genealogies. We considered five initial allele frequencies among the founders, p_f : 1%, 5%, 10%, 25%, and 50%. For each of them, we randomly assigned the genotypes among the

contributing founders (i.e. founders who had descendents and so contributed to a given sample), then dropped the alleles down the genealogies, and finally calculated the allele frequencies in the current generation, p_s. We repeated this process 100,000 times, thus obtaining an empirical probability distribution of allele frequency changes.

Coalescence simulations within fixed genealogy and analysis

We developed a coalescence simulation algorithm conditional on a fixed population genealogy similar to that of Wakeley and al. (2012). In short, the simulations trace the genetic lineages backward in time within the known population genealogical tree. Our simulation algorithm is implemented in the Coalped program available upon request. One simulation starts with a randomly chosen individual belonging to the current generation and consecutively traces back the genetic lineages of his/her maternal allele and paternal allele. The genetic lineages are traced backward in time by assigning the maternal or paternal origin of genes uniformly (i.e. with a 50:50 chance) either until it reaches a founder allele (i.e. a founder allele belongs to a founder individual), or until it reaches an ancestor allele that has already been linked to a founder individual. In that case it coalesces with that lineage. To complete one simulation, this process is repeated over all individuals belonging to a sample. Because we were specifically interested in studying the fate of unique founder alleles, our program records three files (over the *r* iterations of the backward coalescent-like simulations): (i) the overall distribution of genetic contribution (i.e. the distribution of the number of descendants alleles in current generation left by all founder alleles); (ii) the distribution of genetic contribution per founder allele (i.e. the distribution of the number of current generation alleles left by each of the two founder allele) and (iii) the distribution of genetic contribution per founder (i.e. the distribution of the number of current generation alleles left by each individual in the founder generation). To validate our method, we compared founders' mean number of copies transmitted to a sample (calculated using the simulated

distributions) to theoretical expectations of the genetic contribution that were calculated exactly in the genealogy (see Bhérer et al. 2011 and Barton et al. 2011 for a fuller description of this quantity). Simulated results were almost perfectly correlated to expectations, and more so after removing five outlier founders (see text in Supplementary Results and Fig. S2).

RESULTS

A description of the Quebec genealogy

The whole Quebec genealogical sample starts with 2,221 individuals married between 1945 and 1965 and traces backward in time a total of 153,447 distinct ancestors (Fig. 1, Table 1). In total, 8,834 founders were carefully identified to include a maximum number of the actual first immigrants to Quebec (Supporting Methods and Results). Between the current generation and the founders, the genealogical lineages have an average length of 8.4 generations and attain a maximum of 17 generations, thus covering a fairly short evolutionary timescale. As we go back in time, the number of distinct ancestors traced in genealogies grows rapidly, with median year at first marriage reached in 1792 for the whole sample and then collapses because the genealogical lineages reach common ancestors and/or founders (Fig. 1B; Table S2). Indeed, although founders arrived continuously since the 17th century, 50% of founders first married between 1674 and 1771 as measured by 25%-75% interquartiles and the number of founders arrived in the 19th and 20th centuries is small.

Deviation of the whole Quebec site frequency spectrum

We first focused on the impact of demographic history on the site frequency spectrum of the whole population. Starting from the founders, we used the allele dropping procedure to simulate 100 folded site frequency spectrum (hereafter denoted frequency spectrum for simplicity). In the whole Quebec sample, the simulated frequency spectrum is not significantly different from the equilibrium spectrum (chi-square goodness-of-fit test p-value > 0.05 – Fig. 2, Table 1). The frequency spectrum in the whole Quebec therefore appears overall robust to change. However, the total number of polymorphic sites retained in the current generation is significantly lower than equilibrium expectation (S_{ns} = 145,743 95%CI: [145,545; 145,919] vs. $S_{equilibrium}$=158,587)

(Table 1). This difference is mainly explained by a reduction in the number of alleles with frequency below p_s = 5/4,442 = 0.1% (Fig. 2). Consistently, the difference between θ_π and θ_S, two estimators of population diversity parameter based on the frequency spectrum, indicates a significant deficit of rare alleles compared to equilibrium (Fig. 2, Table 1).

Table 1. Summary statistics of current generation SFS.

			S_{ns}				θ_π - θ_S	
	n_s	n_f	Mean	[95% CI]	$S_{equilibrium}$[a]	p-value[b]	Mean	[95% CI]
WQC	2,221	8,834	145,744	[145,545; 145,919]	158,587	ns	394,5	[374; 417]
NW	87	3,761	96,872	[96,656; 97,078]	101,298	<0.001	-252,2	[-282; -211]
NMTL	242	4,525	113,761	[113,600; 113,974]	119,405	ns	-194,7	[-222; -167]
MTL	722	6,785	131,434	[131,194; 131,670]	138,730	ns	-102,4	[-125; -76]
SMTL	178	4,518	109,204	[108,987; 109,392]	113,972	ns	-288,4	[-319; -251]
CTR	348	4,921	119,397	[119,178; 119,577]	125,829	ns	-131,1	[-156; -101]
QUE	272	3,775	113,903	[113,722; 114,113]	121,472	<0.001	61,3	[22; 96]
NE	157	2,717	100,211	[99,971; 100,415]	111,750	<0.001	695,5	[657; 736]
E	215	3,328	109,745	[109,556; 109,975]	117,313	<0.001	85,5	[53; 116]

n_s : Sample size; n_f : number of founders; S_{ns} : number of polymorphic sites in the current generation samples; θ_{ns} = population mutation rate based on heterozygosity; θ_π = population mutation rate based on heterozygosity; θ_π = population mutation rate based on the number of polymorphic sites

[a] S: the number of polymorphic sites at equilibrium was calculated in each sample assuming 2*n_s chromosomes and no loss of diversity since foundation thus θ = 2 x 8,834.

[b] p-values from chi-square goodness-of-fit test between current generation SFS and equilibrium SFS.

Diversifying effects on regional frequency spectra

To study the impact of the specific regional demographic histories on their frequency spectra, we next studied eight regions of Quebec (Fig. 1). The stronger changes between the founders and current generation were observed in the North-East region of Quebec, including a significant difference between simulated and equilibrium frequency spectra (p-value < 0.001 – Table 1) and a significant deficit in rare alleles compared to equilibrium expectations (Table 1, Fig. 2). Significant deviations from equilibrium frequency spectra were also observed for the Quebec City and East region, with slight differences between θ_π and θ_S indicating deficit of rare alleles. These genetic signatures are typically expected under scenarios of

severe reduction in size, and suggest that the regional genealogical histories of the North-East, East and Quebec City regions – all three located in the eastern part of the Quebec territory – may have driven regional founder effects.

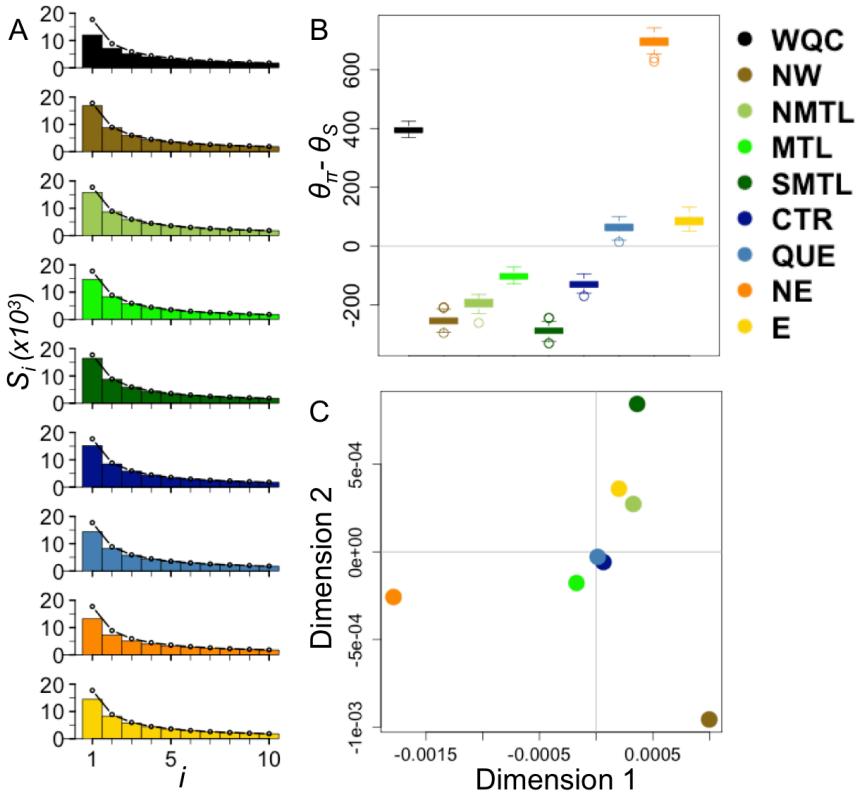

Figure 2. Divergence of whole Quebec and regional frequency spectra from equilibrium.

(A) Histograms of simulated frequency spectra obtained with allele dropping simulations of S_{nf} alleles randomly assigned among the n_f = 8,834 founders assuming θ_f = 2 x n_f. Overlay curves show expected equilibrium frequency spectra assuming no lost of diversity and sample size equal to respective samples. Histograms show the number of polymorphic sites S_i in each frequency classes for i <= 10 calculated as the mean over 100 frequency spectra obtained each with allele dropping simulations nearly 180,000 diallelic

sites. (**B)** Test of equilibrium based on comparison of two estimators of population mutation rate computed. For each sample, boxplot shows the difference between θ_π and θ_s computed over the 100 frequency spectra obtained by allele dropping simulations. **(C)** Multidimensional scaling plot of mean F_{ST} between pairwise regions, averaged over more than 10 millions retained polymorphic sites simulated with the 100 frequency spectra.

In contrast, for regions located west of Quebec City, we observed a diametrically opposite effect: a significant excess in rare alleles based on θ_π and θ_S differences, and for the North-West a significant deviation between simulated and equilibrium frequency spectra (Table 1, Fig. 2). Such an excess of rare alleles from mutation-drift equilibrium expectations is typically interpreted as a signature of demographic expansion, which can cause an accumulation of new mutations (Tajima 1989; Keinan and Clark 2012). However, this interpretation is inapplicable here because in our simulations new alleles entered the population carried by founders and not by *de novo* mutational events. Therefore, the excess of rare alleles in the western regions potentially reflects an enrichment of incoming immigrant founders, introducing rare alleles in larger number than those eroded by drift. This is supported by a larger number of founders in the western regions than in the eastern regions (Fig. 1 – Bhérer et al., 2011).

F_{ST} analysis of the simulated data revealed significant genetic differentiation between pairwise regional populations, with values ranging from 0.0005 (Montreal and Centre) to 0.003 (North-West and North-East) (Table 2). The Montreal region appears the less differentiated region with F_{ST} values below 0.001 with all other regions. Multidimensional scaling analysis of mean F_{ST} highlights three clearly differentiated regions, the North-East, the North-West and, unexpectedly, the South of Montreal region, which is located in the core agricultural territories of the St-Lawrence Valley (Fig. 2). The North-East region is clearly the most differentiated region, with the highest mean F_{ST} values in pairwise comparisons with other regions (most values near or above

0.002 Table 2), as well as the highest maximum per site F_{ST} values reaching up to 0.0768 (North-East and North-West) (Table S3). In general, the allele showing the highest F_{ST} values in the North-East comparisons showed a rise in frequency in that region while remaining inexistent or at lower frequency in the other region, except between NE and NW where fluctuations are seen in both regions (Fig. S3 and S4). Overall, these results indicate that despite their common genealogical origin, the regional populations of Quebec have significantly diverged from each other between the founder and current generations, albeit to different extent.

Table 2. Mean FST values per sites between pairs of current generation samples.

	NW	NMTL	MTL	SMTL	CTR	QUE	NE	E
NW								
NMTL	0,0016							
MTL	0,0006	0,0006						
SMTL	0,0020	0,0013	0,0006					
CTR	0,0012	0,0010	0,0005	0,0010				
QUE	0,0015	0,0012	0,0006	0,0013	0,0009			
NE	0,0029	0,0023	0,0010	0,0025	0,0018	0,0019		
E	0,0018	0,0014	0,0007	0,0015	0,0011	0,0013	0,0023	

Mean F_{ST} values obtained by averaging over more than 10 millions simulated diallelic sites between each pair of current generation samples (Table S3). All pairwise comparisons were significant at the 5% level (see 95%CI in Table S3).

Frequency reshuffling of founders' diversity down genealogies

To investigate the frequency changes that must underlie the regional genetic structure, we next focused on specific initial allele frequencies among the contributing founders, p_f (1%, 5%, 10%, 25% and 50%). For each initial founders' frequency and for the whole Quebec and the eight regions, we obtained an empirical probability distribution of allele frequencies in the current generation with allele dropping simulations, which represents the extent of frequency changes due to random sampling effects constrained by

the genealogical structure. The frequency distributions are illustrated in boxplots in Figure 3 (see Fig. S6 for p_f 25% and 50%).

The amplitude of frequency changes, evaluated using coefficients of variation, decreased exponentially as a function of the initial founders' frequency (Table 3 and S4, Fig. S5). For example, in the whole Quebec, coefficient of variation ranged from 26% for alleles introduced at 1% among the founders, down to 2.6% for alleles introduced at 50% (Table 3 and S4). The exponential decay of the extent of frequency changes emphasizes how the effects of random segregation coupled with genealogical structure affects strongly rare alleles and becomes less important for common alleles. Frequency distributions tend to be skewed towards lower values, meaning that a slightly larger fraction of alleles decreased in frequency, whereas the remaining alleles increased in frequency, with a long tail representing a small fraction of alleles reaching very high frequencies (Table 3 and S4, Fig. 3 and S6). This skew in allele frequency distributions tends to be reduced with increasing initial founders frequencies. However, all frequency distributions were significantly different from normal distribution (Kolmogorov-Smirnov test p-value < 0.001), even for 50% initial frequency among founders.

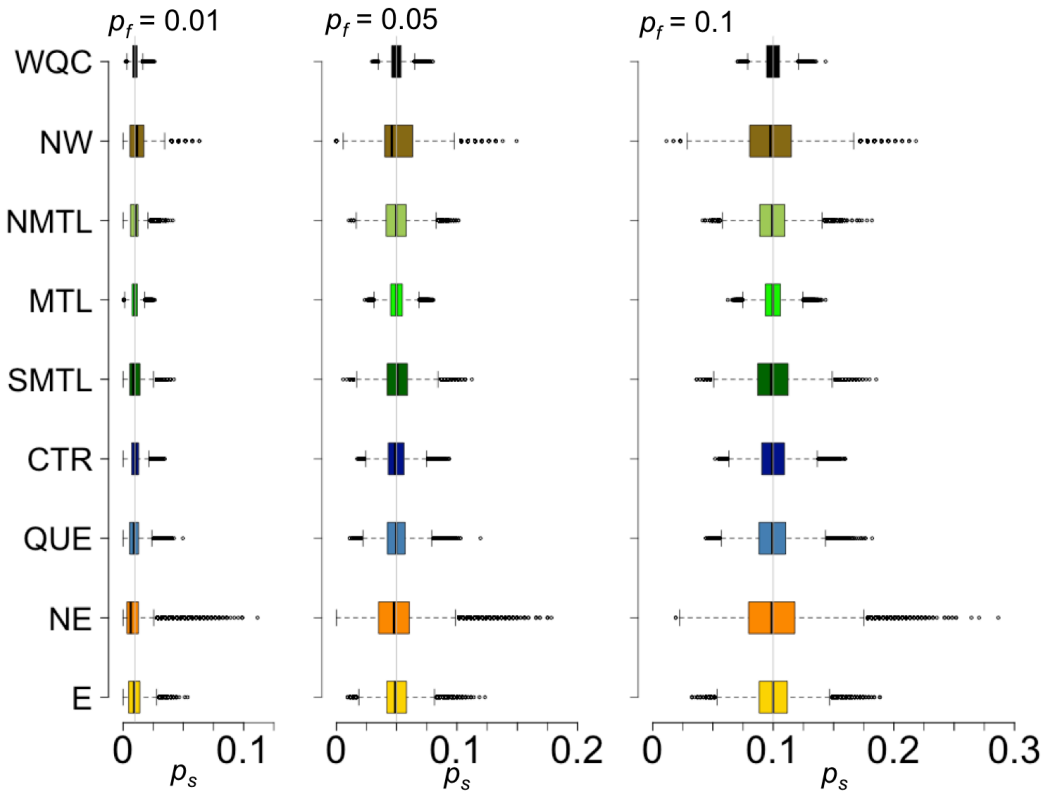

Figure 3. Allele frequency changes between founders and current generation.

For each sample, boxplot shows the probability distribution of allele frequency in the current generation (p_s) given the initial allele frequency among the founders (p_f) that was obtained with in 100,000 iterations of allele dropping simulations. We simulated p_f =0.01 (left panel), p_f =0.05 (middle panel) and p_f =0.1 (right panel). Figure S5 in Supplementary Results shows p_f =0.25 and p_f =0.5.

The extent of frequency changes was significantly different between regional populations in most instances (Fig. 3 and S5, Table S5). The North-East and North-West regions displayed respectively the first and second largest coefficient of variation (Table 3 and S6). Moreover, the variances of their frequency distribution were significantly different from other regional

populations in most comparisons (Kruskall-Walis *p*-value <0.05; Table S5). Conversely, the Montreal and Centre regions displayed the first and second lowest coefficient of variation (Table 3 and S4), with variance significantly different from other regions in most pairwise comparisons (Kruskall-Walis *p*-value <0.05; Table S5). These results demonstrate that within a frequency spectrum that has not changed much (as described above), the genetic diversity introduced by founders is reshuffled in frequency down the genealogical lineages, and so even in a very large population tree such as the whole Quebec genealogical dataset connecting over 150,000 ancestors.

In the whole Quebec, founder alleles introduced at a frequency of 1% or more had a zero probability of loss (Table 3 and S4). Alleles introduced at 5% had zero loss probability in all regions but the North-East and the North-West. Most notably, these results indicate that neutral common genetic variation introduced by the founders is expected to persist in the whole Quebec population. Increase in frequency is depicted in boxplots as the long tail to the right of the frequency distribution (Figure 3). This long tail violates expectations of Wright-Fisher ideal population undergoing pure drift, and reflects the genetic signature of demographic processes acting in natural populations such as Quebec. In the whole Quebec and the eight regions, a substantial fraction of alleles introduced at p_f=1% will double or more in frequency (Table 3). A remarkable increase is observed in the North-East and the North-West, where respectively 11% and 10% of alleles introduced at 1% will at least double in frequency, reaching a maximum frequency of 11.2% and 6.3%. In these two regions, even alleles introduced at 5% and 10% frequency among founders can double or more in frequency. Because the transmission of alleles was modeled in Mendelian ratio, this pattern of increase cannot reflect preferential transmission due to natural selection, but only the effects due to the genealogical history of the population.

Table 3. Allele frequency changes between founders and current generation samples.

	WQC	NW	NMTL	MTL	SMTL	CTR	QUE	NE	E
					p_f = 0.01				
Coefficient of variation (x10^2)	25.9	79.0	50.1	31.6	55.8	43.5	52.6	95.6	59.6
P(loss)	0	18.9	1.3	0	3.5	0.2	1.1	13.1	3.0
P(p_s<p_f)	54.8	49.2	48.6	53.9	53.3	47.9	56.3	67.4	59.1
P(p_s>=$2p_f$)	0.2	10.6	4.0	0.4	3.6	2.6	5.1	11.2	6.4
Max Increase	2.6	6.3	4.2	2.6	4.2	3.5	5.0	11.2	5.4
					p_f = 0.05				
Coefficient of variation (x10^2)	11.3	34.5	21.9	13.8	24.5	19.0	23.0	41.8	26.0
P(loss)	0	0.019	0	0	0	0	0	0.004	0
P(p_s<p_f)	51.2	50.2	54.0	52.9	49.4	50.1	54.0	54.9	52.6
P(p_s>=$2p_f$)	0	0.46	0.002	0	0.019	0	0.005	2.2	0.061
Max Increase	1.6	3.0	2.0	1.6	2.3	1.9	2.4	3.6	2.5
					p_f = 0.1				
Coefficient of variation (x10^2)	7.8	23.8	15.0	9.5	16.8	13.0	15.8	28.8	17.9
P(loss)	0	0	0	0	0	0	0	0	0
P(p_s<p_f)	50.8	52.7	51.8	51.3	50.3	50.9	51.8	54.4	49.2
P(p_s>=$2p_f$)	0	0.013	0	0	0	0	0	0.23	0
Max Increase	1.4	2.2	1.8	1.4	1.9	1.6	1.8	2.9	1.9

p_s = Allele frequency in the current generation sample; p_f = Allele frequency among the founders; P(loss) : Probability of loss; Max Increase : maximum frequency observed in the current generation divided by the frequency among founders, p_f

Transmission fate of unique founders' alleles

We examined the Mendelian transmission of every unique allele carried by the founders along the genealogical lineages down to their descendents in the current generation. To track the fate of unique founders' alleles, we developed an algorithm of coalescence simulations within the fixed population genealogy that models Mendelian segregation backward-in-time conditional on the genealogical structure (Material and Methods). With 100,000 iterations of these simulations, we generated for each founder the distribution of the

number of copies of each of its alleles that were transmitted to a given sample, conditional on the genealogies.

Figure 4 presents each founder's distribution of the number of copies of one of its alleles that was transmitted to the whole Quebec sample, which is a measure of reproductive success (Figure S7 shows the distributions partitioned by classes of maximum number of copies). We observed a high variance in these distributions among the founders (Fig. S8). The vast majority of founders contribute a small number of copies, with 90% of founders contributing less than 10 copies, and 26% contributing zero or one allele copy. However, there is a small fraction of founders who contributed for a much larger number of copies: 8.5% of founders reach a maximum of copies transmitted between 10 and 20 and 1.5% between 20 copies to 57 copies (Figs. 4 and S7). For 94% of the founders, the probability of loss is higher than the probability of survival (Fig. S8). Then as the usual outcome is loss rather than survival, the majority of unique founders' alleles were not transmitted to the sample of individuals drawn from the whole Quebec population. These results mean that in a sample of today's French Canadians, at a randomly chosen genetic locus (or at genome-wide average), we expect the total number of distinct alleles to be lower than expected under equilibrium, because most unique founder alleles are lost.

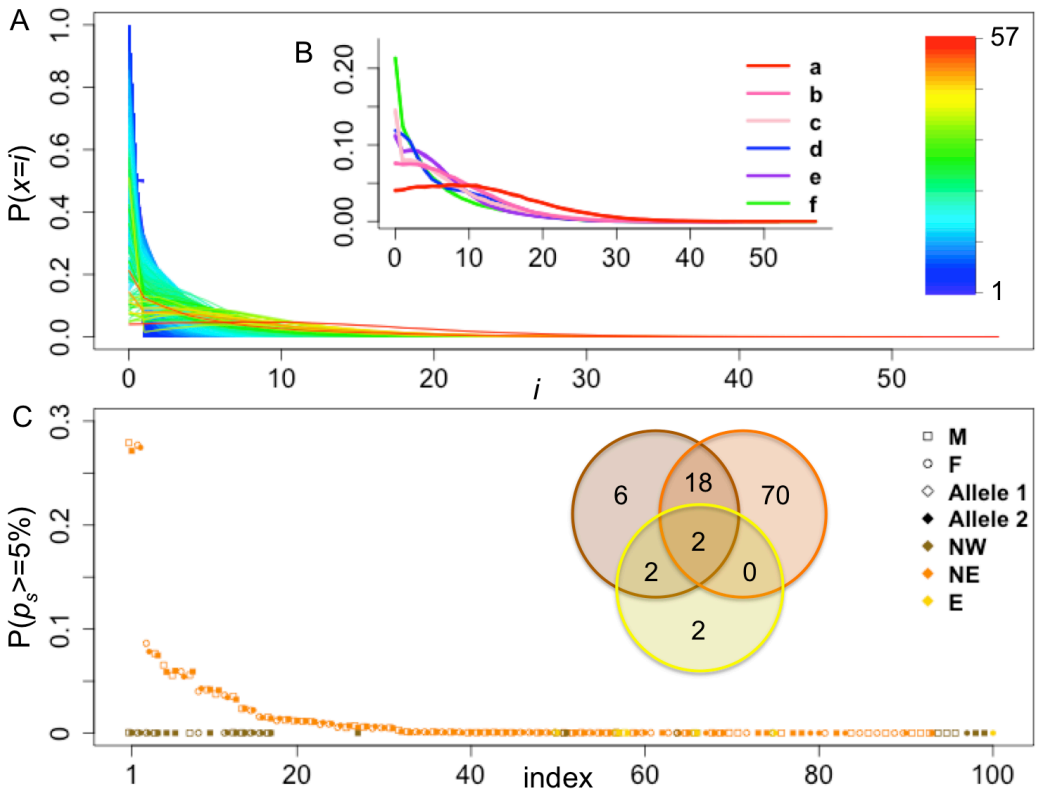

Figure 4. Fate of unique founder allele.

(A) Distribution of number of copies left by a single founder allele in the whole Quebec sample. Each curve represents a different founder. The probability *P(x=i)* is calculated as the count of simulations where we observed exactly *i* copies divided by total number of coalescence simulations (100,000 iterations). Each founder' curve is colored according to the maximum number of copies contributed following color gradient ranging from minimum = 1 copy to maximum = 57 copies. **(B)** The 11 top-contributing founders' distribution of number of copies left by one of their alleles in the whole Quebec sample. The 6 top-contributing couples were labeled from a to f in decreasing rank of genetic contribution. **(C)** The 100 founders with a non-zero probability to reach a carrier frequency of 5% (*$P(p_s>=5\%)$*). Founders are ordered by index of decreasing probability to reach such frequency in the North-East region. Each of the 100 founders are represented by their gender (square: Male; circle: Female) and by two points representing their two alleles (empty: allele

1; filled: allele 2). Venn diagram shows the number of founders shared between the North-West, the North-East and the East.

The observed differences in the distribution of number of copies among founders highlight the effect of variance in reproductive success among founders in shuffling the frequency of alleles. In fact, under our model assumptions (equilibrium founder population, no *de novo* mutations), the changes in allele frequencies are the result of randomness in the reproduction processes that are recorded in the population genealogy. Therefore, results obtained for founders' alleles introduced in single copies can be generalized to founders' alleles introduced at any frequency. As the founders' distribution of number of copies transmitted to the whole Quebec are in majority skewed towards lower values, the proportion of the number of alleles which decreased in frequency is expected to be higher than the proportion which increased, but when allele do increase, they should on average increase to higher frequency. This is how the observed variance in reproductive success among founders explains the skewed distribution in frequency changes observed in the previous section.

In the whole Quebec, the maximum increase in frequency observed for a founder allele was 57 copies among the 2,221 individuals, which translate to 1.28% allele frequency or 2.56% carrier frequency. Our result thus show that at some loci, a founder allele that was initially unique may now be found among 1 out of 50 French Canadian individuals in Quebec. Here we modeled neutral alleles, but as it has been shown that the same frequency fluctuations will hold for a recessive deleterious mutation because homozygotes are rare (Heyer, 1999), this results also means that 1:50 is the maximum carrier frequency of a recessive deleterious mutation inherited identical-by-descent from a single common founder that we can expect in the whole Quebec population.

Table 4. Unique founders' allele expected to reach 5% carrier frequency.

		Carrier frequency = 5%			Founders reaching carrier frequency >= 5%		
	n_s	Nb. carriers[a]	p_s[b] (%)	Overall probability[c]	Nb.	Nb. alleles	Nb. super founders
WQC	2,221	112	2.52	0	0	0	0
NW	87	5	2.87	1,16E-06	28	42	10
NMTL	242	13	2.69	0	0	0	0
MTL	722	37	2.56	0	0	0	0
SMTL	178	9	2.53	0	0	0	0
CTR	348	18	2.59	0	0	0	0
QUE	272	14	2.57	0	0	0	0
NE	157	8	2.55	4,97E-04	90	155	11
E	215	11	2.56	4,06E-07	6	11	0

n_s = Sample size;
[a] Number of current generation individuals simustaneously carriers required to reach 5% carrier frequency.
[b] Samples allele frequency (p_s) corresponding to number of carriers, assuming that homozygotes are rare events.
[c] Overall probability was calculated as the number of founders' alleles reaching target allele frequency or more over the total number of simulated alleles.

In some regions of Quebec, clinical genetic studies have estimated that some Mendelian autosomal recessive diseases have a carrier frequency of 5% (Laberge et al. 2005b). Therefore, we next tested the hypothesis that a unique founder mutation could reach a carrier frequency of 5% among regional populations. We estimated the probability for each founder that one of its unique alleles reach a carrier frequency of 5% or more among the regional samples. Table 4 shows the number of founder allele copies corresponding to an allele frequency of 2.5%, required to reach a carrier frequency of 5%. For five regions (North of Montreal, Montreal, South of Montreal, Centre and Quebec), the overall probability of increase to 5% carrier frequency is null (Table 4). In contrast, in three regions - the North-West, the North-East and East - our simulations demonstrate such increase possible for a total of 100 distinct founders. For these 100 founders, Figure 4C illustrates the probability of reaching a 5% carrier frequency within the

three regions and the Venn diagram shows the sharing of these founders between the three regions. The highest probabilities were found in the North-East region, where two founders, forming a couple, stand out from the others. That couple reached a probability of 28% that at a randomly chosen locus, one of their alleles was found in 1 individual out of 20 in the North-East region. Another way to see this is that 28% of their genome can be expected to reach a frequency of 5% or more in that region. That couple's unique alleles could also reach 5% carrier frequency in the North-West region, but with a much lower probability. In general, the founders whose alleles reach 5% carrier frequency with a non-zero probability in the North-West or in the East also have a non-zero probability in either one or the two other regions.

Top-contributing founders

Figure 4B focuses on the distribution of the 11 top-contributing founders, which we defined as the ones whose unique alleles reach a frequency of at least 1% in the whole Quebec. These top-contributing founders have the highest probability to have passed on their alleles successfully and were the most successful founders in transmitting their genome to the current population. Precisely, simulation results showed that the top-contributing founders transmitted most of their genome to current generation, in total between 79% and 96%. Intriguingly, visual inspection showed an unusual shape of the distribution of the number of copies of some top-contributing founders. This pattern is in striking contrast with the shape of the distribution of the number of copies found for the majority of founders, in which most alleles are lost and a fraction only survives, as expected in a few tens of generations for a Wright-Fisher population (Barton, 2011).

According to BALSAC database, the top-contributing founders comprise five monogamous couples of immigrant founders and one couple formed by one of their daughter - born in Quebec – and an immigrant founder (Table S6). These six couples were ancestors of 48% to 90% of individuals in each region

by at least one genealogical lineage (Table S6). Seven out of 11 top-contributing founders (64%) came from the small Perche region in France from where came 92 (1%) of the 8,834 founders. Sampling six founder couples at random and finding at least seven Perche founders is very unlikely in our data (resampling test p-value < $1x10^{-4}$; Supplementary methods), suggesting this excess of Perche origin among top-contributing founders did not happened by chance alone. The top-founders were all living in Quebec City, having been married there or having their children married there (results not shown). The top-contributing founders were married within 42 years of each other, between 1615 and 1657 (Table S6), which is again unlikely due to chance alone (p-value < $2.26x10^{-2}$). To exclude the effect of the period of arrival – since the first founders have an advantage over latter ones under equal reproductive success (e.g. Labuda et al. 1996; Bhérer et al. 2011) - we restricted our resampling test to the 549 founders married in the same time range than the top-contributing founders and obtained a significant excess of Perche origin (p-value = $2.4x10^{-3}$). This excess of Perche origin among the top-contributing founders is in accordance with the documented high contribution of the Perche founders (De Braekeleer and Dao 1994). Overall, our results show that the top-contributing founders do not appear as a random group of founders that would be expected under equal reproductive success assumptions.

DISCUSSION

In this paper, we studied the genetic changes expected to have happened within the course of history of a young founder population as recorded in its extensive population genealogy. Using different simulation procedures, we followed the allele frequencies changes since the foundation and modeled random effects due to both the sampling of the founders and Mendelian segregation in the subsequent generations. These effects were previously studied in theoretical models of reduction of population size (Nei et al. 1975; De Braekeleer and Dao 1994; Luikart et al. 1998; Marth et al. 2004). Here we studied the impact of the natural and inherently more complex genealogical history that results from the action of both demographic and selective processes. This allowed us not only to verify theoretical predictions of the founder effect but, most importantly, to reveal some unexpected genetic consequences of the natural genealogical history of a young founder population.

For the Quebec and the eight regions studied, our results show that although the frequency spectrum was overall robust to major deviations, the founders' diversity was fully reorganized down the genealogical lineages. This effect, that we coined reshuffling of founders' diversity, acted with variable intensity across regions and leaded to significant genetic differentiation. Furthermore, the distribution of frequency changes between the founders and current population is skewed, with a higher number of alleles decreasing in frequency, but the remaining fraction of alleles increasing in average to much higher frequencies. We have shown how this skewed reshuffling effect relates to skewed probability of gene transmissions among the founders. Importantly, we underlined how the high reproductive success of a small subset of founders can lead to dramatic increase in frequency of alleles introduced in single copies, and may thus explain the clinical founder effects observed in

Quebec. Below we discuss in more details these findings and their implications.

Our results showed that since the birth of the population, the whole Quebec frequency spectrum should not have been dramatically altered. Our simulations suggest that common genetic variation most likely persisted in the whole population, as we have found zero loss of allele at initial frequency of 1% or more in the whole Quebec sample, assuming no purifying selection. Under these predictions, a nearly perfect genetic sharing of common genetic variation between the French Canadians and the French people, who contributed for 90% of Quebec gene pool (Vézina et al. 2005b; Bhérer et al. 2011), is expected. Our results are thus in agreement with the two recent genomic surveys comparing French Canadians and French samples; one which reported a 98% correlation in frequencies for alleles with MAF> 5% (Roy-Gagnon et al. 2011) and the other 99% sharing of common variation (MAF>20%) (Casals et al. 2013). The persistence of common genetic variation might also explain why genetic epidemiology studies conducted with French Canadian samples did not notice any major discrepancy in the common genetic determinants associated to complex disease. This also suggests that the French Canadian population can be used for replication studies of genome-wide association studies performed with European samples. In contrast, our results indicate that rare genetic variation has been subject to extensive frequency changes in the past ~10 generations average between the founders and current population. In the whole Quebec population, our simulations show that if all founders came from the same source equilibrium population, the genealogical history would likely lead to a deficit of rare alleles (MAF<1%), a signature of reduction in population size that may be attributable to substructure and/or founder effect. Our results suggest that this deficit of rare alleles was less pronounced in the whole population than in the regions, the latter showing much larger frequency fluctuations, including greater loss of rare alleles. This is consistent with a fairly large total founding size, estimated at 8,500 founders settlers in the 17th

and 18th centuries, and to a much smaller number of founders in some regions (Moreau et al. 2007; Bhérer et al. 2011).

The variable intensity of reshuffling effect across regions, highlighted by different rates of frequency changes, shows how the different genealogical histories of the regions impacted differently their genomes, giving further support to the hypothesis of the regionalization of the founder effect (Gagnon and Heyer 2001; Scriver 2001; Moreau et al. 2007; Bhérer et al. 2011). Our results confirmed that the frequency changes that occurred since Quebec foundation led to significant genetic differentiation among regional populations. F_{ST} values based on simulated data are of the same order of magnitude as that observed from genome wide SNP data in Quebec (Roy-Gagnon et al. 2011) and in the founder population of Finland (Jakkula et al. 2008). The quick reorganization of the founding diversity in a few generations of descendents who colonized and expanded on a new territory can be considered as a diversifying counterpart to the founder effect. This reshuffling effect goes against the widespread idea of homogeneity of young founder populations. Substructure within a founding population is compatible with models of range expansion where very net frequency clines can be observed between close populations (reviewed in Excoffier and Ray 2008). Moreover, the opposite signatures of regional genealogical histories on their frequency spectra are incompatible with the expectations derived from simpler models of founder effect that do not take into account the temporal and spatial dynamics of colonization processes. Indeed, a deficit of rare alleles was predicted for the eastern regions of Quebec (with the more pronounced effect shown in the North-East region), whereas an excess of rare alleles was found for the western regions of Quebec (North-West, North of Montreal, Montreal, South of Montreal and Centre regions). Because in our simulations new alleles enter the population only by migration events, this excess is likely explained by the documented increased immigration in the western regions (Bhérer et al. 2011), which introduced new rare alleles at greater rate than the rate of drift. This echoes to recent theoretical work showing a persistent peak of diversity

can result from an abrupt migration event in a previously isolated population (Alcala et al. 2013). Most notably, our results are consistent with a recent next-generation sequencing study, which observed an excess of rare private polymorphisms in a sample of French Canadians of unknown regional origin, compared to French individuals (Casals et al. 2013). Our results suggest that an increased migration to the western regions of Quebec is sufficient to cause such excess, however, it is most likely that other mechanisms also contributed. First, de novo mutation events could further contribute an accumulation of new rare alleles, but those mutations introduced by newborns would be transmitted at lesser rate than those carried by adult founders, simply because of the lower probability of survival caused by pre-reproductive mortality (Thompson and Neel 1978). Second, rare variants may have been introduced by admixture due to the 10% genetic contribution of founders of other-than-French origin (Bhérer et al. 2011; Moreau et al. 2013). Understanding the relative contribution of these processes to the excess of rare alleles is crucial to our understanding of human evolutionary history.

We have shown how the genealogical history of a young founder population can lead to a drastic increase in frequency of rare variants, segregating in Mendelian ratio without any preferential transmission. Variant introduced among 1% of founders' chromosome can reach more than 10% carrier frequency in some regions. Furthermore, our simulations prove possible that a founder mutation introduced in a single copy could have increase up to 2.5% carrier frequency in the whole Quebec population and to 5% or more in the North-West, the North-East and East regions. This great increase in a very short evolutionary time can be seen as a hallmark example of genetic signatures caused by the genealogical history of a young founder population. Most importantly, these signatures of neutral demographic processes can be confounded with those of positive selection. The greatest frequency increase was observed in the most remote regions from the core settlement of Quebec, thus suggesting that it may be attributable to surfing at the wave front of range expansion. Further data is required to investigate this question.

Besides, although we simulated neutral variants, these results are expected to hold for slightly deleterious variants and recessive mutations (Heyer 1999). Our results thus show that the genealogical history of Quebec is sufficient to explain the clinical founder effects - the elevated carrier frequency of Mendelian disorders - observed in the North-East, including the regions of Charlevoix and Saguenay-Lac-St-Jean, in agreement with previous studies (Heyer 1999; Tremblay et al. 2003). Although it is difficult to translate these findings in prediction of the burden of Mendelian diseases in regional populations, our results are in agreement with the documented higher concentration of Mendelian diseases at elevated frequency in the North-East and East regions (Scriver 2001; Laberge et al. 2005b), and point to a possible concentration of such diseases in the North-West, emphasizing the need for further clinical studies in that region.

In addition to the significant genetic differentiation proving departure from random mating in historical Quebec, we also report other evidences suggesting that the French Canadian genealogical structure does not conform to assumptions of the Wright-Fisher idealized population model. When restricting to regional genealogical histories, we observed a skewed distribution of frequency changes, with a long tail of alleles reaching high frequency. This is explained by the variance in reproductive success between founders which is very widespread and skewed, with a few top-contributing founders having a nearly flat distribution of reproductive success. This skewed distribution of reproductive success can be the result of the intergenerational correlation in effective family size on the wave front (Moreau et al. 2011a), or of the documented geometric distribution in offspring number (Austerlitz and Heyer 1998), but our analysis of top-contributing founders also suggests that cooperative behavior at the frontier might be involved (see below). In any case, our results stress the importance of finding more realistic models of family structure and reproductive success in humans, especially when modeling human range expansion and recent demography (Simons et al. 2014).

The top-contributing founders do not appear as a random group of founders suggesting that some yet unknown factor may have contributed to their tremendous reproductive success. We found a significant excess of Perche among the top-contributing founders. Furthermore, the top-contributing founders were contemporary to each other and were all living in Quebec City. These facts make it possible that they knew each other. It is indeed the case for the two related super founder couples, one including the daughter of the other, born in Quebec. This implies that top-contributing founders might have cooperated with each other. Recent studies have demonstrated that cooperative behaviors are promoted at the front of a range expansion (e.g. Datta et al. 2013; Van Dyken et al. 2013). Here, we hypothesize that the high reproductive success of the top-contributing founders was driven by cooperative behaviors. Cooperative behaviors among the Perche founders was previously suggested by De Braekeleer and Dao 1994) who have highlighted the high genetic contribution of Perche. However, as shown here, high reproductive success is not restricted to Perche individuals. We propose that in the first few generations following the birth of the population, cooperative behaviors among some pioneers may have promoted their long-term reproductive success. Beyond the particular case of the top-contributing founders, these results show how the foundation of a new population imply unexpected evolutionary mechanisms that cause a perturbation in the variance of reproductive success among individuals and have genetic consequence that extend outside the usual loss of diversity expectations.

In this study, we used an uncommon approach to study recent evolutionary changes in humans. We used extensive population genealogies to simulate and estimate the changes in allele frequencies between the founders of the Quebec population and a sample of contemporary individuals. Our study can thus be seen as a semi-natural experiment in humans, analog to the famous Buri's experiments, who recorded the dynamics of allele frequency changes in captive drosophila populations (Buri 1956). The idea of simulating genes transmission through the population pedigree dates back at least to Edwards

1968), and was applied in early work of MacCluer et al. 1986). Genealogical simulations have been used for decades in epidemiological studies in routine quality analysis of their methods. However, it was fairly rarely applied to address population genetic questions (but see inspiring papers by: Heyer 1999; Austerlitz and Heyer 2000; Helgason et al. 2003; Pardo et al. 2005; Chong et al. 2012). This is surprising knowing that computerized genealogical database have been available in different species and many human populations for quite a while, and that huge internet-based genealogical database are now growing in size and number. Here we exemplified how to use simulation experiments in a population genealogical tree to generate genotypic data and use it to perform population genetic analysis. This approach critically depends on the completeness and reliability of the population genealogies, issues that should be addressed with caution, especially in public based genealogical database. Here we considered that the impact of kinship structure and *de novo* mutations was negligible in Quebec, but those might have an important effect in other populations. For instance, unknown kinship among the founders will affect genetic diversity in the same direction than the founder effect by reducing the effective population size. Taking these issues into consideration, our approach could be applied to study the genetic changes that happened in the recent past of various species, or in human populations with various demographic histories where they have access to genealogies.

SUPPORTING MATERIAL

Supporting Methods

Founders: definitions and identification procedure

For the purpose of this study, we paid special attention to identify a maximum number of “true” founders of the population, i.e. the immigrant individuals who where the first to settle in Quebec. This definition differs from the usual definition of founder in pedigree analysis. Indeed, in genealogical data, a founder is typically defined as the last known ancestor going backward in time in a particular branch of the tree, in other words, an ancestor whose parents are unknown. As we said in the main text, in this study, we termed these individuals “genealogical founders”. This definition is useful, since there can be one and only one genealogical founder per ending branches of a tree. Genealogical founders are thus very easy to identify. However, importantly, in the Quebec genealogical dataset, a genealogical founder is not necessarily an ancestor who founded the population. Indeed, in this dataset, a genealogical founder can either be (i) a true immigrant founder who immigrated to Quebec, (ii) an immigrant founder’s parent who is recorded because she/he appears on a marriage certificate as parent of the bride or the groom but never came to Quebec, (iii) a Native American founder who has descendents among French Canadians or (iv) an ancestor born in Quebec for whom we lack information on the parents, such as adoptees.

To study the immigrants who founded the French Canadian population of Quebec, we used information on the immigrant status (native or immigrant) and on the geographical origins of individuals that was available in the BALSAC database. The immigrant founders of the population where defined as the first ancestors of each genealogical line to settle in Quebec. Following the same logic, we identified the Native Americans founders who were the first to introduce their genetic diversity into the French Canadian gene pool.

However, this definition complicates the task of identifying founders, because there might be zero, one or more than one immigrant or Native American individual along the same ascending lineage. In addition, this definition introduces the additional step of identifying semi-founders, who are immigrant individuals who have one parent connected to some other founders and the other parent disconnected. For example, an individual born in the United States from a mother of French Canadian descent and an American father, who marries and has descendants in Quebec. From a genetic point of view, a semi-founder introduces part of his/her genome to the population. To circumvent these problems in the identification of founders, we developed a procedure to select one and only one founder per genealogical line.

The procedure used to identify the founders in ascending genealogies can be divided in four major steps. First, we separately identify the three categories of founders (genealogical founders, Native American founders and immigrant founders). Genealogical founders are identified as individual with father and mother identification number equal to zero. Native American founders are identified in four substeps: (i) identify the Native American individuals based on the status variable from BALSAC; (ii) calculate the matrix of genetic contribution of these individuals to themselves; (iii) individuals who did not receive any genetic contribution from the other individuals are labeled as founders. The immigrant founders identification is based on the “immigrant status” variable from BALSAC which includes five types: native, immigrant, never came to Quebec, native or immigrant and immigrant or never came to Quebec. For the three types that include immigrants, we identified the founders following the same substeps as for Native American founders, plus one step for the semi-founders: (iv) we label as semi-founders individuals who received a genetic contribution from the other founders but only through one parental line. Second, we select the founders among the genealogical lines following an order of priority between the categories of founders, each time selecting the founders by checking at their mutual genetic contribution. We applied the same order of priority for the founders and the semi-founders, so

"founders" refers here to both. We prioritize (a) the Native American founders and immigrant founders (all 3 types) over the genealogical founders; (b) the Native American founders and the immigrant founders with "immigrant status" = "immigrant" over the immigrant founders with "immigrant status" = "native or immigrant" or = "immigrant or never came to Quebec"; (c) the oldest founder between a Native American and a immigrant founders with "immigrant status" = "immigrant"; (d) the oldest between a immigrant founders with "immigrant status" = "native or immigrant" and an immigrant founders with "immigrant status" = "immigrant or never came to Quebec". Third, for sake of the simulation experiments that cannot take into account semi-founders, we replaced each semi-founder by its parent who did not receive any genetic contribution from any other founder. Four, to make sure we have one founder in each line, we climb up the genealogical lineages until we reach a selected founder, and if we don't find any and reach a genealogical founder, we select this one as a founder.

Supporting Results

Genealogical analysis of the founders

In the whole Quebec genealogical dataset, we identified 8,834 founders. Among them we found 7,703 immigrant founders and 95 Native American founders that were described in Bhérer et al. 2011 (or one of their parent in cases of semi-founders). Immigrant and Native American founders cumulated 99.5% of the explained genetic contribution and were found among 99.8% of the ascending lineages (Bhérer et al. 2011). Among the remaining 0.2% of ascending lineages, we identified 1,036 genealogical founders. Although genealogical founders represent 11.7% of the total number of founders, overall they had a very low genetic contribution to the whole Quebec gene pool, with an explained genetic contribution of 0.5%. We confirmed that there was a single founder per genealogical lineage by checking that founders had zero genetic contribution to themselves (results not shown). The total genetic contribution of the 8,834 founders to the 2,221 individuals belonging to the

current generation was, as desired exactly equal to 1 for 80% of individuals, between 0.99 and 1.01 for 98% of them, between 0.96 and 0.99 for 10 individuals and between 1.01 and 1.03 for 26 individuals. Manual examination these last 36 outlier individuals indicated the lower/higher total genetic contribution total was likely related to the identification of semi-founders in the tree. Replacing semi-founders by their ancestors who are disconnect to the Quebec genealogical tree (and instead selecting these ancestors as founders) during the first prioritizing phase of the identification procedure should in principle allow a genetic contribution equal to one.

Validation of coalescence simulations within fixed population genealogy

To validate our simulation algorithm, we compared our simulation results to theoretical expectations of founders' genetic contribution (theoretical GC) calculated exactly in the genealogy (Roberts 1968; Bhérer et al. 2011). Mean number of founders' allele copies obtained with the coalescence simulations (simulated GC) was almost perfectly correlated to theoretical GC (for both alleles: Pearson' r = 0.9997; p-value < 2.2×10^{-16}), thus suggesting that 100,000 iterations allowed for accurate approximation of transmission of founders alleles (Figs. S2). We excluded five outlier founders in the following analysis because the correlation between their simulated and theoretical GC was less than 0.8. The mean number of copies of founders' allele 1 and 2 was also almost perfectly correlated, further validating our approach (Pearson' r = 0.99998; p-value < 2.2×10^{-16}) (Fig. S2). We also compared the simulated and theoretical GC for the 13,000 genealogical founders, for which we are sure there is exactly one and only one founder per genealogy. In that case, the correlation was even higher (for both alleles: Pearson' r = 0.999994; p-value < 2.2×10^{-16}), indicating that the potential errors in the identification of founders had negligible effect and further confirmed the almost perfect agreement between simulations and theoretical expectations of genetic contribution with 100,000 iterations.

Supporting Tables

Table S1. Descriptive statistics of the Quebec sample.

	BALSAC region[a]	Population 1956[b]		1945-65 sample		Ancestors	Founders
		N	%	n_s[c]	%	n_a[d]	n_f[e]
WQC	Whole Quebec	4,628,378	100	2,221	100	153,447	8,834
NW		157,239	3.4	87	3.9	29,557	3,761
	Abitibi	99,578	2.2	56	2.5		
	Témiscamingue	57,661	1.2	31	1.4		
NMTL		542,041	11.7	242	10.9	43,709	4,525
	Outaouais	189,477	4.1	92	4.1		
	Laurentides	109,821	2.4	58	2.6		
	Lanaudière	114,377	2.5	57	2.6		
	North Shore of Montreal	128,366	2.8	35	1.6		
MTL		1,507,653	32.6	722	32.5	87,706	6,785
	Montreal Island	1,507,653	32.6	722	32.5		
SMTL		395,453	8.5	178	8.0	42,114	4,518
	South Shore of Montreal	114,317	2.5	56	2.5		
	Richelieu	281,136	6.1	122	5.5		
CTR		752,706	16.3	348	15.7	54,478	4,921
	Mauricie	225,594	4.9	111	5.0		
	Bois-Francs	218,092	4.7	104	4.7		
	Estrie	309,020	6.7	133	6.0		
QUE		531,054	11.5	272	12.2	40,605	3,775
	Quebec City	288,754	6.2	140	6.3		
	Quebec Region	123,053	2.7	64	2.9		
	Beauce	94,649	2.0	51	2.3		
	Côte-de-Beaupré	24,598	0.5	17	0.8		
NE		322,299	7.0	157	7.1	22,478	2,717
	Charlevoix	30,263	0.7	19	0.9		
	Saguenay-Lac-Saint-Jean	234,672	5.1	106	4.8		
	Côte-Nord	57,364	1.2	32	1.4		
E		419,933	9.1	215	9.7	27,193	3,328
	Côte-du-Sud	76,219	1.6	41	1.8		
	Bas-Saint-Laurent	157,536	3.4	80	3.6		
	Gaspésie	174,622	3.8	83	3.7		
	Îles-de-la-Madeleine	11,556	0.2	11	0.5		

[a] BALSAC regions are composed of 24 geographical units (http://balsac.uqac.ca)
[b] Reported in the 1956 Canadian Census (www.statcan.gc.ca)
[c] n_s = size of current generation samples.
[d] n_a = number of ancestors identified in the ascending genealogies of current generation samples.
[e] n_f = number of founders identified in the ascending genealogies of current generation samples.

Table S2. Distribution of year at first marriage of ancestors and founders.

	WQC	NW	NMTL	MTL	SMTL	CTR	QUE	NE	E
ANCESTORS									
median	1792	1723	1741	1766	1736	1751	1749	1733	1747
upper	1946	1893	1924.5	1927	1906.5	1923	1937.5	1943	1943
lower	1627	1600	1600	1607	1600	1600	1600	1600	1600
q1	1750	1678	1697	1727	1694	1708	1700	1680	1690
q3	1832	1764	1788	1807	1779	1794	1795	1794	1800
mean	1789.6	1725.6	1744.4	1766.4	1738.9	1751.9	1749.4	1739.9	1747.4
n	153,209	29,429	43,575	87,516	41,970	54,339	40,480	22,365	27,066
FOUNDERS									
median	1731	1672	1679	1706	1678	1685	1673	1670	1675
upper	1916.5	1757.5	1804.5	1887.5	1796	1824.5	1782.5	1758.5	1850
lower	1600	1609.5	1600	1600	1600	1600	1600	1602.5	1602
q1	1674	1665	1667	1670	1666	1667	1665	1661	1665
q3	1771	1702	1722	1757	1718	1730	1712	1700	1739
mean	1733.6	1685.7	1696.8	1717.3	1693.8	1699.4	1690.7	1686.4	1701.1
n	8,804	3,755	4,517	6,767	4,509	4,913	3,767	2,711	3,315

Upper and lower notches of boxplot are calculated in R as +/-1.58 IQR/sqrt(n).

n : number of ancestors (or founders) for whom we could determine year at first marriage after 1600.

Table S3. Maximum F_{ST} values per sites [lower matrix] between pairs of current generation samples and number of simulated diallelic sites [upper matrix] considered.

	NW	NMTL	MTL	SMTL	CTR	QUE	NE	E
NW		11,889,372	13,302,816	11,573,730	12,282,450	11,851,537	10,980,498	11,585,353
NMTL	0.0459		13,548,368	12,298,583	12,809,282	12,520,327	11,992,887	12,354,582
MTL	0.0356	0.0189		13,465,239	13,697,974	13,547,289	13,322,430	13,476,089
SMTL	0.0582	0.0369	0.0261		12,626,035	12,303,399	11,674,735	12,099,843
CTR	0.0368	0.0260	0.0140	0.0339		12,729,782	12,346,102	12,634,159
QUE	0.0591	0.0380	0.0184	0.0303	0.0249		11,893,543	12,268,056
NE	0.0768	0.0740	0.0587	0.0712	0.0672	0.0723		11,636,560
E	0.0577	0.0396	0.0287	0.0406	0.0357	0.0370	0.0708	

Table S4. Allele frequency changes between founders and current generation samples given p_f = 25% and 50%.

	WQC	NW	NMTL	MTL	SMTL	CTR	QUE	NE	E
p_f = 0.25									
Coefficient of variation (x10^2)	4,5	13,7	8,7	5,5	9,7	7,6	9,1	16,6	10,4
P(loss)	0	0	0	0	0	0	0	0	0
P(p_s<p_f)	50,9	50,8	48,8	49,5	48,6	49,2	49,0	51,8	51,0
P(p_s>=$2p_f$)	0	0	0	0	0	0	0	0	0
Max Increase	1,2	1,7	1,4	1,3	1,4	1,3	1,5	1,8	1,5
p_f = 0.5									
Coefficient of variation (x10^2)	2,6	7,9	5,0	3,2	5,6	4,4	5,3	9,6	5,9
P(loss)	0	0	0	0	0	0	0	0	0
P(p_s<p_f)	49,8	47,2	48,4	49,0	48,0	49,1	48,3	49,0	49,1
P(p_s>=$2p_f$)	0	0	0	0	0	0	0	0	0
Max Increase	1,1	1,4	1,2	1,1	1,3	1,2	1,2	1,4	1,2

Table S5. Summary of Kruskal-Wallis test for difference in the distribution of frequency changes between pairs of samples for the five initial founders' allele frequencies considered.

1%	WQC	NW	NMTL	MTL	SMTL	CTR	QUE	NE	E
WQC		1259,1	431,5	18,7	582,7	221,2	909,4	7180,2	1191,8
NW	8,93E-276		1059,2	1209,2	374,9	755,2	0,5	55,3	614,0
NMTL	7,72E-96	2,44E-232		942,4	7,4	15,2	9,7	2411,6	72,0
MTL	1,50E-05	6,04E-265	6,02E-207		62,2	103,1	685,2	6369,8	650,1
SMTL	9,61E-129	1,62E-83	6,57E-03	3,03E-15		1028,8	173,3	1303,5	44,5
CTR	5,06E-50	3,00E-166	9,72E-05	3,21E-24	1,00E-225		150,8	4568,2	315,7
QUE	9,01E-200	0,4989	1,86E-03	4,84E-151	1,40E-39	1,16E-34		3074,0	8,4
NE	0	1,04E-13	0	0	1,94E-285	0	0		1904,4
E	3,71E-261	1,49E-135	2,19E-17	2,11E-143	2,61E-11	1,27E-70	3,74E-03	0	
5%	WQC	NW	NMTL	MTL	SMTL	CTR	QUE	NE	E
WQC		287,0	54,6	1,7	100,9	45,8	106,0	1709,7	212,7
NW	2,23E-64		56,9	265,0	8,9	150,2	29,3	303,1	41,9
NMTL	1,47E-13	4,57E-14		56,4	4,4	2,7	0,1	878,3	14,1
MTL	1,96E-01	1,43E-59	6,01E-14		73,9	25,9	61,9	1623,8	157,2
SMTL	9,58E-24	2,82E-03	0,0362	8,02E-18		4,2	8,6	643,3	0,3
CTR	1,29E-11	1,57E-34	0,0983	3,65E-07	0,0407		22,6	1104,1	81,5
QUE	7,24E-25	6,25E-08	0,8011	3,53E-15	0,0033	2,00E-06		960,2	9,0
NE	0	6,80E-68	5,04E-193	0	6,41E-142	4,34E-242	8,16E-211		613,0
E	3,45E-48	9,48E-11	1,71E-04	4,55E-36	5,67E-01	1,78E-19	2,70E-03	2,46E-135	
10%	WQC	NW	NMTL	MTL	SMTL	CTR	QUE	NE	E
WQC		120,7	19,7	1,8	9,3	12,9	25,7	689,9	81,5
NW	4,40E-28		45,2	116,1	216,9	69,5	24,9	102,6	5,7
NMTL	9,23E-06	1,75E-11		1,9	0,1	0,7	0,0	398,5	7,4
MTL	0,1824	4,46E-27	0,1730		15,5	12,9	33,4	685,9	79,6
SMTL	0,0023	4,19E-49	0,7385	8,40E-05		0,8	1,1	324,6	20,0
CTR	3,25E-04	7,83E-17	0,3887	3,30E-04	0,3710		3,4	470,1	33,7
QUE	4,04E-07	5,96E-07	0,9334	7,32E-09	0,3006	0,0669		393,2	15,1
NE	4,69E-152	4,02E-24	1,18E-88	3,41E-151	1,47E-72	3,00E-104	1,71E-87		261,6
E	1,75E-19	0,0168	0,0065	4,67E-19	7,69E-06	6,58E-09	1,03E-04	7,67E-59	
25%	WQC	NW	NMTL	MTL	SMTL	CTR	QUE	NE	E
WQC		23,1	1,8	1,0	14,8	5,5	4,9	143,7	19,7
NW	1,57E-06		12,3	26,4	4,3	9,0	8,4	23,8	1,8
NMTL	0,1816	4,63E-04		3,0	4,9	0,7	0,5	92,4	7,4
MTL	0,3180	2,81E-07	0,085		17,9	8,9	6,6	142,5	22,4
SMTL	1,21E-04	0,0386	0,027	2,29E-05		1,6	2,2	58,2	0,4
CTR	0,0188	0,0027	0,391	2,87E-03	0,2018		0,0	88,9	4,0
QUE	0,0264	0,0038	0,475	9,96E-03	0,1355	0,9020		80,1	4,2
NE	4,09E-33	1,07E-06	7,10E-22	7,52E-33	2,40E-14	4,09E-21	3,49E-19		50,9
E	9,21E-06	0,182	0,006	2,25E-06	0,5538	0,0465	0,0415	9,77E-13	
50%	WQC	NW	NMTL	MTL	SMTL	CTR	QUE	NE	E
WQC		0,2	3,7	3,4	0,5	3,7	8,5	4,6	18,0
NW	0,6191		3,8	1,8	0,5	0,2	5,4	1,6	5,3
NMTL	0,0544	0,0528		0,6	1,8	10,8	0,4	10,0	27,4
MTL	0,0658	0,1744	0,4231		0,2	9,6	2,5	7,5	26,0
SMTL	0,4972	0,4959	0,1809	0,6652		3,0	3,5	4,2	13,1
CTR	0,0536	0,6978	0,0010	0,0019	0,0827		14,6	1,5	6,3
QUE	0,0036	0,0200	0,5363	0,1130	0,0626	0,0001		12,3	31,5
NE	0,0320	0,2069	0,0016	0,0063	0,0413	0,2155	0,0005		0,3
E	2,19E-05	0,0207	1,66E-07	3,48E-07	0,0003	0,0121	1,97E-08	0,5959	

The lower matrix shows the p-value of the test (in red : not significant at the 0.05 level) and the upper matrix shows the F statistic of the test.

Table S6. Characterization of the 11 top-contributing founders.

Couple		a		b		c		d		e		f	
Individual		Ozanne Achon	Pierre Tremblay	Marie Michel	Louis Gagné	Marguerite Langlois	Abraham Martin	Marguerite Martin	Étienne Racine	Mathurine Robin	Jean Guyon	Madeleine Giguère	Jean Roussin
Sex		F	M	F	M	F	M	F	M	F	M	F	M
Year of marriage		1657		1638		1620		1638		1615		1622	
Place of marriage		Quebec City		Perche		France		Quebec City		Perche		Perche	
Migrant status		Immigrant	Immigrant	Immigrant	Immigrant	Immigrant	Immigrant	Born in Quebec	Immigrant	Immigrant	Immigrant	outside Quebec	Immigrant
Origin		Perche	Perche	Perche	Perche	France	Great Brit	Quebec	Normandie	Perche	Perche	France	Perche
Max Nb Copies	Allele 1	52	57	43	46	48	50		45	47	45	45	57
	Allele 2	57	60	45	41	48	43		41	51	48	41	47
Mean Nb Copies	Allele 1	11.66	11.66	7.56	7.58	6.87	6.86		6.56	6.24	6.19	5.21	5.25
	Allele 2	11.72	11.60	7.58	7.58	6.86	6.88		6.54	6.24	6.21	5.25	5.28
Relative genetic contribution to current generation sample (%)	WQC	0.52	0.52	0.34	0.34	0.31	0.31		0.30	0.28	0.28	0.24	0.24
	NW	0.48	0.48	0.38	0.38	0.37	0.37		0.34	0.30	0.30	0.24	0.24
	NMTL	0.09	0.09	0.08	0.08	0.16	0.16		0.08	0.15	0.15	0.08	0.08
	MTL	0.33	0.33	0.22	0.22	0.22	0.22		0.18	0.21	0.21	0.16	0.16
	SMTL	0.23	0.23	0.17	0.17	0.18	0.18		0.16	0.19	0.19	0.12	0.12
	CTR	0.24	0.24	0.21	0.21	0.26	0.26		0.21	0.26	0.26	0.18	0.18
	QUE	0.49	0.49	0.44	0.44	0.51	0.51		0.53	0.45	0.45	0.41	0.41
	NE	3.41	3.41	1.77	1.77	0.79	0.79		1.34	0.53	0.53	0.77	0.77
	E	0.32	0.32	0.25	0.25	0.34	0.34		0.13	0.34	0.34	0.25	0.25
Proportion of descendants among current generation samples (%)	WQC	47.7	47.7	58.8	58.8	85.1	85.1		56.2	89.7	89.7	74.6	74.6
	NW	49.4	49.4	70.1	70.1	95.4	95.4		74.7	94.3	94.3	75.9	75.9
	NMTL	26.4	26.4	43.8	43.8	81.8	81.8		39.3	84.3	84.3	62.0	62.0
	MTL	42.4	42.4	50.7	50.7	79.5	79.5		47.1	86.6	86.6	67.6	67.6
	SMTL	42.7	42.7	50.0	50.0	78.1	78.1		51.7	83.1	83.1	65.2	65.2
	CTR	42.0	42.0	55.5	55.5	85.6	85.6		56.9	93.4	93.4	75.9	75.9
	QUE	66.2	66.2	78.3	78.3	95.6	95.6		77.6	98.9	98.9	95.2	95.2
	NE	91.7	91.7	92.4	92.4	96.8	96.8		92.4	97.5	97.5	94.9	94.9
	E	46.5	46.5	61.9	61.9	86.0	86.0		47.4	87.0	87.0	76.7	76.7

Supporting Figures

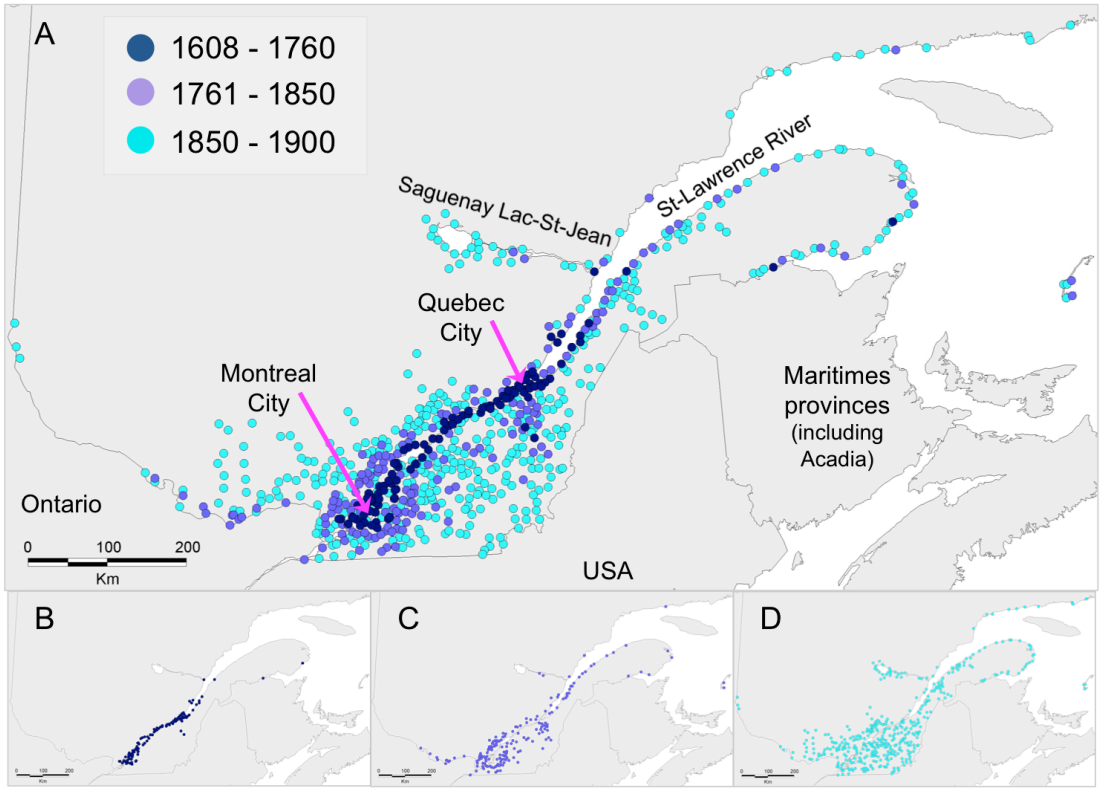

Figure S1. Progression of Quebec settlement.

(A) Map of the progression of Quebec settlement from the foundation of the colony in 1608 to the 20th century. Each dot pinpoints the opening of a new locality and is colored according to period of settlement illustrated in the following maps. (**B) 1608-1760:** *Nouvelle-France* period, between the foundation of Quebec City in 1608 and the British Conquest of 1760. Population size in 1760: 70,000. (**C) 1760-1850:** After the Conquest, French immigration stopped and growth of the population relied mostly on high fertility rates. Estimated number of French Canadians in 1850: 670,000. (**D) 1850-1900:** Expansion in the backcountry. It is the time of great population movements including colonization of remote territories such as the Saguenay-Lac-St-Jean region but also of major emigration to USA. In 1900, 1,350,000 French Canadians are counted among a total of 1,600,000 inhabitants. Data for the maps comes from (Brais et al. 2007).

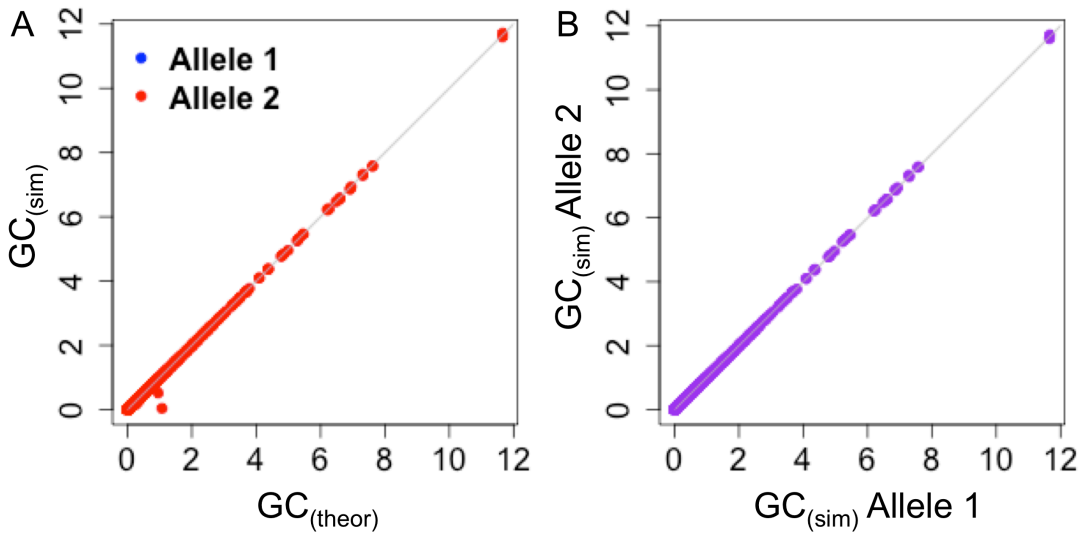

Figure S2. Validation of backward coalescent-like simulations.

(A) Correlation between the founders' simulated genetic contribution ($GC_{(sim)}$) and theoretical expectation of genetic contribution ($GC_{(theor)}$) to the whole Quebec sample. $GC_{(sim)}$ was calculated for each of the two founders' allele as the mean number of copies of an allele in the WQC obtained with 100,000 iterations of backward coalescent-like simulations. Line of perfect correlation is drawn in grey. **(B)** Correlation between the mean number of copies obtained for allele 1 ($GC_{(sim)}$ Allele 1) and allele 2 ($GC_{(sim)}$ Allele 2).

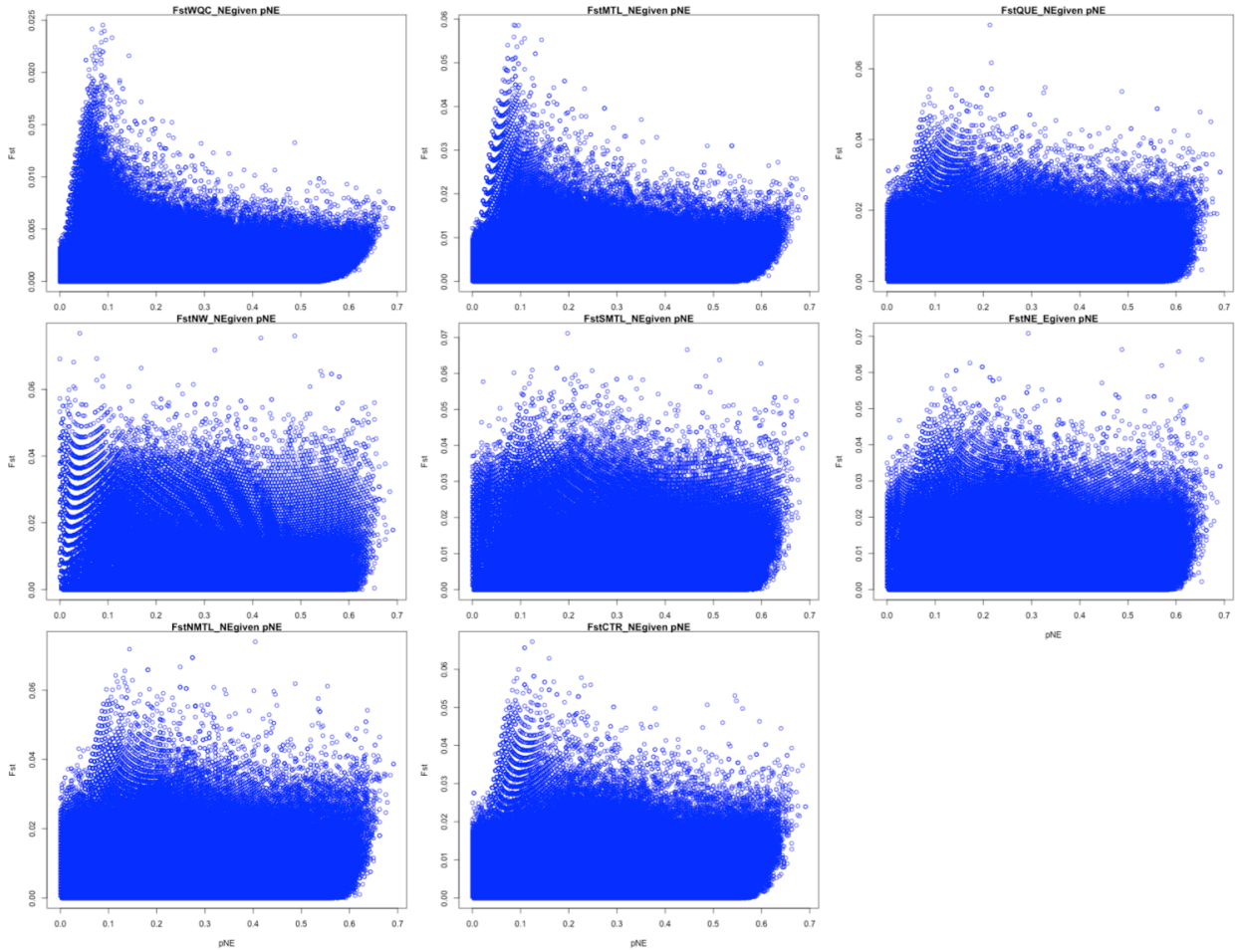

Figure S3. F_{ST} per locus between the NE and the eight other samples given allele frequency in the NE (*pNE*).

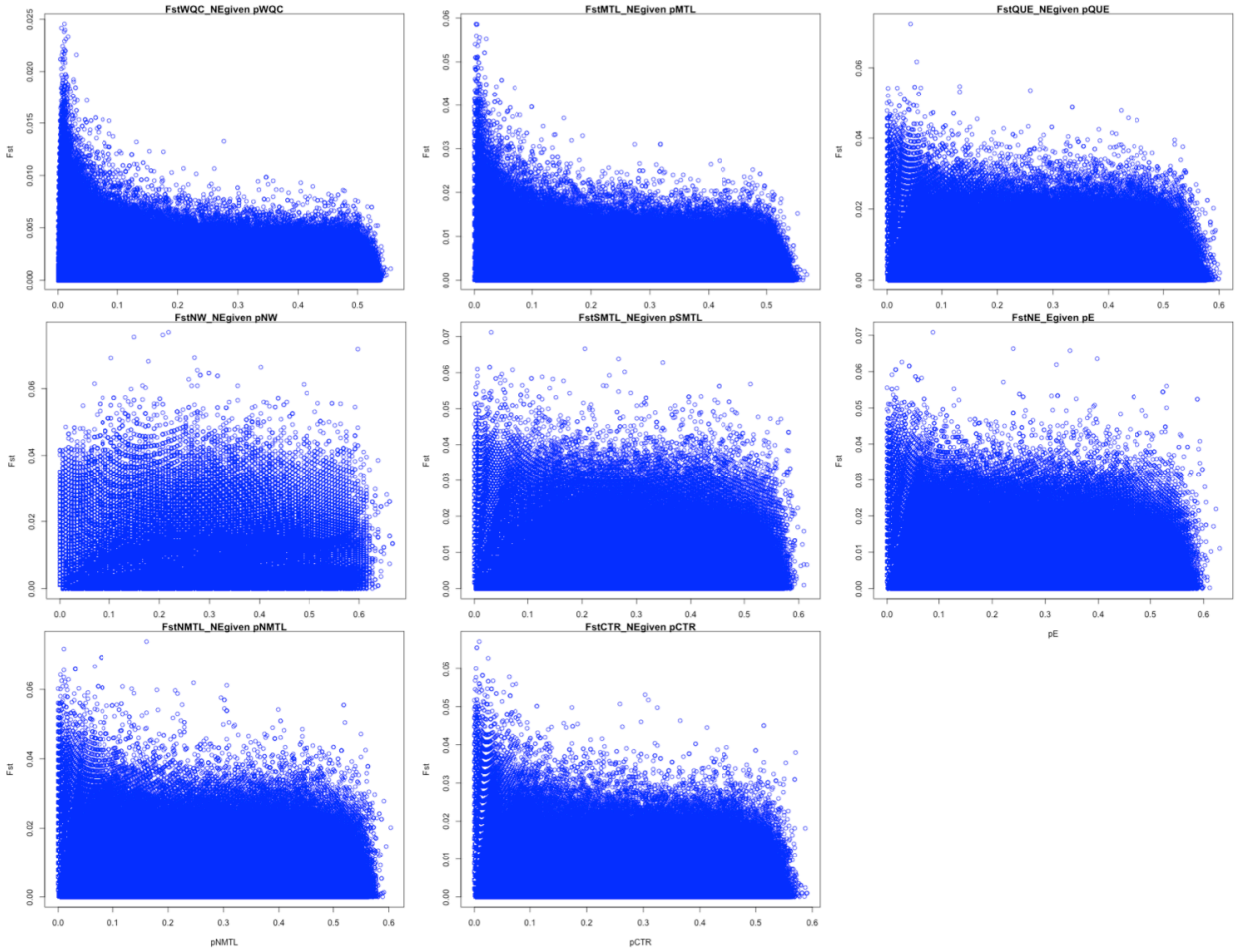

Figure S4. F_{ST} per locus between the NE and the eight other samples given allele frequency in the other samples.

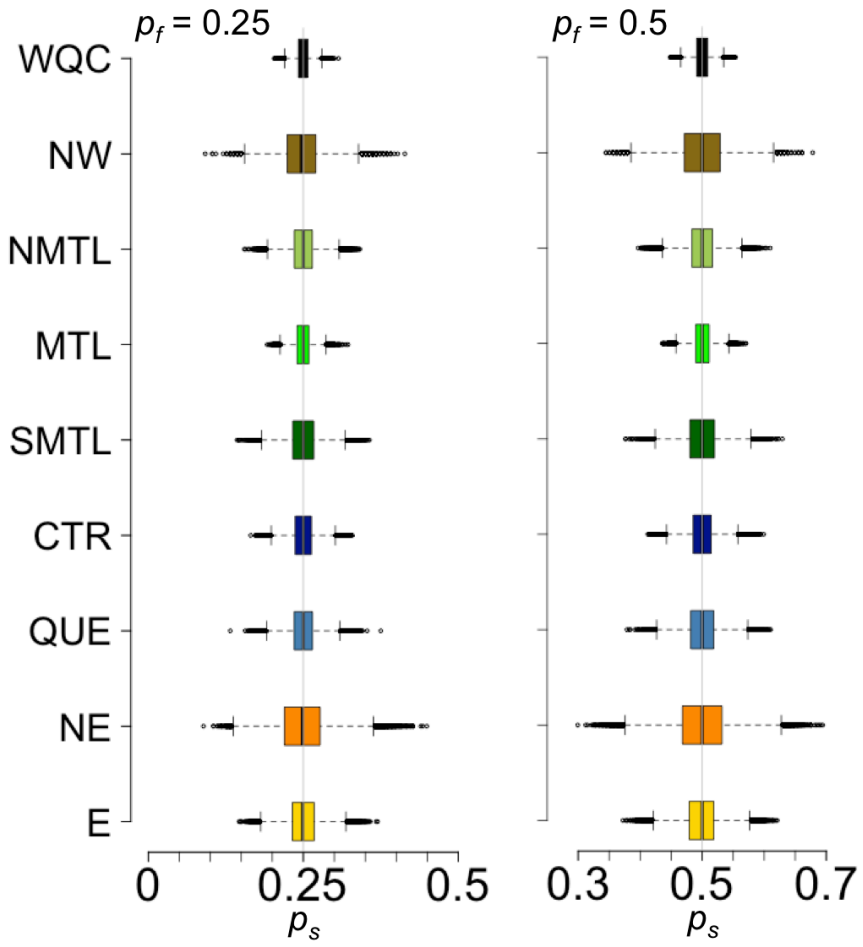

Figure S5. Allele frequency changes between founders and current generation.

For each sample, the probability distribution of current generation' allele frequency (p_s) conditional on founders' allele frequency (p_f) was estimated with 100,000 AD simulations. We assumed p_f =0.25 (left panel) and p_f =0.5 (right panel). The following abbreviations are used: WQC: Whole Quebec, NW: North-West, NMTL: North of Montreal, MTL: Montreal City Area, SMTL: South of Montreal, CTR: Centre, QUE: Quebec City area, NE: North-East, E: East.

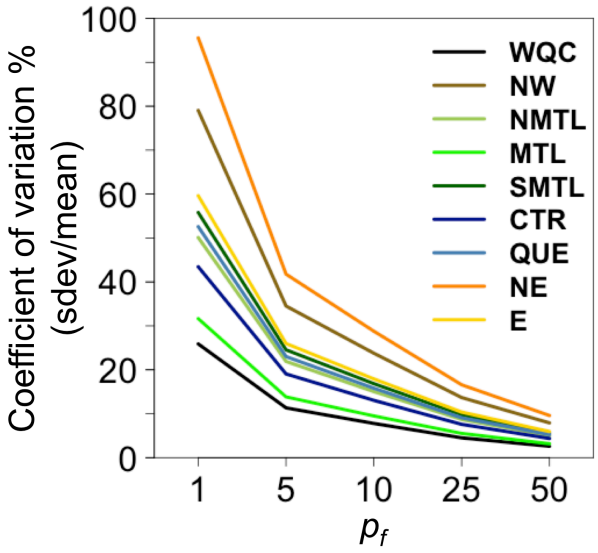

Figure S6. Variance of frequency changes distribution.

For each sample, coefficient of variation was calculated as the standard deviation of the distribution of allele frequencies (please refer to Figure 3 and S5), divided by the mean (%). This measure of variance was calculated for five founders' initial allele frequency, p_f (1%, 5%, 10%, 25%, 50%). The following abbreviations are used for samples: WQC: Whole Quebec, NW: North-West, NMTL: North of Montreal, MTL: Montreal City Area, SMTL: South of Montreal, CTR: Centre, QUE: Quebec City area, NE: North-East, E: East.

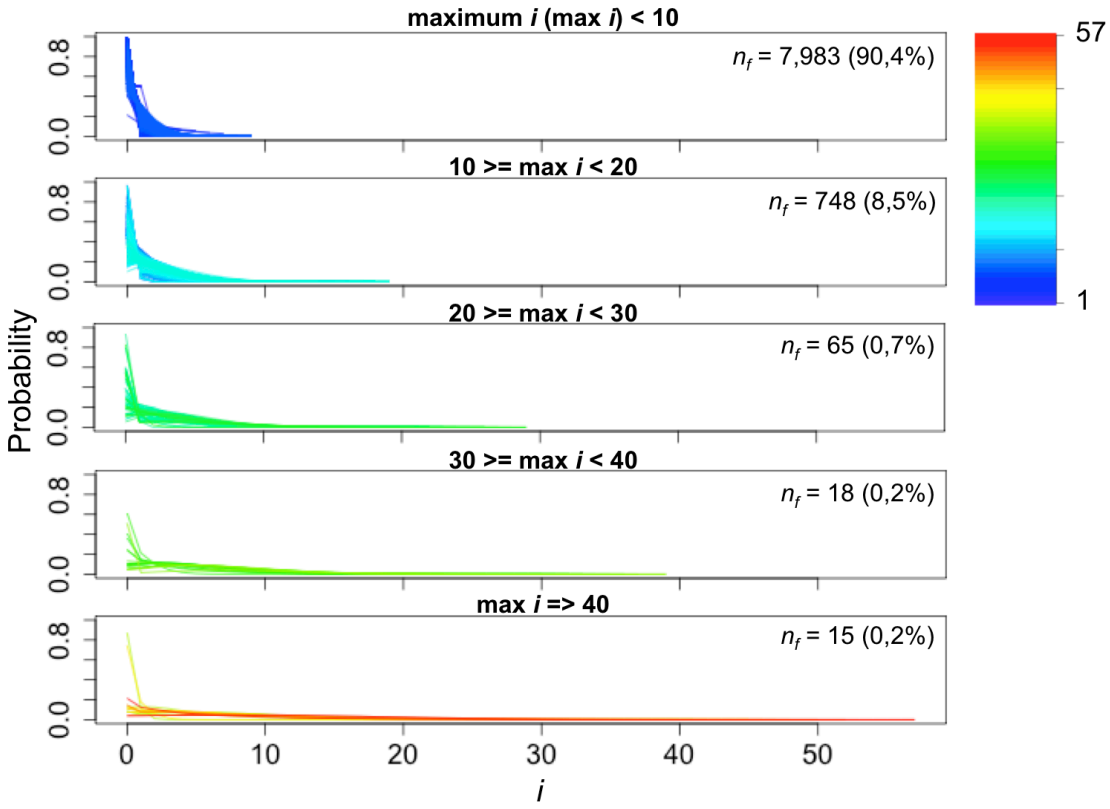

Figure S7. Distribution of genetic contribution to WQC per unique founder allele.

Each curve represents the distribution of the number of unique founders' allele contributed to the WQC current generation sample obtained with 100,000 iterations of backward coalescent-like simulations. For simplicity, results obtained for only one of the two founders alleles are presented, but note that their mean number is almost perfectly correlated (Figure S2). We excluded five outliers founders identified in Figure S2A. Each founder' curve is colored according to the maximum number of copies contributed; following the colorstrip ranging from minimum = 1 copy (dark blue) to maximum = 57 copies (red). The founders are divided in five categories of increasing contribution (from top to bottom) according to the maximum number of allele copies contributed.

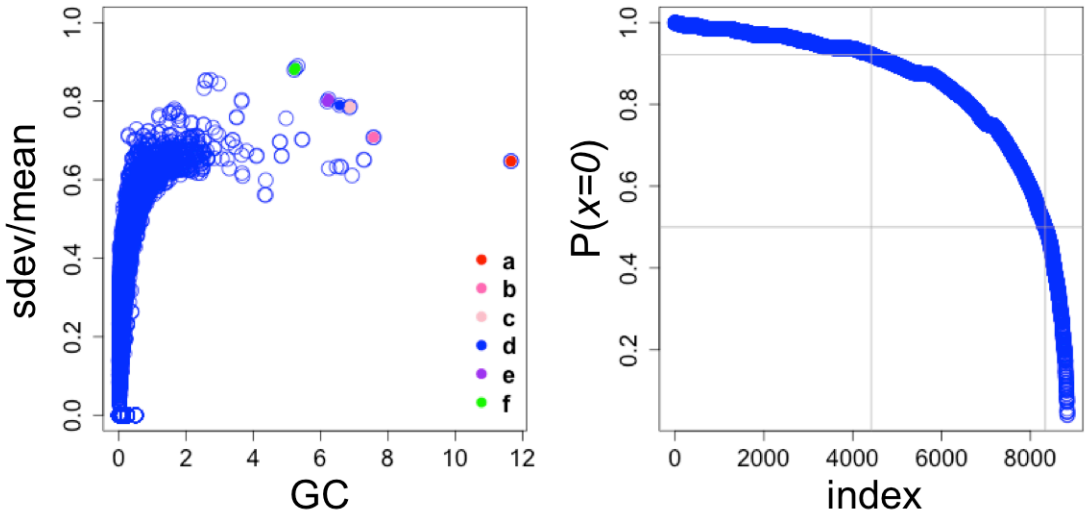

Figure S8. Distribution of number of copies of unique founders' allele in the whole Quebec sample.

Left panel shows the coefficient of variation of the founders' distribution of number of copies of alleles transmitted in the whole Quebec sample, conditional on survival, plotted against the genetic contribution (GC) of each founder and calculated as the mean of the distribution. The 6 colored points represent the 6 top-contributing founders highlighted in Figure 4B (from a to f). The right panel shows the probability of extinction for the 8,834 founders, ranked in decreasing order of probability. The first vertical line represents 50% of founders (whose alleles have over 92% probability of extinction). The second, at the 94% of founders, represents the threshold under which the probability of extinction of founders' alleles is higher than the probability of survival.

CHAPITRE V:

Deep human genealogies reveal a selective advantage to be on an expanding wave front

Claudia Moreau, Claude Bhérer, Hélène Vézina, Michèle Jomphe, Damian Labuda, Laurent Excoffier

Référence:

Moreau C, Bhérer C, Vézina H, Jomphe M, Labuda D, Excoffier L. 2011. Deep human genealogies reveal a selective advantage to be on an expanding wave front. *Science* **334**(6059): 1148-1150.

CONTRIBUTION DES CO-AUTEURS

Pour cet article, ma contribution est la suivante:

- design de l'étude avec LE et DL;
- définition des unités géographiques étudiées avec CM, LE et DL;
- définition géographique du front versus cœur du peuplement avec LE, CM et DL;
- conseils et support pour les analyses statistiques généalogiques réalisées par CM (notamment : contribution génétique, consanguinité, définition et identification des fondateurs)
- cartographie de l'expansion de la population (Figure 1);
- analyse et interprétation des résultats;
- révision du manuscript.

La contribution des co-auteurs est la suivante: LE est l'auteur principal de l'article. Il a dirigé l'étude (incluant le design, la conception, la réalisation et l'interprétation de l'ensemble des analyses) et il a rédigé le manuscrit. CM a réalisé les analyses statistiques et bioinformatiques ainsi que les figures pour l'article à l'exception de la figure 1. HV et MJ ont construit l'ensemble de données généalogiques tiré du fichier de population BALSAC. DL a eu l'idée originale d'étudier le phénomène de surfing dans la population du Saguenay. Il a participé au design de l'étude, à l'analyse et à l'interprétation des résultats. CM, CB, HV, MJ et DL ont révisé le manuscrit.

ACKNOWLEDGMENTS

We thank L. Barreiro, D. Bhérer, H. Harpending, Y. Idaghdour, D. Reich, R. Shine, and M. Slatkin for their helpful comments on the manuscript. LE was supported by a Swiss NSF grant No 3100A0-126074, DL and HV by the

Réseau de Médecine Génétique Appliquée of the Fonds de Recherche en Santé du Québec (FRSQ), CB was a recipient of an FRSQ studentship. The raw genealogical data used in this study are available upon request from BALSAC project.

ABSTRACT

Since their origin, human populations have colonized the whole planet, but the demographic processes governing range expansions are mostly unknown. We analyzed the genealogy of more than 1 million individuals resulting from a range expansion in Quebec between 1686 and 1960, and reconstructed the spatial dynamics of the expansion. We find that a majority of the present Saguenay Lac Saint-Jean population can be traced back to ancestors having lived directly on or close to the wave front. Ancestors located on the front contributed significantly more to the current gene pool than those from the range core, likely due to a 20% larger effective fertility of women on the wave front. This fitness component is heritable on the wave front and not in the core, implying that this life-history trait evolves during range expansions.

REPORT

Most species go through environmentally induced range expansions or range shifts (Hewitt 2000), promoting the evolution of traits associated to dispersal and reproduction (Phillips et al. 2010). Humans likely colonized the world by a series of range expansions from Africa (Cavalli-Sforza et al. 1994), possibly with episodes of interbreeding with now extinct hominins (Green et al. 2010; Reich et al. 2010b), leading to allele frequency and heterozygosity clines from entry points into several continents (e.g. Prugnolle et al. 2005a; Wang et al. 2007). Range expansions can also lead to drastic changes in allele frequencies, sometimes mimicking the effect of positive selection in recently colonized habitats (Edmonds et al. 2004; Klopfstein et al. 2006), through a process called gene surfing (Klopfstein et al. 2006). Neutral, favorable or even deleterious mutations can surf and increase in frequency (Travis et al. 2007; Hallatschek and Nelson 2010), implying that wave fronts may harbor mutations with a wider range of selective coefficients than core populations. The evolutionary consequences of range expansions have been studied in a wide array of species (Phillips et al. 2010; Hill et al. 2011), but studies of the dynamics of range expansions have been generally restricted to species with short generation times (Biek et al. 2007; Hallatschek et al. 2007) or to invasive species (Pysek and Hulme 2005; Estoup et al. 2010), because both spatial and temporal sampling are required to understand the dynamics of wave fronts.

Deep-rooted human genealogies in recently expanded populations may offer an opportunity to study the wave front demographics and its genetic consequences on present day populations. We studied the genealogies reconstructed from Quebec parish registers that document the recent temporal and spatial expansion of the settlement of the Charlevoix Saguenay-Lac-Saint-Jean (ChSLSJ) region, North-East of Quebec City, Canada: a prime example of a recent, fast, and well-documented range expansion

(Lavoie et al. 2005) (Figure 1). The European colonization of Quebec was initiated in 1608 with the foundation of Quebec city, and the colony was well established by the end of the 17th century (Charbonneau et al. 2000). The peopling of the Charlevoix region started from Baie-Saint-Paul, and both a rapid demographic growth and the development of the timber industry promoted further expansions after 1838 up the Saguenay River and the Lac-Saint-Jean region (SLSJ) (Bouchard 1983; Gauvreau et al. 1991). The spatial and temporal dynamics of the peopling of the whole ChSLSJ region can be reconstructed by tracing back the founding events of new localities. As shown on Fig. 1, the inferred colonization process is a mixture of long-distance settlements creating an irregular wave front, followed by further, more progressive, short-range expansions, which then filled gaps and created a more regular wave front.

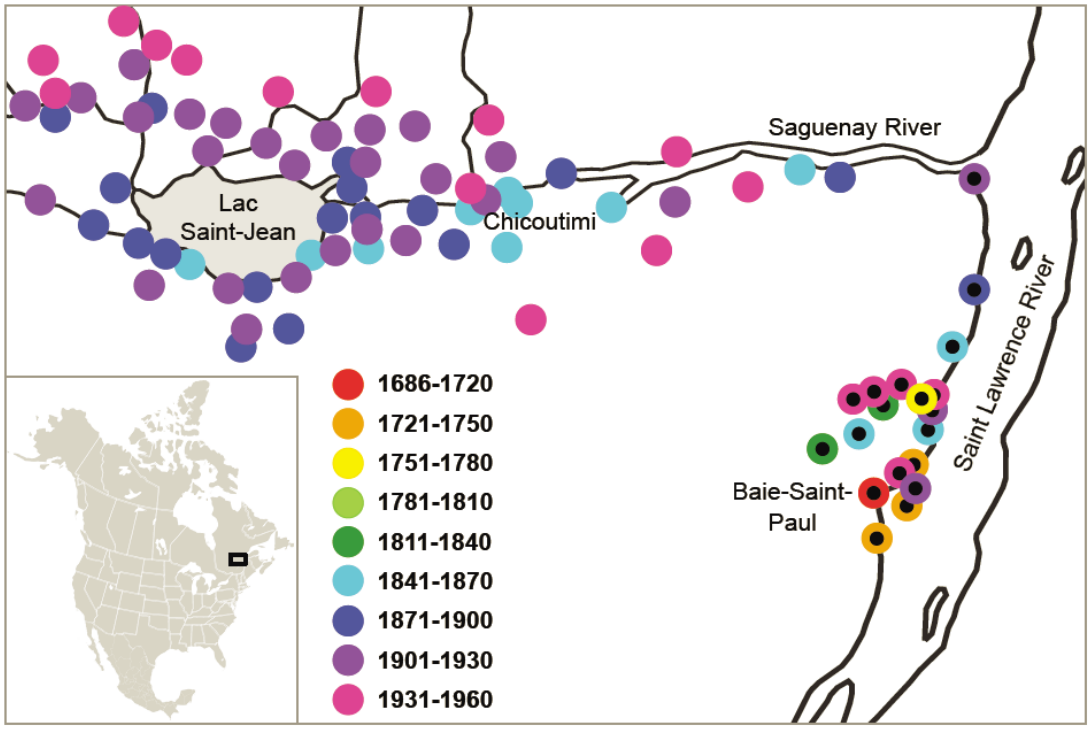

Figure 1. Map of Charlevoix Saguenay Lac-Saint-Jean region showing the range expansion dynamics and the wave front at different periods.

Each filled circle represents a locality and its color indicates its age. Localities from the Charlevoix region are indicated by a black dot.

On the basis of the computation of a wave front index WFI (see Materials and methods), we find that the ancestors of the Saguenay and the Lac-Saint-Jean people lived more often on or close to the wave front than expected by chance (WFI p-value<0.001 in both regions, Figure S1). Indeed the very high WFI of 0.75 observed in Lac-Saint-Jean corresponds to a situation where half of the Lac-Saint-Jean ancestors had lived directly on the wave front and the other half just one generation away from it. In contrast, WFI is significantly lower in the Charlevoix region (p-value=0.003, Fig. S1). These results are consistent with different colonization dynamics of Saguenay Lac-Saint-Jean (SLSJ) and Charlevoix. The wave front was always widespread in SLSJ where new localities were continuously settled, while it was much smaller in Charlevoix where most localities remained in the range core until the 20th century (see Fig. 1). New immigrants from outside ChSLSJ constituted an important minority of the people getting married, with a greater proportion of immigrants settling on the wave front than on the range core especially prior to 1900 (up to 20% on the wave front and up to 10% in the range core, Table S2). Generally more male than female immigration occurred in all regions, and this bias towards males is significantly higher in the core than on the wave front (Table S3). Nevertheless, the new territories of SLSJ have been largely colonized by people recruited directly on the wave front or next to it, and not by people from the range core (Table S4).

Table 1. Genetic contribution (GC) of ancestors having lived in the ChSLSJ region to individuals from the 1931-1960 generation found anywhere in the Quebec province.

	Wave front (WF)			Range core (RC)				
Generation	Total GC	No. of ancestors in genealogy	Mean GC	Total GC	No. of ancestor in genealogy	Mean GC	% ancestors on wave front	**Mean GC Ratio (WF/RC)**
	ChSLSJ							
1686-1720	19298	48	402.0	612	6	102.0	88.9	3.94***
1721-1750	19263	104	185.2	16833	106	158.8	49.5	1.17 *
1751-1780	22119	196	112.9	25990	373	69.7	34.4	1.62***
1781-1810	21696	364	59.6	35613	1069	33.3	25.4	1.79***
1811-1840	30504	1383	22.1	27061	1815	14.9	43.2	1.48***
1841-1870	56589	6555	8.6	10175	2438	4.2	72.9	2.07***
1871-1900	40386	8757	4.6	25619	8784	2.9	49.9	1.58***
1901-1930	23370	10034	2.3	44408	26255	1.7	27.7	1.38***
Total ChSLSJ		*27441*			*40846*		*40.2*	
	SLSJ							
1841-1870	27833	3743	7.4	39	15	2.6	99.6	2.8***
1871-1900	33917	7300	4.6	15444	4420	3.5	62.3	1.3***
1901-1930	21061	8832	2.4	35777	19726	1.8	30.9	1.3***
Total SLSJ		*19875*			*24161*		*45.1*	

*: p<0.05; ***p<0.001
Note that GCs of different generations are not independent

We computed the expected number of genes left by a given ancestor to the current generation [its genetic contribution (GC)] (Bhérer et al. 2011) of all ancestors of ChSLSJ, distinguishing between those having reproduced on the wave front and those in the range core (Table 1). We find that over the entire

studied period individuals on the front have contributed significantly more genes to the present generation than those in the core, in line with theory predicting that surfing alleles should be traced back to ancestors living on or close to the wave front (Hallatschek and Nelson 2008). We find similar results when we restrict the analysis to the SLSJ region (Table 1), which has been colonized more recently. Overall, ancestors on the edge contributed 1.2 to 3.9 times more genes to the current generation than ancestors from the core, the oldest ancestors generally passing on more genes than more recent ones, in keeping with previous results (see, e.g., Fig. 4 in Bhérer et al., 2011). In addition, 40.2% of all ancestors of the ChSLSJ living between 1686 and 1930 were on the wave front, reaching 45.1% for the SLSJ region (Table 1). For SLSJ, the number of ancestors living directly on the front or just one generation away from it even reaches 81% (Table S4), showing the importance of this moving edge for this region.

Table 2. Age of reproduction and number of children of women from SLSJ in the period 1840-1900.

	No. of women	Mean no. of children (FS)	Mean no. of married children (EFS)	Mean age at marriage	FS ratio WF/RC	EFS ratio WF/RC	Marriage age ratio WF/RC
Wave front (WF)	2663	9.1	4.9	20.5	1.15***	1.20***	0.95 ***
Range core (RC)	1783	7.9	4.1	21.6			

*** *t*-test of difference between means, p<0.001
Note that this table only includes women with known birth date, such that age at marriage can be computed.

We compared the reproductive success of women on the edge to the ones in the core, considering both the number of their children (family size, FS) and the number of their married children (effective family size, EFS). SLSJ female ancestors living on the edge had on average 15% more children than core SLSJ female ancestors (Table 2, *p*-value<0.001) and even 20% more married

children (p-value <0.001). These results show that women's fertility was significantly higher on the wave front than in the range core, and that the larger genetic contribution of ancestors reproducing on the wave front is likely not purely due to a neutral surfing process, but also to a net effect of positive selection on the front.

Women on the front overall had a slightly higher modal FS value (Fig. S2A) and larger EFS due to a right shift of the whole distribution towards higher values (Fig. S2B), leading to a larger proportion of women on the front having more than 5 married children (40% on the front vs. 26% in the core).We find only a slightly lower mortality rate of children under the age of five on the wave front (23.6% vs. 25.1% in the core), implying that the increase in EFS compared to FS on the front is likely due to facilitated access to reproduction (marriage). Interestingly, women on the front married almost one year earlier than women in the core (Table 2), increasing their reproductive life, which may partly explain their overall higher fertility. This is in line with Charbonneau's observations (Charbonneau et al. 2000) that Quebec women had an overall longer reproductive life compared to French women of the same period, due to both an earlier age at first child and a later age for their last child. However, our results suggest that the larger fertility of women in Quebec is mainly due to a front effect. An analysis of covariance reveals that the number of children per women actually depends significantly both on the age of marriage and on the spatial location of reproduction (front or core) (p<0.001 for the two effects), but that there is no interaction between these factors (p=0.46). For the number of married children, the two factors (p<0.001) and their interaction (p=0.02) are significant. We conclude that even though women on the front reproduce earlier than women from the core, this contrast does not fully explain the difference in their fertility. Note that the advantage of being on the wave front remains if we use less informative criteria to assign individuals to the front in SLSJ (see Tables S6, and S7), and that it is not due to a higher fertility of new immigrants settling preferentially on the wave front (Table S8).

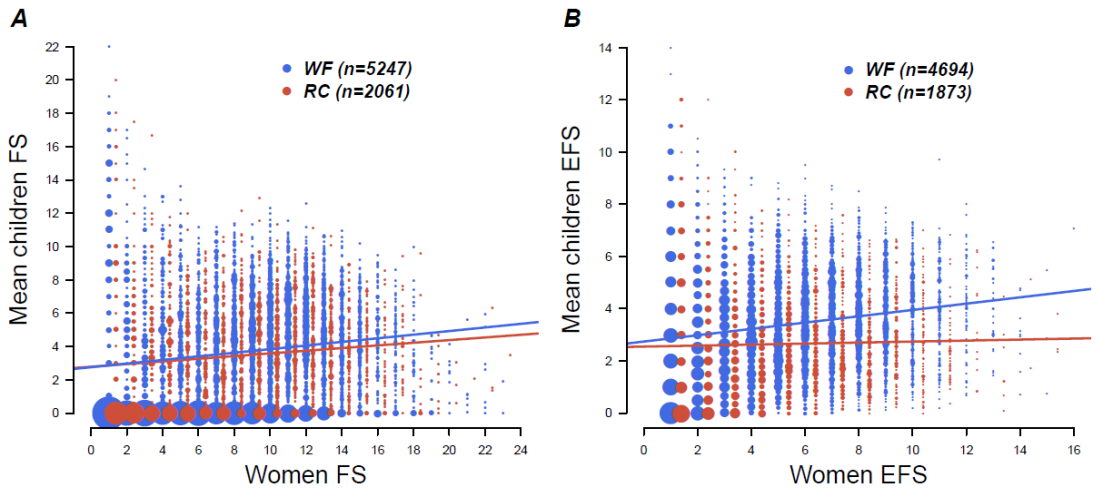

Figure 2. Intergenerational correlation in family size in SLSJ between 1840 and 1900.

(A) Family size (FS) or number of children per woman. WF: r=0.15, h^2=0.22, p-value<0.001; RC: r=0.12, h^2=0.16, p-value<0.001. **(B)** Effective family size (EFS) approximated by the number of married children per woman. WF: r=0.18, h^2=0.24, p-value<0.001; RC: r=0.027, h^2=0.04, p-value=0.23.

We compared the fertility of women to the average fertility of their offspring (Austerlitz and Heyer 1998), using the fact that the regression slope b gives us directly a measure of heritability as $h^2=2b$ (Falconer and McKay 1996). A women's FS is correlated with that of her own children on the wave front, with a non-significantly different heritability on the wave front (Fig. 2A, h^2=0.22) and in the range core (h^2=0.16) (ancova test of slope difference, p-value=0.07). In contrast, if we consider the trans-generational correlation in EFS, the correlation is only significant on the wave front (Fig. 2B, h^2=0.24, ancova test of slope, p<0.001) and not in the range core (h^2=0.04, ancova test of slope, p=0.23). Note that the significant heritability on the front is not due to the larger number of women-children comparison on the SLSJ front (4694 vs. 1873 in the range core, see Fig. 2B), as heritability on the front was always significant among 1000 bootstrap regressions performed on 1873

women-children comparisons (Fig. S3). This suggests that the absence of EFS heritability in the range core and its preservation on the wave front are due to post-zygotic effects (i.e. environmental, economic, social or behavioral) affecting the viability of the children and their access to partners. In other words, women on the wave front and in the range core have the same potential to produce children, but the possible extrinsic factors that condition their reproductive fitness are preserved between generations on the wave front but not on the core.

The larger genetic contribution of individuals at the wave front could be considered as a long-term selective advantage of these individuals due to surfing events having effects similar to selective sweeps (Excoffier et al. 2009), but our results suggest that selective processes based on differential reproduction may also be directly involved. Indeed, we evidence that women's fitness measured by EFS and its heritability are significantly higher on the wave front (Table 2, Fig. S2). More precisely, the heritability of FS is similar over the whole SLSJ, but the heritability of EFS is only significant and high (h^2=0.24) on the wave front, suggesting that the same fertility potential can only be transmitted over several generations on the wave front. This may be due to the fact that individuals belonging to large families had more difficulty to have access to marriage and reproduction than individuals from small families in the range core. This trade-off between fertility and children access to reproduction would explain the absence of EFS correlation between generations in the range core, in contrast to the wave front. The lowering of the age at first reproduction (as measured by age at first marriage) is the life-history trait that most effectively increases reproductive rate (Lewontin 1965). It has been shown to have evolved in several invading species (Phillips et al. 2010), but also in human populations. For instance, it recently decreased in an Ethiopian agro-pastoralist society with facilitated access to cultivable land (Gibson and Gurmu 2011), as well as in a small and isolated population from Quebec over the period 1800-1939 (Milot et al. 2011) potentially due here to positive selection. The evolution of population growth rate in SLSJ thus

follows the prediction that high growth strategies should evolve on an expansion wave front as they are not hindered by density regulation occurring in the range core (Phillips et al. 2010).

The fact that EFS is correlated between generations has been shown to lead to highly unbalanced gene genealogies and to an increase in the frequency of rare variants over a few generations (Austerlitz and Heyer 1998; Sibert et al. 2002), which may resemble the action of selection. However, at odds with previous results (Austerlitz and Heyer 1998), we find that the heritability of EFS does not occur in the whole SLSJ region but is restricted to the wave front, and is thus potentially one of its important properties. It remains to be seen if fitness heritability occurred in other human range expansions. If it was the case, it could mean that allele frequency changes in spatially expanding populations could have been even more drastic than expected under pure surfing.

SUPPORTING MATERIAL

Material and Method

Data

Individual information

The BALSAC genealogical database of Quebec (http://balsac.uqac.ca) covers all regions of Quebec and contains information on about 5 million individuals spreading between the beginning of the 17th century up to the 1970's. This database has been fruitfully used in several studies linking genetics and demography (e.g. Heyer et al. 1997; Austerlitz and Heyer 1998; Heyer et al. 2001; Moreau et al. 2011b; Roy-Gagnon et al. 2011). For the purpose, of this study we extracted genealogical information from the 88,157 marriages that were recorded in the Charlevoix Saguenay Lac-Saint-Jean (ChSLSJ) region between 1686 and 1960. We traced the descendents of all these marriages, which amounts to 1,294,367 individuals. Among these, 367,947 individuals married in Québec outside ChSLSJ, and 761,049 individuals either never got married, married outside Quebec, or married after 1960. In addition to their genealogical connections, we got information for each of these individuals on the date and location of their marriage if it occurred in Quebec. For the individuals of SLSJ, we also got information on their place and year of birth and we could compute their total number of children from genealogical information. We also obtained the age at death of individuals deceased in the SLSJ region between 1840 and 1970. For some analyses (such as reported in Table 1), we allocated individuals to non-overlapping 30 years intervals going backward in time and starting in 1960 (see Table 1). Note that 30 years corresponds approximately to the duration of a human generation (Tremblay and Vézina 2000). Individual allocation to a given generation was then made on the basis of marriage date.

Wave front definition and proxy for the place of reproduction

The date of the settlement of a new locality was estimated as the date of the first marriage celebrated in a parish of this locality. Twenty new localities were founded in Charlevoix between 1686 and 1953, and 64 localities were founded in Saguenay-Lac-Saint-Jean between 1842 and 1957. The dynamics of the settlement of these localities can be seen on Figure 1. The wave front of the colonization of the ChSLSJ region was defined at any time as the set of all localities settled in the last 30 years (1 generation ago). An individual was then assigned to the wave front if it reproduced on the front, and to the range core if it did not. The place of reproduction of an individual was estimated in three different ways: 1) the place of birth of the majority of his/her children; 2) the place of first marriage of the majority of his/her children; 3) the place of his/her first marriage that led to descendants. The quality of the estimation decreases from 1) to 3), but we always used the best quality estimate for a given individual depending on the available information (see Table S1). The best quality information was then used to define if the individual reproduced on the front or in the core. In case of ties in procedures 1) or 2), individuals were discarded if there was a conflict on imputed wave front allocation, which overall represented less than 5% of all married individuals (see Table S1).

For Table S4, we also computed the distance (in generations) between the individuals and the front. We used a procedure to allocate individuals to a given generation away from the front that is similar to that described to infer the front status of the individuals. We also removed individuals with ambiguities in front status, but we kept individuals with several possible reproduction localities away from the front (506 cases out of 44036 ancestors). In those cases, we applied the following procedure 1000 times:for each ambiguous individual, we chose a reproduction site at random among the possible alternatives, and then recomputed the distribution of distances between the individuals and the wave front. The final distribution was obtained by averaging over the 1,000 random distributions.

Computations

Genetic contribution of ancestors

In this study, we define the genetic contribution GC (James 1972) of an individual to a given generation as the expected number of gene copies ascending to that individual from that generation. GC is measured by examining all potential transmission paths between an individual and his descendents and summing over its transmission probabilities (e.g. Bhérer et al. 2011). It measures the genetic impact of an individual to future generations, which can be considered as a long-term measure of the genetic success of an individual. The genetic contribution of all the ancestors married in ChSLSJ between 1686 and 1930 to the last studied generation, which consisted in all individuals married in ChSLSJ between 1931 and 1960, was calculated with the TIBCO Spotfire S+ package GenLib 8.4.18 (available on http://balsac.uqac.ca).

Wave front index

In order to test if the ancestors of the individuals married in a given region had been reproducing more on the wave front than expected by chance, we computed a wave front index (*WFI*) and obtained its null distribution under the hypothesis of random reproduction site location. *WFI* for a given ancestor *i* is simply obtained as $WFI_i = 1/(g_i + 1)$ where g_i is the number of generations since the locality where the individual reproduced was settled (in numbers of generations, assuming a generation time of 30 years). For instance, WFI_i has a value of 1 if individual *i* reproduced in a recently settled locality (located on the front), and 0.5 if (s)he reproduced in a locality settled one generation ago. Note that front allocation of the individuals was inferred by following the rules mentioned earlier (major place of birth of children, major place of marriage of children, or place of marriage). *WFI* is then averaged over all the ancestors of individuals married in a given region in the last generation 1931-1960, even if these ancestors lived in different regions of ChSLSJ. *WFI* was computed

separately for the 3 regions Charlevoix, Saguenay, and Lac-Saint-Jean. The null distribution of *WFI* was obtained by randomly assigning ancestors' reproduction site to any locality settled at the time of their marriage in ChSLSJ, with a probability proportional to the number of marriages having occurred in that locality, and recalculating *WFI* on this randomized set of reproduction sites. The null distributions were then obtained by repeating this procedure 10,000 times, and the results for each region are presented in Figure S1.

Fertility and its heritability

The fertility of each woman was measured either as the total number of her children (family size, FS) or as the number of her married children (effective family size, EFS, used as a proxy for the number of reproducing children). If a woman was married several times, we used all her descendents, but we used the date of her first fertile marriage to determine her position relative to the front (and to assign her to the wave front or range core, see above). FS was only computed in the SLSJ since we did not have access to the unmarried children born or deceased outside this region. For consistency reasons, FS and EFS were computed on all SLSJ women married in the period 1840-1900, as SLSJ colonization started around 1840 and to ensure that all women children would marry before 1960 which marks the end of our parish records.

The narrow-sense heritability (h^2) of fertility (FS or EFS) was computed by regressing the average fertility of offspring (men and women) on the mother fertility and using the fact that in this case the slope of the regression $b=h^2/2$ (see e.g. Falconer and McKay 1996).

Supporting Figures

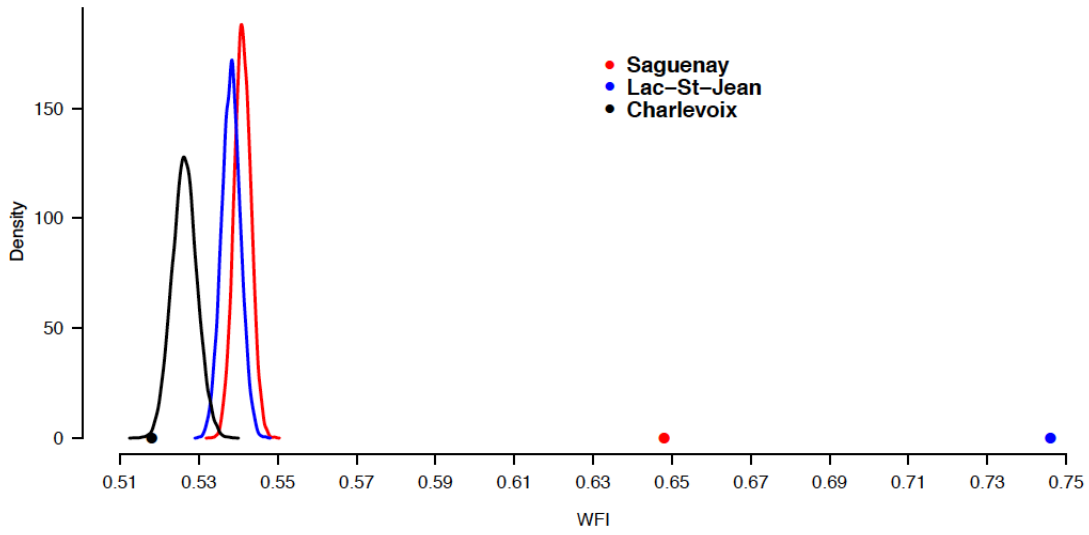

Figure S1. Empirical null distributions (lines) of the Wave Front Index (WFI) and observed values (filled circles) in Saguenay, Lac Saint-Jean, and Charlevoix.

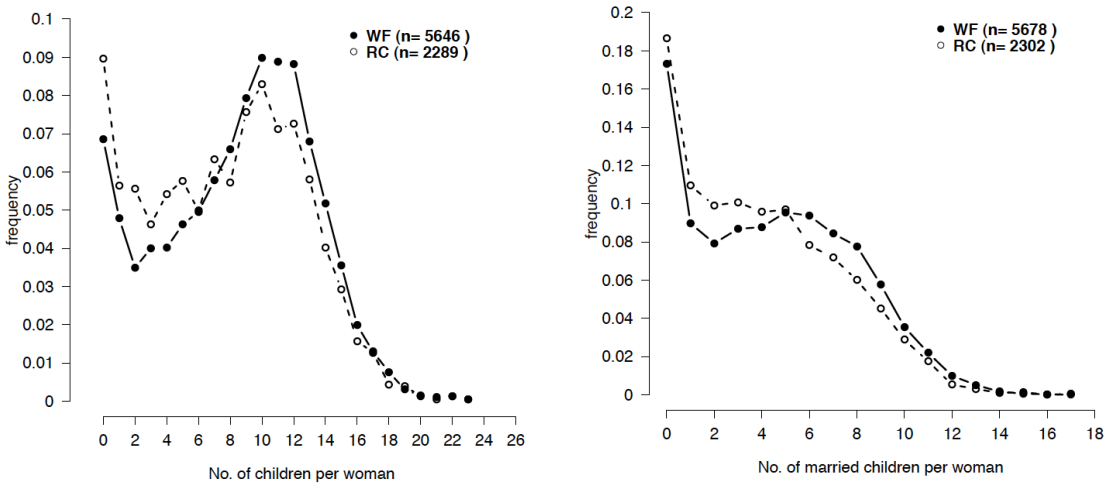

Figure S2. Family size distributions in SLSJ between 1840 and 1900. We contrast the number of children per woman (family size, FS) and the number of married children per woman (effective family size (EFS) on the wave front (WF) and in the range core (RC).

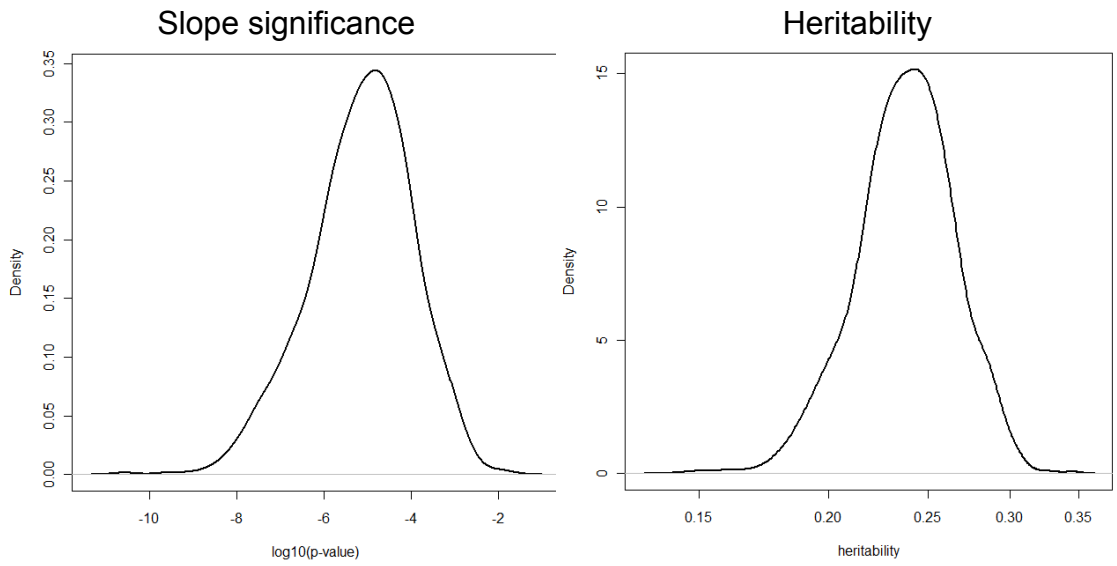

Figure S3. Bootstrap distributions of regression analyses between the EFS of women and that of their children.

In order to see if the significant heritability of EFS on the wave front is not due to a larger sample size (n=4694) than in the range core (n=1873), we performed 1000 regressions of EFS on 1873 women-children pairs taken at random from those available on the front, and computed each time the significance of the regression slope b and the heritability $h^2=2b$.

Supporting Tables

Table S1. Determination of the wave front status for individuals married between 1686 and 1960.

Regions	Number of married individuals	Fraction of individuals with defined front status	Criterion used to define reproduction place		
			Children main place of birth	Children main place of marriage	Place of own marriage
ChSLSJ	165371	95.2	67.9	10.1	17.2
Charlevoix	40909	86.9	11.3	39.4	36.2
SLSJ	124462	97.9	86.5	0.4	11.0
		Fraction of individuals with front status undefined	Cause of absence of definition		
			Conflict among places of birth	Conflict among places of marriage	Children married out of ChSLSJ
ChSLSJ	165371	4.8	1.4	0.6	2.8
Charlevoix	40909	13.1	0.2	2.5	10.5
SLSJ	124462	2.1	1.8	0.0	0.3

Table S2. Proportion of immigrants among married individuals between 1686 and 1960.

	Immigration into ChSLSJ		Immigration into SLSJ from outside ChSLSJ		Immigration into SLSJ	
1686- 1900	WF	RC	WF	RC	WF	RC
All individuals	0.163	0.099	0.197	0.096	0.704	0.405
Women	0.073	0.041	0.087	0.040	0.344	0.186
Men	0.090	0.058	0.109	0.056	0.359	0.219
1900 - 1930	WF	RC	WF	RC	WF	RC
All individuals	0.212	0.165	0.223	0.186	0.328	0.272
Women	0.096	0.066	0.102	0.077	0.152	0.117
Men	0.116	0.100	0.122	0.110	0.176	0.155
1930 - 1960	WF	RC	WF	RC	WF	RC
All individuals	0.255	0.170	0.270	0.175	0.305	0.212
Women	0.105	0.062	0.113	0.065	0.128	0.082
Men	0.150	0.108	0.157	0.110	0.177	0.130

Table S3. Test for sex differences in immigration rates between the wave front (WF) and the range core (RC) for the period 1686-1960.

	Number of immigrants into ChSLSJ		Number of immigrants into SLSJ		Number of immigrants into SLSJ from outside ChSLSJ	
1686-1900	WF	RC	WF	RC	WF	RC
Women	1271	595	3801	824	963	176
Men	1567	845	3969	966	1209	247
$\chi 2$ test p-value	<0.0001	<0.0001	0.0567	0.0008	<0.0001	0.0006
Fisher exact test p-value	0.031		0.029		0.309	
1900-1930	WF	RC	WF	RC	WF	RC
Women	963	1726	1348	2309	899	1511
Men	1168	2620	1556	3054	1077	2170
$\chi 2$ test p-value	<0.0001	<0.0001	0.0001	<0.0001	<0.0001	<0.0001
Fisher exact test p-value	<0.0001		0.0034		0.0013	
1930-1960	WF	RC	WF	RC	WF	RC
Women	1413	4686	1508	5434	1333	4315
Men	2020	8153	2083	8565	1850	7267
$\chi 2$ test p-value	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
Fisher exact test p-value	<0.0001		<0.0001		<0.0001	

The χ^2 test tests for differences in immigration rates between sexes.
The Fisher exact test tests for different immigrant proportions on the wave front and on the range core.

Table S4. Fraction of SLSJ ancestors living directly or some distance away from the wave front.

Distance from the wave front [a]	Number of ancestors	Percentage of ancestors assigned to the front
0	19875	45.13
1	15694	35.64
2	8256	18.75
3	174	0.40
4	15	0.03
5	8	0.02
6	8	0.02
7	6	0.01
Total	44036	100

[a] The distance from the front is expressed in numbers of generations since the settlement of a given locality. A distance of zero indicates the wave front.

Table S5. Test for sex differences in emigration rates between the wave front (WF) and the range core (RC) for the period 1686-1960.

	Number of emigrants from ChSLSJ out of ChSLSJ		Number of emigrants from SLSJ out of SLSJ		Number of emigrants from SLSJ out of ChSLSJ	
1686-1900	WF	RC	WF	RC	WF	RC
Women	1025	253	650	34	616	18
Men	1035	494	545	15	509	7
χ2 test p-value	0.8256	<0.0001	0.0023	0.0066	0.0014	0.0278
Fisher exact test p-value	<0.0001		0.0408		0.1044	
1900-1930	WF	RC	WF	RC	WF	RC
Women	1094	612	933	392	911	356
Men	1449	860	1207	464	1118	411
χ2 test p-value	<0.0001	<0.0001	<0.0001	0.0139	<0.0001	0.0470
Fisher exact test p-value	0.3891		0.2898		0.4958	
1930-1960	WF	RC	WF	RC	WF	RC
Women	429	705	410	633	398	557
Men	650	1057	625	1004	575	826
χ2 test p-value	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
Fisher exact test p-value	0.9058		0.6255		0.7656	

The χ^2 test for differences in immigration rates between sexes.
The Fisher exact test tests for different immigrant proportions on the wave front and on the range core.

Table S6. Genetic contribution (GC) of ancestors having lived in the ChSLSJ region to individuals from the 1931-1960 generation found anywhere in the Quebec province.

	Wave front (WF)			Range core (RC)				
Generation	Total GC	No. of ancestors in genealogy	Mean GC	Total GC	No. of ancestor in genealogy	Mean GC	% ancestors on wave front	Mean GC Ratio (WF/RC)
Charlevoix-Saguenay-Lac-Saint-Jean								
1686-1720	19298	48	402.0	612	6	102.0	88.9	3.9***
1721-1750	19263	104	185.2	16833	106	158.8	49.5	1.2*
1751-1780	22119	196	112.9	25990	373	69.7	34.4	1.6***
1781-1810	21696	364	59.6	35613	1069	33.3	25.4	1.8***
1811-1840	29862	1324	22.6	27197	1835	14.8	41.9	1.5***
1841-1870	53347	5506	9.7	10631	2588	4.1	68.0	2.4***
1871-1900	40297	7879	5.1	21937	7721	2.8	50.5	1.8***
1901-1930	22973	9372	2.5	41242	24415	1.7	27.7	1.5***
Total CSLSJ		24793			38113		39.4	
Saguenay-Lac-Saint-Jean								
1841-1870	25941	3053	8.5	114	46	2.5	98.5	3.4***
1871-1900	34040	6570	5.2	11874	3361	3.5	66.2	1.5***
1901-1930	20731	8217	2.5	32730	17970	1.8	31.4	1.4***
Total SLSJ		17840			21377		45.5	

This table differs from Table 1 by assigning individuals to the wave front or range core without considering the main place of birth of their children, but only the date and place of marriage of their children or of their own marriage (criteria 2 and 3, respectively), in order to have more comparable information between the SLSJ and Charlevoix regions.

*: p<0.05; ***p<0.001

Note that GCs of different generations are not independent

Table S7. Age of reproduction and number of children of individuals from SLSJ in the period 1840-1900.

	No. of women	Mean no. of children (FS)	Mean no. of married children (EFS)	Mean age at marriage	FS ratio WF/RC	EFS ratio WF/RC	Marriage age ratio WF/RC
Wave front (WF)	2408	9.1	5.2	20.5	1.21***	1.30***	0.94***
Range core (RC)	1370	7.5	4.0	21.8			

This table differs from Table 2 by assigning individuals to the wave front or range core without considering the main place of birth of their children, but only the date and place of marriage of their children or of their own marriage (criteria 2 and 3, respectively), in order to have more comparable information between the SLSJ and Charlevoix regions.

*** *t*-test of difference between means, $p<0.001$
Note that this table only includes women with known birth date, such that age at marriage can be computed.

Table S8. Age of reproduction and number of children of SLSJ individuals for the period 1840-1900, considering immigrants and non-immigrants separately.

	No. of women	Mean no. of children (FS)	Mean no. of married children (EFS)	Mean age at marriage	FS ratio WF/RC	EFS ratio WF/RC	Marriage age ratio WF/RC
Immigrants [a]							
Wave front (WF)	54	9.4	4.6	20.4	1.11 NS	0.94 NS	0.96 NS
Range core (RC)	16	8.5	4.9	21.3			
Non-immigrant[a]							
Wave front (WF)	2609	9.1	4.9	20.5	1.15***	1.20***	0.95***
Range core (RC)	1767	7.9	4.1	21.6			

[a] Individuals were considered a new immigrant if they came from a region outside ChSLSJ
This table is analogous to Table 2
*** *t*-test of difference between means, $p<0.001$, NS : Not significant
Note that this table only includes women with known birth date, such that age at marriage can be computed.

CHAPITRE VI: Discussion

LA DÉMOGRAPHIE D'UN PEUPLEMENT FONDATEUR
ET SES CONSÉQUENCES ÉVOLUTIVES

DISCUSSION ET PERSPECTIVES

Les processus de colonisation de nouveaux territoires tels que les événements fondateurs et les expansions territoriales sont à l'origine de la formation d'un grand nombre de populations contemporaines, dont les Canadiens français du Québec. Dans cette thèse, j'ai contribué à construire un échantillon de la variation génétique et à caractériser en profondeur le patrimoine génétique du Québec et de ses régions. De plus, j'ai documenté certains processus démographiques gouvernant les phénomènes de colonisation ainsi que leurs conséquences évolutives. Le défi méthodologique était de taille : utiliser les généalogies de la population du Québec pour réaliser des analyses de génétique des populations. Ce chapitre de discussion est une occasion de proposer un nouveau portrait synthèse du patrimoine génétique du Québec. Je discute ici des limites et du potentiel des ensembles de données, je propose une solution à la question de la diversité génétique des Canadiens français et je caractérise un nouveau portrait de la structure de la population du Québec. Enfin, je discute des retombées de nos résultats dans un contexte plus large, d'une part pour la compréhension du rôle des peuplements fondateurs au cours de l'évolution humaine, d'autre part pour les études d'épidémiologie génétique.

1. Un échantillon génétique et généalogique du Québec

Échantillon de référence du Québec

Dans le chapitre III, nous avons présenté les données génomiques et généalogiques de 140 individus provenant de l'Échantillon de référence du Québec. Cet échantillon a été construit dans le cadre d'un vaste programme de recherche ayant pour objectif d'échantillonner et de caractériser la diversité génétique des populations régionales du Québec et de la lier aux généalogies reconstruites à l'aide du fichier BALSAC.

Le recrutement des participants est certainement un des aspects les plus limitatifs et laborieux de ce type de projet. J'ai pu y contribuer pour les régions de Lanaudière, de Montréal, de l'Outaouais et de la Ville de Québec. Dans ces régions, le recrutement a notamment fait appel à des contacts établis lors d'une étude précédente (Loggia et al. 2009), des contacts personnels et à la participation de membres de sociétés de généalogies. Les participants inclus dans l'étude ont leurs ancêtres présents dans une région donnée depuis au moins deux générations, avec trois ou quatre grands-parents mariés et/ou nés dans cette même région. Cette stratégie d'échantillonnage nous a permis de documenter soigneusement la provenance et les origines des participants. Cette stratégie a aussi été adoptée par un projet britannique de grande envergure nommé « People of the British Isles » (Winney et al. 2012).

Étant donné la nature du recrutement, l'échantillon n'offre pas un portrait représentatif de l'ensemble de la population du Québec d'aujourd'hui, mais peut plutôt être considéré comme un échantillon de référence, ou une collection, du Québec historique. À ce titre, l'Échantillon de référence du Québec est complémentaire au projet CARTaGENE, où les participants sélectionnés sont des résidents du Québec appartenant à une tranche d'âge précise, sans égard à leurs origines (Awadalla et al. 2013). L'ancrage géographique des participants à notre étude permet de réaliser des analyses de grande qualité sur l'évolution et la structure fine de la population, en excluant notamment les effets des migrations récentes. L'Échantillon de référence du Québec, qui est disponible pour la communauté scientifique (www.quebecgenpop.ca), a un important potentiel de recherche, autant en génétique des populations qu'en épidémiologie génétique. Par exemple, il pourrait être utilisé pour estimer la distribution et la fréquence de certains polymorphismes ou servir de référence pour inférer l'origine géographique ou ethnique la plus probable de sujets d'étude dont l'origine est inconnue.

L'échantillon généalogique de l'ensemble du Québec

Deux articles, présentés aux chapitres II et IV, ont été réalisés avec l'échantillon généalogique de l'ensemble du Québec qui provient du fichier de population BALSAC. Cet échantillon offre une perspective globale de la population puisqu'il couvre l'ensemble du territoire et est représentatif de la distribution de la population au recensement canadien de 1956. Les résultats de ces articles se basent donc sur le Québec contemporain mais non actuel. Puisqu'il s'agit de sujets mariés dans une paroisse catholique entre 1945 et 1965, on peut estimer qu'ils représentent les grands-parents des jeunes adultes d'origine canadienne-française de la population actuelle (en supposant 30 ans comme intervalle intergénérationnel moyen). La période de mariage des individus sélectionnés précède donc les changements démographiques importants qui ont eu lieu depuis 50 ans. Ainsi, de nouvelles études, réalisées avec des échantillons représentatifs de l'ensemble des habitants du Québec, demeurent nécessaires pour appréhender le patrimoine génétique du Québec d'aujourd'hui dans toute sa globalité et complexité. Néanmoins, en fournissant un portrait détaillé du Québec tel qu'il l'était au milieu du 20^{e} siècle, nos travaux pavent la voie aux études de l'évolution récente du patrimoine génétique. Notamment, on peut se demander si l'intensification des mouvements migratoires internationaux et interrégionaux observés au Québec depuis 50 ans ont bouleversé la structure de la population, tel qu'il a été montré dans la région de Lanaudière (Bhérer et al. 2008). Comme ces mouvements récents sont largement liés au phénomène d'urbanisation, on s'attend en effet à ce que les régions près des grands centres urbains aient été plus touchées que les régions périphériques (Helgason et al. 2005; Ashrafian-Bonab et al. 2007; Bhérer et al. 2008).

2. L'effet fondateur et la diversité génétique au Québec

Au cœur de cette thèse se trouve une question maintes fois soulevée, mais demeurée ouverte : comment expliquer que la population fondatrice

canadienne-française du Québec ait un niveau de diversité génétique similaire à une grande population telle que la France, alors que son bagage particulier de maladies mendéliennes témoigne de l'effet fondateur et de l'homogénéité qu'il entraine? La question de la diversité génétique des Canadiens français est intrigante puisque les études menées à ce jour suggèrent que la population, dans son ensemble, a échappé à une des principales prédictions de l'effet fondateur : la perte de diversité. Pour chercher une explication à cette apparente contradiction, nous sommes d'abord remonté aux origines de la population du Québec.

Un pool de fondateurs suffisamment large et hétérogène

Une des principales hypothèses proposées pour expliquer les niveaux de diversité des Canadiens français est que le pool de fondateurs de l'ensemble du Québec ait été suffisamment large et hétérogène pour limiter la perte de diversité génétique (Bouchard and De Braekeleer 1990; De Braekeleer 1990; Bouchard and De Braekeleer 1991b; Gagnon and Heyer 2001; Moreau et al. 2007). Au chapitre II, nous avons étudié en détail l'immigration fondatrice, incluant non seulement les fondateurs bien documentés de la Nouvelle-France, mais aussi ceux arrivés après la Conquête britannique de 1760.

Nos résultats réaffirment la prépondérance de la France dans les origines fondatrices des Canadiens français. Dans l'échantillon généalogique de l'ensemble du Québec, les Français représentent 70% des immigrants fondateurs et ont contribué pour 90% du génome canadien-français, en accord avec les estimations précédentes (Charbonneau et al. 1987; Charbonneau et al. 2000; Vézina et al. 2005b; Bergeron et al. 2008; Bhérer et al. 2008; Tremblay et al. 2009; Tremblay 2010). Ces proportions sont sensiblement les mêmes d'une région à l'autre sauf dans l'Est où la contribution génétique des Français est réduite à 70%. D'un côté, ces observations peuvent être interprétées comme indicatrices d'homogénéité. Cependant, on sait que les fondateurs français sont venus en majorité seuls plutôt qu'en famille et en provenance de toutes les régions de la France

(Charbonneau et al. 1987; Guillemette and Légaré 1989; Vézina et al. 2005b). De plus, notre analyse globale des fondateurs souligne que le fait de concentrer son interprétation sur une majorité de fondateurs d'origine française amène à négliger une part substantielle (10%) du génome canadien-français.

En effet, nous avons montré que virtuellement tous les Canadiens français sont métissés. Ceci est certainement un de nos résultats les plus influents pour interpréter l'impact de l'histoire démographique sur la diversité de la population. Avec l'échantillon généalogique du Québec, nous avons estimé que les Canadiens français ont en moyenne 6,5 origines ancestrales distinctes, incluant entre autres des fondateurs acadiens, britanniques, allemands et autochtones. Nous avons fourni des estimations de l'apport génétique des autochtones, qui est depuis longtemps un sujet de spéculation et même de controverse (Vézina et al. 2012). Nous avons montré qu'au moins la moitié des Canadiens français ont des origines amérindiennes documentées dans leurs généalogies. Nos estimations généalogiques ont évalué à 0,2% la contribution génétique amérindienne dans l'ensemble de la population. Dans une étude subséquente réalisée avec les données génomiques de l'échantillon de référence du Québec, nous avons estimé à 1% la proportion du génome canadien-français d'origine autochtone (Moreau et al. 2013). Les contributions d'autres groupes de fondateurs sont aussi significatives (Chapitre II, Bergeron et al. 2008; Tremblay et al. 2009; Tremblay 2010). Ainsi, des fondateurs d'origine autre que française se sont métissés avec les Canadiens français tout au long des 400 ans d'histoire du Québec. Au chapitre IV, nous avons montré qu'un immigrant anglais du 17^{e} siècle figure parmi le top six des couples fondateurs ayant eu le plus important succès reproducteur au Québec. Ces résultats vont à l'encontre d'un isolement génétique de la population et démontrent, au contraire, que les Canadiens français forment un peuple métissé depuis la période de la Nouvelle-France.

L'apport non français au patrimoine génétique des Canadiens français a sans aucun doute enrichi leur diversité génétique, tant au niveau de l'hétérozygotie que de la richesse allélique. Comme une petite fraction du génome canadien-français provient de chacune des origines autres que françaises (ex. contribution acadienne de 4%, britannique de 1,8%, autochtone de 1%, irlandaise de 0,9%), on peut s'attendre à ce que ce métissage ait surtout enrichit la diversité en allèles rares. Il est toutefois difficile de traduire les estimations généalogiques du métissage en valeurs génomiques concrètes. Des études génomiques seront nécessaires pour évaluer directement l'ampleur du métissage et sa contribution à la diversité canadienne-française. Par exemple, les données de séquençage, qui sont de plus en plus disponibles au Québec et ailleurs, devraient permettre d'estimer pour un génome individuel quel est le nombre d'allèles rares ou communs qui provient en moyenne d'une origine donnée.

Ceci étant dit, notre étude de simulation au chapitre IV démontre que même si tous les fondateurs provenaient de la même population source, leur effectif dans l'ensemble du Québec a été suffisamment grand pour empêcher la perte des allèles dont la fréquence initiale parmi les fondateurs était de 1% ou plus. Ces allèles, que l'on peut définir comme communs, devraient avoir survécu dans la population jusqu'à aujourd'hui. Ce résultat prédit donc un partage quasi parfait des allèles communs parmi les Canadiens français et les Français, qui, rappelons-le, ont contribué pour 90% du génome dans l'ensemble du Québec. Nos simulations montrent aussi que le spectre de fréquences alléliques n'a pas connu de changements majeurs dans l'ensemble du Québec. Ceci explique donc les résultats d'une des premières études qui a suscité une remise en question du paradigme d'homogénéité des Canadiens français, soit le partage parfait des allèles principaux du système HLA entre la France et le Québec et leur fréquence similaire (De Braekeleer 1990). De plus, la persistance des allèles communs est conforme aux analyses génomiques récentes qui ont comparé des échantillons de la France et du Québec : une étude de séquençage qui a montré un partage à

99% des allèles dont la fréquence est de 20% ou plus (Casals et al. 2013) ainsi que notre étude au chapitre III qui montre une corrélation de 98% dans la fréquence des allèles communs (fréquence de 5% ou plus).

En somme, nous avons démontré que dans l'ensemble du Québec le nombre de fondateurs a été suffisamment grand pour limiter une perte de diversité et appuient l'hypothèse que les origines diversifiées des fondateurs ont pu enrichir la diversité du Québec, notamment en introduisant des allèles rares.

La régionalisation de l'effet fondateur

Au Québec, il est depuis longtemps documenté que certaines populations régionales présentent des signatures de l'effet fondateur, incluant des indices de diversité génétique réduits et une incidence accrue de certaines maladies mendéliennes autrement rares. De nombreux auteurs ont proposé que cette régionalisation des signatures de l'effet fondateur a été causée par le mode de peuplement des régions (Labuda et al. 1996; Gagnon and Heyer 2001; Scriver 2001; Yotova et al. 2005; Gerbault 2006; Moreau et al. 2007). Cette hypothèse de la régionalisation de l'effet fondateur pourrait expliquer l'apparente contradiction entre les niveaux de diversité et les preuves de l'effet fondateur. Si les conséquences génétiques que l'on attribue aux événements fondateurs sont limitées à certaines régions et ne sont pas retrouvées dans l'ensemble de la population, alors il n'y a pas de raison de s'attendre à une perte de diversité due à un effet fondateur panquébécois. Néanmoins, jusqu'à récemment, l'hypothèse de la régionalisation de l'effet fondateur s'était appuyée essentiellement sur les observations réalisées dans la région du Saguenay-Lac-St-Jean et l'applicabilité de ce paradigme saguenéen[19] aux autres régions du Québec restait à évaluer. Nos travaux ont

[19] Le paradigme saguenéen a été défini par Moreau, Vézina et Labuda essentiellement comme l'ensemble des signatures de l'effet fondateur observées dans cette région. Ils ont montré que le paradigme saguenéen ne peut s'appliquer à l'ensemble du Québec (Moreau C, Vézina H, Labuda D. 2007. Effet fondateur et variabilité génétique au Québec [Founder effects and genetic variability in Quebec]. *Med Sci (Paris)* **23**(11): 1008-1013.).

permis de mieux caractériser l'impact de l'histoire démographique sur les patrons de diversité génétique de l'ensemble du Québec et de ses régions.

Chaque population régionale du Québec a connu un mode de peuplement unique et une histoire démographique subséquente qui lui est propre. À l'aide de l'information contenue dans les données généalogiques, nous avons observé et mesuré directement plusieurs paramètres de ces processus démographiques. Des différences interrégionales notables ont été observées sur le plan de la profondeur des généalogies, de l'apparentement et de la consanguinité (Chapitre III, Vézina et al. 2004). Nous avons aussi montré que la composition et la contribution génétique de l'immigration fondatrice varient substantiellement d'une région à l'autre (Chapitre II), tel qu'il a été documenté plus en détail pour les fondateurs acadiens (Bergeron et al. 2008), irlandais (Tremblay et al. 2009) et allemands (Tremblay 2010). Tous les fondateurs n'ont pas des descendants dans toutes les régions du Québec et même les fondateurs communs n'ont pas contribué également d'une région à l'autre (Chapitre II). De plus, on observe un nombre variable de fondateurs et d'ancêtres ayant eu des descendants dans les populations régionales contemporaines (Chapitre II, Chapitre IV, Chapitre V). Ces importantes différences interrégionales dans la structure généalogique prouvent que des facteurs démographiques déterminants pour la diversité génétique peuvent varier de façon importante à l'intérieur même d'une population fondatrice.

Nous avons évalué la diversité génétique des populations régionales à l'aide de différents indices reposant sur les données généalogiques et génomiques. Une réduction importante de la diversité n'est retrouvée que dans quelques régions. Au chapitre II, un nombre réduit de fondateurs utiles a été trouvé dans la région de Québec ainsi que dans les régions du Nord-Est et de l'Est. Au chapitre III, l'analyse des segments d'homozygotie (« runs of homozygosity ») a fait ressortir une similitude entre les échantillons des villes de Montréal et de Québec par rapport aux échantillons français du HGDP et CEU de HapMap, alors qu'une plus grande homozygotie a été observée pour

les échantillons de la Gaspésie, de la Côte-Nord et du Saguenay-Lac-St-Jean. Au chapitre IV, nous avons montré que seules les histoires généalogiques du Nord-Ouest et du Nord-Est sont compatibles avec une perte importante des allèles introduits à une fréquence de 1% parmi les fondateurs. De la même façon, d'autres signatures de l'effet fondateur sont limitées à certaines régions, notamment un déficit en allèles rares par rapport à l'hétérozygotie (Chapitre IV), une augmentation du déséquilibre de liaison sur de longues distances (Chapitre III) et une déviation significative du spectre de fréquences alléliques (Chapitre IV). En somme, ces résultats appuient l'hypothèse de la régionalisation de l'effet fondateur et démontrent qu'il est important de considérer la spécificité de peuplement des régions du Québec pour rendre compte de leur diversité génétique.

Nos analyses comparatives permettent d'affirmer que parmi toutes les populations régionales, celles de Charlevoix et Saguenay-Lac-St-Jean ont connu le plus fort effet fondateur. Au sein des sous-populations étudiées, il n'y a que les Acadiens de la Gaspésie qui présentent d'aussi fortes signatures attribuables à l'effet fondateur. Plusieurs explications ont été proposées pour expliquer ces patrons dans Charlevoix et au Saguenay-Lac-St-Jean, dont la figure emblématique est leur bagage de maladies héréditaires rares. D'abord, l'effectif limité de fondateurs dans Charlevoix a pu laisser un net avantage aux premiers pionniers dans la transmission leur bagage génétique (Labuda et al. 1996; Labuda et al. 1997; Yotova et al. 2005). Ensuite, la marche séquentielle du peuplement a pu causer des effets fondateurs successifs, incluant les migrations majeures France > Québec > Charlevoix > Saguenay (e.g. Bouchard and De Braekeleer 1991a). Enfin, la transmission intergénérationnelle des comportements reproducteurs, documentée dans la région, a aussi pu renforcer les conséquences de l'effet fondateur (Tremblay 1997; Austerlitz and Heyer 1998; Austerlitz and Heyer 1999; Austerlitz and Heyer 2000; Heyer et al. 2005). Notre étude de l'expansion démographique dans cette région, présentée au chapitre V, démontre que la dynamique démographique du peuplement, où les

colonisateurs ont été recrutés en majorité directement sur le front de l'expansion, combinée à une plus grande fécondité sur le front que dans le cœur du peuplement, a pu rapidement propager la diversité génétique des ancêtres sur le front et contribuer à l'augmentation en fréquence de maladies mendéliennes rares.

L'effet fondateur à l'époque de la Nouvelle-France

Bien que la plupart des prédictions de l'effet fondateur soient limitées à quelques populations régionales ou ethniques et qu'une réduction majeure de la diversité ne soit pas observée à l'échelle de l'ensemble du Québec, nos travaux ont pourtant révélé des signatures génétiques durables du peuplement initial de la Nouvelle-France, partagées par toutes les populations régionales contemporaines. Notre analyse des couches de fondateurs a montré l'importance primordiale et une étonnante persistance de la contribution génétique des premiers fondateurs (Chapitre II). En particulier, pour les fondateurs mariés avant 1660, soit avant l'arrivée importante d'immigrants durant la période des Filles du Roy, on n'observe jusqu'à la période contemporaine aucune perte, et même parfois une augmentation, de leur contribution génétique. Ceci démontre que le succès reproducteur à long terme des premiers pionniers s'est établi dès les toutes premières générations suivant la fondation et a laissé une empreinte durable sur la constitution génétique de l'ensemble de la population. Au chapitre IV, nous avons montré qu'un allèle initialement rare peut augmenter en fréquence de façon drastique. Ceci constitue la première preuve que cette signature de l'effet fondateur est retrouvée dans l'ensemble du Québec et de ses régions, et que son intensité variable culmine dans les régions du Nord-Ouest, du Nord-Est et de l'Est. Ces résultats appuient l'hypothèse d'un effet fondateur initial dans la Vallée du St-Laurent, où la diversité génétique apportée par les premiers pionniers aurait rapidement remanié en fréquence. Cet effet global fût fort mais n'a vraisemblablement agit que sur une très courte durée puisque la quasi-totalité des fondateurs ayant des descendants dans toutes

les régions se sont mariés avant 1700. Ceci pourrait notamment expliquer l'ancienneté des liens d'apparentement au Québec (Chapitre III, Vézina et al. 2004; Tremblay et al., 2008; Tremblay et Vézina 2010) et leur importance cardinale dans la composition génétique de la population, qui est telle que virtuellement tous les Canadiens français sont généalogiquement apparentés entre eux (Tremblay et al. 2008). Identifier les traces génomiques de l'effet fondateur de la Vallée du St-Laurent, notamment celles laissées par les pionniers ayant eu le plus fort succès reproducteur, est une avenue de recherche fascinante.

Vers une résolution de la question de la diversité génétique

Comment réconcilier l'impact de l'effet fondateur avec les niveaux de diversité génétiques observés dans les populations fondatrices? Cette question trouve un écho chez plusieurs populations fondatrices récentes, notamment chez les Islandais et les Costa Ricains de la Vallée centrale, où la question de la diversité génétique a aussi été débattue. Nous offrons dans ce chapitre une explication qui s'appuie sur le peuplement du Québec et permet du même coup de définir l'héritage génétique des Canadiens français.

Après l'effet fondateur initial du peuplement de la Vallée du St-Laurent au 17^{e} siècle, les spécificités génétiques régionales se sont probablement établies rapidement dès le début du 18^{e} siècle. Certaines conséquences génétiques du peuplement initial ont pu être renforcées dans certaines régions par la colonisation successive des territoires plus éloignés et par la dynamique de la reproduction de la population, tel qu'illustré par l'exemple classique de l'expansion démographique de Charlevoix-Saguenay-Lac-St-Jean. En revanche, même si l'effet fondateur initial a pu entrainer une perte de diversité sur le front pionnier de la Vallée du St-Laurent au 17^{e} siècle, cette perte a été compensée à l'échelle du Québec dans les siècles suivants grâce à l'apport génétique des immigrants d'origines de plus en plus variées. Cet apport des immigrants plus tardifs, qui a été plus important dans les régions à l'ouest de la ville de Québec, a même pu y engendrer un excès d'allèles

rares. Ceci pourrait expliquer, du moins en partie, l'excès d'allèles rares observés récemment dans des données de séquençage (Casals et al. 2013). Ces conséquences diversifiées - parfois antagonistes - du peuplement des régions ne peuvent pas être résumées par des modèles simples d'effet fondateur. Nous avons illustré au chapitre V comment l'expansion de Charlevoix-Saguenay-Lac-St-Jean est conforme aux prédictions de modèles d'expansion territoriale. Il reste à évaluer si ces modèles, qui prennent en considération les dynamiques spatiales et temporelles des expansions, peuvent expliquer les patrons de diversité de l'ensemble du Québec. Enfin, avant même de discuter de cet aspect important, on sait d'ores et déjà que pour rendre compte de la diversité génétique des Canadiens français, on doit considérer la population à différentes échelles – régionale, locale, individuelle – qui définissent la structure du paysage génétique québécois. En effet, dans une population structurée, on s'attend à ce que les niveaux de diversité génétique diminuent avec une réduction de l'échelle d'observation.

3. La structure de la population du Québec

Dans cette thèse, j'ai tenté de mieux caractériser la structure de la population du Québec et de mieux comprendre les mécanismes l'ayant façonnée. À l'aide de différentes méthodes statistiques appliquées aux données généalogiques et génomiques, nous avons montré que la population du Québec est significativement structurée à l'échelle régionale et sous-régionale. Nos études ajoutent ainsi de nouvelles preuves de structure en appui à celles avancées précédemment et basées notamment sur la distribution non-uniforme des maladies mendéliennes sur le territoire (Vézina 1996; Scriver 2001; Laberge et al. 2005b; Brais et al. 2007; Laberge 2007), la variation génétique de marqueurs génétiques (Gerbault 2006; Moreau et al. 2007; Moreau et al. 2009) et de protéines sériques (Magnan and Benoist 1969; De Braekeleer 1990) ainsi que les mesures de diversité basées sur les

données généalogiques et patronymiques (Bouchard et al. 1985; Bouchard et al. 1995; Gagnon et al. 1998; Gagnon and Heyer 2001; Vézina et al. 2004). L'idée que les Canadiens français forment une population homogène apparaît aujourd'hui tout à fait erronée. Les travaux présentés dans cette thèse démontrent clairement que la population du Québec représente un assemblage complexe de sous-populations régionales et ethniques qui, à partir d'un tronc ancestral commun, se sont significativement différenciées génétiquement.

Une approche méthodologique originale

Une population est dite structurée, ou stratifiée, lorsqu'elle est composée de sous-groupes d'individus qui se marient préférentiellement entre eux. La structure d'une population dépend donc de la structure des unions qui est directement retracée dans son arbre généalogique. Nous avons exploitée cette idée dans les chapitres II, III et IV, où nous avons développé de nouvelles approches d'analyse de structure à l'aide de données généalogiques, en plus d'utiliser les méthodes existantes. Au chapitre III, nous avons montré que la structure inférée à partir des données généalogiques est extrêmement similaire à celle inférée par les données génomiques. Ce résultat peut être considéré comme une validation empirique des méthodes génomiques d'inférence de structure, réalisée avec des données indépendantes. Colonna et ses collaborateurs (2009) ont montré précédemment que l'apparentement généalogique au sein de deux villages italiens est corrélé avec les clusters d'individus inférés par la méthode STRUCTURE. Au chapitre III, l'analyse en composantes principales de la variation génomique neutre capture effectivement les liens d'apparentement connus dans la généalogie. Ceci est attendu en théorie, puisque les projections des individus sur les composantes principales peuvent être prédites par les temps de coalescence entre les paires de génomes haploïdes (McVean 2009) et ces derniers sont résumés par le coefficient d'apparentement généalogique. Cependant, en pratique, la concordance

observée dépend de la qualité et de la couverture des deux types de données. Nos résultats démontrent donc, d'une part, la grande validité et complétude des généalogies québécoises reconstruites dans le fichier de population BALSAC. D'autre part, il nous amènent à conclure que la structure observée au Québec a nécessairement été générée durant la période couverte par ces généalogies, soit depuis l'arrivée des premiers fondateurs d'origine européenne au 17^{e} siècle.

Comment se dessine la structure du Québec ?

Avec nos trois nouvelles études et la synthèse des études précédentes présentée ici, on peut certes affirmer que la présence de structure au Québec est bien documentée. Pourtant, il reste encore à définir un consensus sur la nature de la structure de la population. En me basant sur nos résultats, j'avance que la variation génétique et la variation généalogique s'articulent autour de deux axes principaux, ou autrement dit, qu'il existe deux dimensions principales de la structure de la population du Québec.

Premièrement, nos avons montré un axe de variation d'ouest en est du Québec sur lequel les populations régionales du Québec s'ordonnent en accord général avec leur position géographique et qui représente un continuum de partage génétique. L'analyse en composante principale présentée au chapitre II a montré que les Canadiens français au milieu du 20^{e} siècle tendent à se regrouper selon leur région de mariage et selon leur position géographique dans l'axe ouest/est de la province. L'adéquation entre l'axe ouest/est de variation observé et l'ordre des individus selon leur région de mariage n'est pas parfaite, mais bien corrélée pour les régions ceinturant le fleuve St-Laurent (Nord de Montréal, région de Montréal, Sud de Montréal, Centre, Québec, Est). Cet axe de variation n'a cependant pas été répliqué aussi clairement dans le chapitre III. Spécifiquement, bien que les sujets provenant de la Côte-Nord et du Saguenay-Lac-St-Jean forment un cluster qui se distingue de celui formé par les sujets de Montréal et Québec, la position des individus dans l'axe ouest/est n'influence pas leur distribution à

l'intérieur de ces deux clusters, sur aucune des trois premières composantes principales de la variation génomique. Cependant, l'axe ouest/est a pu être brouillé dans cette analyse par l'inclusion d'échantillons d'origines ethniques variées (en Gaspésie), par la taille ou même la distribution spatiale des échantillons qui sont des facteurs reconnus pour influencer cette méthode graphique (McVean 2009; Novembre et Stephens 2008; Patterson 2006). Au chapitre IV, nous avons montré des signatures génétiques opposées (surplus/déficit d'allèles rares) dans l'ouest et l'est du Québec. Une preuve encore plus solide de la présence d'un axe de variation ouest/est est l'observation d'un gradient de diversité décroissant d'ouest en est du Québec. Ainsi, au chapitre II, nous avons montré que les indices de diversité génétique basés sur l'effectif efficace de fondateurs diminuent d'ouest en est. Un tel gradient de diversité avait été observé dans la population contemporaine sur la base de la diversité des patronymes (Bouchard et al. 1995) et à l'aide de mesures généalogiques d'apparentement et de consanguinité (Vézina et al. 2004). À la lumière de ces résultats, je propose d'appeler cette dimension de la structure du Québec, « le continuum spatial » de variation génétique des populations régionales dans l'axe ouest/est du Québec.

Deuxièmement, nos études suggèrent que certaines sous-populations régionales et ethniques ressortent, sous certains angles, comme des groupements distincts du reste du Québec. Le Nord-Ouest du Québec, le Sud de Montréal, le Nord-Est et l'Est du Québec, ainsi que les trois sous-populations ethniques de la Gaspésie (les Acadiens, les Loyalistes et les Canadiens français) ont montré une différenciation génétique significative dans une ou plusieurs analyses de structure. Nos résultats appuient ainsi certaines conclusions des premières études réalisées avec l'Échantillon de référence du Québec (Gerbault 2006; Moreau et al. 2007; Moreau et al. 2009) et révèlent pour la première fois la différentiation du Nord-Ouest et du Sud de Montréal. Par ailleurs, les sous-populations différenciées ne sont pas toutes égales en terme de différenciation génétique. Nos résultats laissent supposer

que les Acadiens de la Gaspésie et le Nord-Est (qui inclut Charlevoix, Saguenay Lac-St-Jean et Côte-Nord) occupent respectivement le premier et le deuxième rang de différenciation génétique parmi les populations étudiées. De surcroît, la différenciation de ces deux groupes explique une plus grande proportion de la variation génétique et généalogique entre les individus que tout autre axe de variation, même l'axe ouest/est. Précisément, la variation expliquée par le regroupement des sujets mariés au Nord-Est du Québec sur la première composante principale est plus du double de la variation expliquée par l'axe ouest/est (Chapitre II). De nouvelles analyses génomiques, sur un plus large éventail de populations régionales, seraient nécessaires pour ordonner les autres populations différenciées selon une échelle de différenciation, puisque nos diverses analyses suggèrent plusieurs classements hiérarchiques. Nos résultats indiquent cependant que tous les groupements différenciés se rattachent au tronc ancestral québécois, que je définis comme un partage génétique commun provenant des ancêtres ayant le plus contribué à la population. En effet, au chapitre III, dans les comparaisons avec les populations asiatiques et africaines, les échantillons québécois forment clairement un groupe uni. De plus, tel que révélé au chapitre II, les fondateurs qui ont eu le plus grand poids dans le façonnement de la structure se sont mariés avant la Conquête de 1760, sont communs à toutes les populations régionales et ont eu une forte contribution génétique. En somme, la deuxième dimension de la structure du Québec comprend de multiples segments : des sous-populations différenciées qu'on peut voir comme des ramifications asymétriques divergentes à partir du tronc ancestral québécois.

Une nouvelle métaphore

Par le passé, de nombreuses représentations ont été évoquées pour décrire la structure de la population du Québec. Celles-ci comprennent : deux « stratifications » de l'espace québécois, un axe nord/sud, un axe ouest/est, un clivage ouest/est, deux pôles, une structure bipartite, une structure

tripartite, une courtepointe (« patchwork ») ou mosaïque des pools géniques régionaux (Bouchard 1990; Gagnon and Heyer 2001; Brais et al. 2007). Comment nos résultats se situent-ils vis-à-vis de ces représentations?

L'axe nord/sud n'est pas ressorti dans nos analyses. Ceci peut signifier que cet axe a eu peu d'influence dans la structuration de la population, ou que notre échantillonnage et/ou catégorisation géographique n'ont pas permis de le mettre en valeur. Bouchard (1990) a présenté cet axe géographique comme celui contrastant les premières terres colonisées le long du Fleuve St-Laurent à la périphérie colonisée plus tardivement. Or, au chapitre IV, quelques analyses sur les changements de fréquences alléliques appuient une certaine distinction entre la vallée laurentienne et la périphérie (avec des signatures plus prononcées pour les régions les plus éloignées), mais celles-ci n'ont pas révélé, à strictement parler, un axe nord-sud. L'axe de variation « vielles populations laurentiennes » versus « régions périphériques » resurgira peut-être avec des analyses spatiales plus détaillées considérant directement les coordonnées géographiques des individus échantillonnés, plutôt que des ensembles régionaux comprenant à la fois des vieilles et des nouvelles paroisses.

Nous avons confirmé l'importance de l'axe ouest/est dans la structure de la population canadienne-française du Québec. Cependant, un clivage net entre l'Ouest et l'Est du Québec n'a pas été observé, ni une ligne Beauce-Portneuf séparant le segment ouest du segment est de la population. Cette structure bipartite a notamment été proposée sur la base de la variation dans les parlers québécois, mais plus souvent pour expliquer la distribution des maladies héréditaires (Bouchard 1990; Bouchard and De Braekeleer 1991b; Vézina 1996). Or, en plus d'être remis en doute par l'observation de maladies mendéliennes ayant une fréquence élevée dans l'Ouest du Québec (e.g. Roddier et al. 2005), un clivage artificiel entre l'Est et l'Ouest a peut-être pu être causé par la concentration des études de génétique médicale dans l'Est du Québec. Gagnon et Heyer (2001) ont observé une structure tripartite de la

population au Québec ancien (1608-1800), centrée sur les trois ports d'entrée des immigrants (Montréal, Trois-Rivières, Québec), qui est aussi en contraste avec le continuum spatial de variation observé ici. La différence entre la structure du Québec ancien et du Québec contemporain s'explique probablement par une évolution temporelle de la structure la population. Cette différence est intrigante puisqu'elle pourrait être représentative des signatures attendues des processus d'expansion territoriale, où des clines de fréquences alléliques d'abord très prononcés s'effacent graduellement sous l'effet du flot génétique (Ray and Excoffier 2009).

Les sous-populations plus différenciées du reste du Québec montrent qu'au continuum génétique d'ouest en est du Québec se rattache des groupements d'individus plus différenciés génétiquement. En conséquence, la métaphore qui me semble la plus juste pour représenter la structure de la population est celle d'une grappe de sous-populations qui s'ordonnent le long d'un tronc commun, mais dont certaines s'en détachent à distances variables et dans des directions multiples. Cette nouvelle métaphore représente les facettes de continuité et de segmentation de la population, tout comme la mosaïque ou courtepointe des bassins génétiques régionaux (e.g. Brais et al. 2007), mais elle permet de mieux représenter le chevauchement entre les clusters régionaux et leurs origines ancestrales communes.

4. L'évolution d'une population nouvelle

Conquérir de nouveaux territoires et y établir de nouvelles populations sont certainement des traits caractéristiques de l'espèce humaine. Quels sont les coûts et les bénéfices adaptatifs à fonder une population nouvelle? Quels sont les facteurs qui contribuent au succès d'une telle entreprise? Ces questions importantes sont débattues depuis longtemps par les biologistes évolutifs, mais les données sur les populations naturelles demeurent largement limitées aux organismes ayant un intervalle intergénérationnel

court (Roman and Darling 2007; Waters et al. 2013; Rius and Darling 2014). Dans cette thèse, nous avons montré comment utiliser les généalogies profondes pour étudier la dynamique démographique et spatiale des phénomènes de colonisation chez l'humain. Nos travaux ont révélé certains processus évolutifs en jeu lors de peuplements fondateurs ainsi que leurs conséquences génétiques. À la lumière de nos résultats, un peuplement fondateur représente une période d'évolution accélérée qui peut causer une rapide structuration spatiale du paysage génétique et influencer l'évolution de traits liés à la reproduction de la population.

Une rapide structuration spatiale

La structuration spatiale du paysage génétique québécois détaillée dans la section précédente démontre que la population fondatrice canadienne-française est fractionnée en sous-populations qui, à partir d'une origine commune, se sont rapidement différenciées génétiquement. Cette importante différenciation génétique a été créée et maintenue malgré des apports migratoires internationaux et interrégionaux substantiels (Chapitre II, Chapitre V, Bhérer et al. 2008). Nos résultats démontrent que cet effet diversificateur du peuplement a été entraîné principalement par la réorganisation en fréquence des allèles fondateurs dans les régions par la dérive génétique (le « reshuffling effect ») et que la présence/absence des fondateurs dans les généalogies régionales joue un rôle moins déterminant (Chapitre II, Chapitre IV). Les fondateurs ayant eu le plus grand poids dans le façonnement de la structure se sont mariés avant la Conquête, sont partagés par un maximum de populations régionales et ont eu une forte contribution génétique (Chapitre II, Chapitre IV). Ainsi, l'héritage génétique qui unit et qui différencie les populations régionales du Québec provient essentiellement des mêmes fondateurs.

L'importance de la géographie dans la structure génétique de la population du Québec est en accord avec le principe que « les gènes sont le miroir de la géographie » chez l'Homme (e.g. Cavalli-Sforza et al. 1994; Novembre et al.

2008). Plusieurs facteurs contribuent à cette structuration spatiale, mais elle est d'abord le reflet d'un phénomène d'endogamie géographique où les unions ont lieu préférentiellement entre les individus appartenant à une même unité géographique. Au Québec, ceci est confirmé par l'observation de plus forts indices de consanguinité que d'apparentement qui montre qu'en moyenne les époux sont plus apparentés entre eux que deux individus pris au hasard dans une région donnée (Chapitre III, Vézina et al. 2004; Bhérer et al. 2008). Ensuite, une endogamie ethnique ou religieuse peut jouer un rôle si les groupes sont distribués de façon non uniforme sur le territoire, tel que les lignées maternelles le suggèrent pour les métisses autochtones de la Gaspésie (Moreau et al. 2011b). Enfin, une expansion territoriale peut créer une structuration spatiale complexe de la variation génétique. Une marche régulière du peuplement dans l'espace, qui cause des effets fondateurs successifs au front de la vague d'expansion, peut notamment générer des secteurs de diversité génétique réduits et des clines de fréquences alléliques dans l'axe de l'expansion (Excoffier and Ray 2008). L'orientation et la force de l'axe ouest/est de variation génétique, la drastique augmentation en fréquence de certains allèles fondateurs (Chapitre IV), de même que notre étude de la dynamique de l'expansion dans les régions de Charlevoix et Saguenay-Lac-St-Jean (Chapitre V) appuient l'hypothèse que les processus d'expansions territoriales ont joué un rôle primordial dans le façonnement des patrons de diversité du Québec.

Nos résultats démontrent donc que la dynamique du peuplement des régions du Québec a causé une structure significative au sein même d'une population fondatrice. Une telle structuration génétique régionale et des indices de différenciation du même ordre de grandeur sont documentés en l'Islande (Helgason et al. 2005; Price et al. 2009) et en Finlande (Jakkula et al. 2008). Cette conséquence des processus de colonisation a pu être répandue au cours de la grande expansion de l'Homme moderne sur la planète (Henn et al. 2012) et contribuer substantiellement à la variation génétique observée entre les peuples du monde.

Par ailleurs, nos résultats soutiennent que les migrations plus tardives ont aussi contribué à enrichir la diversité des populations régionales et à accentuer les différences interrégionales (Chapitre II, IV). À ce sujet, une question intéressante qui demeure est d'évaluer dans quelle mesure ce métissage, en créant des nouvelles combinaisons alléliques, a pu être favorable au succès de l'entreprise de peuplement du Québec, tel qu'il a été suggéré chez des populations invasives (Rius and Darling 2014). On peut se demander, par exemple, si certains variants génétiques d'origine autochtone, en conférant une adaptation quelconque à l'environnement, se sont préférentiellement propagé dans la population canadienne-française.

L'avantage reproductif des premiers pionniers

Migrer dans un territoire nouveau n'est pas sans risques ni périls pour les migrants. L'exemple du voyage de la France vers la Nouvelle-France illustre clairement ceci avec 2 000 personnes qui ont péri sur les bateaux parmi les 27 000 embarqués (Charbonneau et al. 2000). D'un point de vue de l'évolution, les coûts adaptatifs associés au peuplement d'un territoire nouveau doivent être balancés par certains avantages adaptatifs pour expliquer que les comportements de dispersion soient aussi répandus chez l'humain. Nos travaux ont révélé un avantage reproductif à fonder une nouvelle population. Nous avons montré que les premiers pionniers ont une plus grande contribution génétique à long terme (Chapitre II, IV, V), vraisemblablement grâce à une plus grande fécondité utile sur le front que dans le cœur du peuplement (Chapitre V). De plus, nous avons montré que cette composante du fitness est héritable seulement sur le front, ce qui implique que ce trait a évolué durant les expansions territoriales (Chapitre V).

Les descriptions généalogiques ont confirmé le plus grand succès reproducteur des fondateurs les plus anciens par rapport aux immigrants plus récents (Chapitre II, IV, V), un phénomène qui a notamment été appelé « premiers arrivés, premiers servis » (Gagnon et al. 2001) et « founder takes all » (Waters et al. 2013). Au Québec, de nombreuses études ont fait état de

cet avantage des premiers fondateurs (e.g. Labuda et al. 1996; Labuda et al. 1997; Charbonneau et al. 2000; Vézina et al. 2005b). Dans notre étude au Chapitre V, tout en fournissant la première description de la dynamique démographique d'une expansion humaine, nous avons montré que le succès reproducteur est associé à la marche du peuplement sur le territoire. Les individus provenant des paroisses nouvellement ouvertes ont contribué plus de gènes à la génération présente que les individus dans les plus vieilles paroisses (Chapitre V), en accord avec la théorie qui prédit que les gènes qui surfent sur l'expansion devraient remonter aux ancêtres sur (ou près de) la vague d'expansion (Hallatschek and Nelson 2008). Néanmoins, nous avons montré que ce succès reproducteur accru n'est pas seulement causé par le phénomène de surfing génétique, mais est aussi vraisemblablement engendré par une plus forte fécondité utile sur le front que dans le cœur du peuplement (Chapitre V). Par ailleurs, nos résultats suggèrent aussi que le spectaculaire succès reproducteur des top-fondateurs du Québec est associé à des comportements de coopération (Chapitre IV), en accord avec les récentes preuves que ces comportements sont favorisés sur le front des expansions (e.g. Datta et al. 2013).

Nos résultats suggèrent que la fécondité utile accrue sur le front du peuplement est associée, du moins en partie, à la relaxation des contraintes liées à l'accès au territoire dans les territoires nouveaux et donc au mariage et à la reproduction. Alors que le nombre total d'enfants par femme et son héritabilité sont similaires sur le front et dans le cœur du peuplement, nous avons montré que le nombre d'enfants mariés (la fécondité efficace) est plus grand sur le front que dans le cœur du peuplement et héritable seulement sur le front (Chapitre V). L'augmentation de la fécondité utile dans les paroisses nouvellement fondées est en effet associée à une réduction de l'âge au mariage (Chapitre V), qui est le trait qui augmente le plus efficacement le taux de reproduction (Lewontin 1965). Celui-ci a été montré pour évoluer chez les populations invasives (Phillips et al. 2010) et aussi chez des populations humaines, notamment dans la population québécoise de l'Île aux Coudres,

dans ce cas potentiellement dû à la sélection positive (Milot et al. 2011). L'évolution des comportements de reproduction que nous avons montrés au Saguenay-Lac-Saint-Jean suit donc la prédiction que des stratégies de fortes croissances devraient être sélectionnées dans les territoires nouveaux parce qu'elles ne sont pas soumis à la compétition pour les ressources liées à la densité de la population au cœur du peuplement (Phillips et al. 2010; Waters et al. 2013).

Néanmoins, cet avantage reproductif des populations frontières ne vient pas sans coût : le surfing génétique de certains allèles délétères. Au chapitre IV, nous avons montré comment les gènes de certains fondateurs peuvent augmenter drastiquement en fréquence, jusqu'à atteindre les fréquences de porteur connues pour certaines maladies mendéliennes rares. Une récente étude génomique par Casals et al. (2013) a montré que la population canadienne-française présente un excès d'allèles potentiellement délétères par rapport à la France. Cet accroissement du fardeau de mutations délétères peut s'expliquer par les faibles densités de population au front de l'expansion et les forts taux d'accroissement qui augmentent la dérive génétique et diminuent l'efficacité de la sélection (Alves et al. 2012; Casals et al. 2013). Cet effet de balance évolutive entre l'avantage reproductif des premiers pionniers et l'accumulation des mutations délétères pourrait avoir été important au cours de l'évolution humaine, notamment au cours de la colonisation de l'Europe (Lohmueller et al. 2008). Plus d'études seront nécessaires pour relier ces prédictions à des phénotypes concrets (autres que les maladies mendéliennes) qui auraient pu conférer un avantage ou un désavantage adaptatif au cours de l'adaptation d'une population fondatrice à son nouvel environnement. Cependant, ces études suggèrent que le fardeau de mutation délétères varient d'une sous-population à une autre, notamment à cause de l'effet fondateur et donc que la démographie a bel et bien un impact sur le fardeau mutationnel, du moins lorsqu'on définit celui-ci non seulement par le nombre, mais aussi par la variance de ce nombre entre les individus, contrairement à une idée récemment avancée (Simons et al. 2014).

5. Retombées pour les études en épidémiologie génétique

Durant plusieurs décennies, les populations fondatrices ont été au cœur des efforts en génétique médicale pour identifier les mutations responsables des maladies mendéliennes (Arcos-Burgos and Muenke 2002; McKusick 2007). Au cours des années 2000, la recherche des déterminants génétiques des maladies complexes a détourné l'attention des variants délétères rares aux variants de susceptibilité communs (Reich & Lander, 2001) et a orienté le design vers des études d'association dans des échantillons populationnels de plus en plus grands et diversifiés. Néanmoins, au cours des dernières années, l'importance des variants rares dans la susceptibilité aux maladies complexes a été mise en évidence et a ravivé l'intérêt pour les populations fondatrices.

Dans ce contexte, la présente étude a mis en lumière des avantages potentiels à étudier la population fondatrice canadienne-française. D'abord, l'accessibilité aux données généalogiques est certainement une des ressources les plus exceptionnelles pour les recherches au Québec. Nous avons fait la preuve de l'excellente qualité de ces données par leur concordance avec les analyses génomiques au Chapitre III. Cependant, il ne faut pas sous-estimer la difficulté d'exploiter cette information complexe en combinaison avec les données génomiques ni le besoin énorme de développements méthodologiques à ce sujet. Par ailleurs, en caractérisant les patrons de variation génétique des régions du Québec, nos résultats ont révélé les sous-populations plus diversifiées et celles plus marquées par l'effet fondateur. Cette variabilité des populations régionales devrait permettre un éventail de design d'études à l'intérieur même du Québec. Les études cherchant à maximiser la diversité génétique pourront se tourner vers la région métropolitaine de Montréal dont les patrons de diversité sont similaires à ceux de la France. Les populations du Saguenay-Lac-Saint-Jean et d'origine acadienne ressortent quant à elles comme idéales pour les études qui chercheraient à exploiter les avantages potentiels des populations

fondatrices (Jorde et al. 2000; Peltonen et al. 2000; Shifman and Darvasi 2001; Heutink and Oostra 2002; Newman et al. 2004; Bourgain and Genin 2005; Kristiansson et al. 2008).

À la lumière de nos résultats, nous pouvons conclure qu'au Québec la structure de la population et l'apparentement cryptique (ou éloigné) sont des facteurs importants à considérer pour optimiser les études en épidémiologie génétique. En plus des origines ancestrales, la géographie explique une large proportion de la variation les fréquences alléliques (Chapitre II, III), de sorte que deux voisins ont plus de chance de partager des allèles identiques par ascendance (« IBD ») que des individus séparés par une plus grande distance géographique. Également, les fréquences alléliques sont plus similaires entre des régions voisines qu'entre des régions plus éloignées l'une de l'autre (Chapitre II, III, IV). De nombreuses méthodes statistiques permettent de contrôler ces facteurs dans les études d'association (Chapitre III, Price et al. 2010). Or, lorsque possible, il demeure bénéfique de collecter l'information sur l'origine régionale et ethnique des participants à une étude. Ceci pourrait par exemple permettre de retracer les débalancements régionaux entre les cas et les contrôles qui peuvent causer des résultats faux-positifs dans les études d'association (Cardon and Palmer 2003; Marchini et al. 2004). La stratification régionale de la population canadienne-française entraine aussi la possibilité que l'incidence de certains traits puisse varier significativement d'une région à l'autre. Ceci devrait surtout affecter la distribution des traits dont l'étiologie est basée sur les variants rares, tel qu'il est connu pour les maladies monogéniques rares au Québec.

Concernant les maladies rares, cette thèse confirme l'importance de l'histoire du peuplement pour expliquer les concentrations régionales et locales de certains phénotypes mendéliens, décrits comme des « effets fondateurs cliniques ». Nos simulations ont même permis d'identifier les populations régionales les plus « à risque » de présenter des maladies mendéliennes ayant une incidence accrue (Chapitre IV). La recension des écrits présentée

en introduction a fait ressortir certaines lacunes dans nos connaissances du bagage génétique des Canadiens français. En particulier, notre connaissance de la fréquence de porteurs, de la prévalence et de la distribution géographique des maladies caractéristiques des Canadiens français est essentiellement limitée aux individus atteints. La récente disponibilité des techniques de séquençage à haut débit devrait faciliter des sondages des mutations causales des maladies mendéliennes et offrir des estimations non biaisées des facteurs de risque génétiques ségrégrants au sein de cette population fondatrice. Ceci devrait pouvoir guider les politiques de santé publique au Québec, notamment au sujet des tests de dépistage néonataux qui sont présentement déficients (Côté and Gosselin 2013).

CONCLUSION

Dans cette thèse, j'ai contribué à échantillonner la variation génétique et à mieux caractériser les patrons de diversité génétique de la population du Québec et de ses régions. J'ai développé et utilisé des approches méthodologiques originales pour extraire l'information contenue dans l'arbre généalogique de la population. Nos travaux pionniers dans ce domaine pourront servir de modèle aux études de généalogie génétique qui gagneront sans doute en importance avec la croissance des bases de données généalogiques disponibles sur le web (Larmuseau et al. 2013).

Vingt-trois ans après la publication du livre « Histoire d'un génome » (Bouchard and De Braekeleer 1991a), cette thèse offre une nouvelle synthèse des travaux de génétique des populations au Québec. Alors que le livre traitait essentiellement des régions de Charlevoix et du Saguenay-Lac-St-Jean, cette thèse porte sur l'ensemble des populations régionales du Québec. J'ai pu y dresser un portrait détaillé de la majorité canadienne-française et de quelques sous-populations ethnoculturelles recensées dans l'Échantillon de référence du Québec. J'ai omis de discuter d'un grand nombre de sous-populations, tels que les Irlandais, les Juifs, les Italiens, qui forment pourtant une fraction importante du Québec d'hier et d'aujourd'hui. Ce tissu génétique québécois, vraisemblablement métissé de façon complexe, mérite d'être caractérisé dans des recherches futures, notamment pour mieux connaître la distribution des facteurs de risque génétique dans la population et d'ajuster en conséquence le panel de tests de dépistage.

Le portrait nouveau du patrimoine génétique du Québec qui émerge de cette thèse permet de trancher la question longtemps controversée et difficile à cerner de la diversité génétique. J'ai soutenu que la population du Québec forme un assemblage complexe de sous-populations régionales et ethniques, réunies par des liens d'apparentement ancien et de métissage continu. À la lumière de ces résultats, on peut supposer que la population du Québec

forme un peuple « métissé serré »[20], depuis sa fondation il y a 400 ans. La structure régionale de la population d'origine canadienne-française, qui compose aujourd'hui environ 80% de la population du Québec, témoigne d'une histoire de peuplement complexe et intrigante puisqu'elle ne se conforme pas aux modèles simples associés à l'effet fondateur.

Les conséquences des processus de colonisation de nouveaux territoires ont sans aucun doute laissé leurs traces dans les lignées ancestrales de tous les peuples humains. Les travaux présentés dans cette thèse ont mis en lumière différents mécanismes d'évolution jouant un rôle lors de la formation des populations nouvelles. Aujourd'hui, le Québec peut être considéré comme un laboratoire de génétique humaine exceptionnel grâce aux larges effectifs et à la grande diversité de ses sous-populations, au bassin grandissant de chercheurs en génétique ainsi qu'à l'accessibilité à des données généalogiques d'une profondeur et d'une qualité exceptionnelles.

[20] J'emprunte ici l'expression de Boucar Diouf qui fait référence à l'expression des Québécois « tricotés serrés » (Diouf B. 2013. Vive le Québec métissé serré! La Presse, Section Débats, 19 octobre 2013.).

Bibliographie

Abecasis GR, Burt RA, Hall D, Bochum S, Doheny KF, Lundy SL, Torrington M, Roos JL, Gogos JA, Karayiorgou M. 2004. Genomewide scan in families with schizophrenia from the founder population of Afrikaners reveals evidence for linkage and uniparental disomy on chromosome 1. *Am J Hum Genet* **74**(3): 403-417.

Agarwala R, Biesecker LG, Schaffer AA. 2003. Anabaptist genealogy database. *American journal of medical genetics Part C, Seminars in medical genetics* **121C**(1): 32-37.

Alcala N, Streit D, Goudet J, Vuilleumier S. 2013. Peak and persistent excess of genetic diversity following an abrupt migration increase. *Genetics* **193**(3): 953-971.

Allendorf FW. 1986. Genetic drift and the loss of alleles versus heterozygosity. *Zoo Biology* **5**(2): 181-190.

Alves I, Sramkova Hanulova A, Foll M, Excoffier L. 2012. Genomic data reveal a complex making of humans. *PLoS Genet* **8**(7): e1002837.

Anderson EC, Slatkin M. 2007. Estimation of the number of individuals founding colonized populations. *Evolution* **61**(4): 972-983.

Anderson KG. 2006. How well does paternity confidence match actual paternity? Evidence from worldwide nonpaternity rates. *Curr Anthropology* **47**: 513-520.

Arcos-Burgos M, Muenke M. 2002. Genetics of population isolates. *Clin Genet* **61**(4): 233-247.

Ardlie KG, Kruglyak L, Seielstad M. 2002. Patterns of linkage disequilibrium in the human genome. *Nature reviews Genetics* **3**(4): 299-309.

Ashrafian-Bonab M, Lawson Handley LJ, Balloux F. 2007. Is urbanization scrambling the genetic structure of human populations? A case study. *Heredity* **98**(3): 151-156.

Austerlitz F, Heyer E. 1998. Social transmission of reproductive behavior increases frequency of inherited disorders in a young-expanding population. *Proc Natl Acad Sci U S A* **95**(25): 15140-15144.

-. 1999. Impact of demographic distribution and population growth rate on haplotypic diversity linked to a disease gene and their consequences for the estimation of recombination rate: example of a French Canadian population. *Genet Epidemiol* **16**(1): 2-14.

-. 2000. Allelic association is increased by correlation of effective family size. *Eur J Hum Genet* **8**(12): 980-985.

Austerlitz F, Jung-Muller B, Godelle B, Gouyon P. 1997. Evolution of coalescence times, genetic diversity and structure during colonization. *Theoretical population biology* **51**: 148-164.

Awadalla P, Boileau C, Payette Y, Idaghdour Y, Goulet JP, Knoppers B, Hamet P, Laberge C, Project CA. 2013. Cohort profile of the CARTaGENE study: Quebec's population-based biobank for public health and personalized genomics. *International journal of epidemiology* **42**(5): 1285-1299.

Barrett JC, Fry B, Maller J, Daly MJ. 2005. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**(2): 263-265.

Barton N, Charlesworth B. 1984. Genetic revolutions, founder effects and speciation. *Annual Review of Ecology and Systematics* **15**: 133-164.

Barton NH. 2008. The effect of a barrier to gene flow on patterns of geographic variation. *Genetics research* **90**(1): 139-149.

Barton NH, Etheridge AM. 2011. The relation between reproductive value and genetic contribution. *Genetics* **188**(4): 953-973.

Bergeron J, Vézina H, Houde L, Tremblay M. 2008. La contribution des Acadiens au peuplement des régions du Québec. *Cahiers québécois de démographie* **37**: 181-204.

Bernard G, Thiffault I, Tetreault M, Putorti ML, Bouchard I, Sylvain M, Melancon S, Laframboise R, Langevin P, Bouchard JP et al. 2010. Tremor-ataxia with central hypomyelination (TACH) leukodystrophy maps to chromosome 10q22.3-10q23.31. *Neurogenetics* **11**(4): 457-464.

Bhérer C. 2006. Caractérisation du pool génique de Lanaudière : Analyse démogénétique et étude épidémiogénétique de la névrite héréditaire NHSA2. Mémoire de maîtrise présenté à la Faculté des études supérieure de l'Université Laval. Université du Québec à Chicoutimi. Octobre 2006. 131 pages.

Bhérer C, Brais B, Vézina H. 2008. Impact des récentes transformations démographiques sur le bassin génétique de la région de Lanaudière. *Cahiers québécois de démographie* **37**(2): 211-236.

Bhérer C, Labuda D, Roy-Gagnon MH, Houde L, Tremblay M, Vezina H. 2011. Admixed ancestry and stratification of Quebec regional populations. *Am J Phys Anthropol* **144**(3): 432-441.

Biek R, Henderson JC, Waller LA, Rupprecht CE, Real LA. 2007. A high-resolution genetic signature of demographic and spatial expansion in epizootic rabies virus. *Proc Natl Acad Sci U S A* **104**(19): 7993-7998.

Boerwinkle E, Chakraborty R, Sing CF. 1986. The use of measured genotype information in the analysis of quantitative phenotypes in man. I. Models and analytical methods. *Annals of human genetics* **50**(Pt 2): 181-194.

Boleda M. 1990. Trente mille Français à la conquête du Saint-Laurent. *Histoire Sociale - Social History* **23**(45): 153-177.

Bonnen PE, Pe'er I, Plenge RM, Salit J, Lowe JK, Shapero MH, Lifton RP, Breslow JL, Daly MJ, Reich DE et al. 2006. Evaluating potential for whole-genome studies in Kosrae, an isolated population in Micronesia. *Nat Genet* **38**(2): 214-217.

Bosch E, Calafell F, Rosser ZH, Norby S, Lynnerup N, Hurles ME, Jobling MA. 2003. High level of male-biased Scandinavian admixture in Greenlandic Inuit shown by Y-chromosomal analysis. *Human genetics* **112**(4): 353-363.

Bouchard G. 1983. Le peuplement blanc. In *Les Saguenayens Introduction à l'histoire du Saguenay XVIe-XXe siècles*, (ed. C Pouyez, Y Lavoie), pp. 125-180. Presses de l'Université du Québec, Sillery, Québec, Canada.

-. 1990. Représentations de la population et de la société québécoises : l'apprentissage de la diversité. *Cahiers québécois de démographie* **19**(1): 7-28.

-. 2004. Information génétique et risque de stigmatisation collective. L'exemple du Saguenay-Lac-St-Jean. *Med Sci (Paris)* **20**: 933-934.

Bouchard G, Charbonneau H, Desjardins B, Heyer E, Tremblay M. 1995. Mobilité géographique et stratification du pool génique canadien-français sous le Régime français. In *Landry Y, Dickinson J, Pasleau S, Desama C, editeurs Les chemins de la migration en Belgique et au Québec XVIIe - XXe siècles*, pp. 51-59. Publications MNH, Beauport.

Bouchard G, De Braekeleer M. 1990. Homogénéité ou diversité? L'histoire de la population du Québec revue à travers ses gènes. *Histoire Sociale - Social History* **23**(46): 325-361.

-. 1991a. *Histoire d'un génome. Population et génétique dans l'est du Québec*. Presses de l'Université du Québec, Sillery, Québec.

-. 1991b. Mouvements migratoires, effets fondateurs et homogénéisation génétique. In *Histoire d'un génome Population et génétique dans l'est du Québec*, (ed. G Bouchard, M De Braekeleer), pp. 282-321. Presses de l'Université du Québec, Sillery, Québec.

Bouchard G, Desjardins-Ouellette M-A, Markowski F, Kouladjian K. 1985. La distribution des patronymes au Québec : témoins des dynamiques de population. *Anthropologie et Sociétés* **9**(3): 197-218.

Bouchard G, Morissette J, Kouladjian K. 1987. La statistique agrégée des patronymes du Saguenay et de Charlevoix comme indicateurs de la structure de la population aux XIXe et XXe siècles. *Cahiers québécois de démographie* **16**(1): 67-98.

Bouchard G, Vézina H. 2009. Projet BALSAC - Rapport annuel 2008-2009. Chicoutimi (Québec), Canada.

Bourgain C, Genin E. 2005. Complex trait mapping in isolated populations: Are specific statistical methods required? *Eur J Hum Genet* **13**(6): 698-706.

Boycott KM, Parboosingh JS, Chodirker BN, Lowry RB, McLeod DR, Morris J, Greenberg CR, Chudley AE, Bernier FP, Midgley J et al. 2008. Clinical genetics and the Hutterite population: a review of Mendelian disorders. *American journal of medical genetics Part A* **146A**(8): 1088-1098.

Brais B, Desjardins B, Labuda D, St-Hilaire M, Tremblay M, Vézina H. 2007. The genetics of French Canadians. In *Human population genetics : evolution and variation*, The Biomedical & Life Science Collection (ed. L Cavalli-Sforza, M Feldman). Henry Stewart Talks Ltd, London.

Brandenburg JT, Austerlitz F, Toupance B. 2012. Impact of fertility transmission and other sociodemographic factors on reproductive success and coalescent trees. *Genetics research* **94**(3): 121-131.

Buri P. 1956. Gene frequency in small populations of mutant drosophila. *Evolution*: 367-402.

Calboli FC, Sampson J, Fretwell N, Balding DJ. 2008. Population structure and inbreeding from pedigree analysis of purebred dogs. *Genetics* **179**(1): 593-601.

Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, Bodmer WF, Bonne-Tamir B, Cambon-Thomsen A et al. 2002. A human genome diversity cell line panel. *Science* **296**(5566): 261-262.

Cannon Albright LA. 2008. Utah family-based analysis: past, present and future. *Human heredity* **65**(4): 209-220.

Cannon-Albright LA, Farnham JM, Thomas A, Camp NJ. 2005. Identification and study of Utah pseudo-isolate populations-prospects for gene identification. *American journal of medical genetics Part A* **137A**(3): 269-275.

Cardon LR, Palmer LJ. 2003. Population stratification and spurious allelic association. *Lancet* **361**(9357): 598-604.

Carothers AD, Rudan I, Kolcic I, Polasek O, Hayward C, Wright AF, Campbell H, Teague P, Hastie ND, Weber JL. 2006. Estimating human inbreeding coefficients: comparison of genealogical and marker heterozygosity approaches. *Annals of human genetics* **70**(Pt 5): 666-676.

Carvajal-Carmona LG, Ophoff R, Service S, Hartiala J, Molina J, Leon P, Ospina J, Bedoya G, Freimer N, Ruiz-Linares A. 2003. Genetic demography of Antioquia (Colombia) and the Central Valley of Costa Rica. *Human genetics* **112**(5-6): 534-541.

Casals F, Hodgkinson A, Hussin J, Idaghdour Y, Bruat V, de Maillard T, Grenier JC, Gbeha E, Hamdan FF, Girard S et al. 2013. Whole-exome sequencing reveals a rapid change in the frequency of rare functional variants in a founding population of humans. *PLoS Genet* **9**(9): e1003815.

Cavalli-Sforza L, Menozzi P, Piazza A. 1994. *The History and Geography of Human Genes*. Princeton University Press, Princeton.

Chakravarti A. 1999. Population genetics--making sense out of sequence. *Nat Genet* **21**(1 Suppl): 56-60.

Chapdelaine C. 2007. Cliche-Rancourt, un site du Paléoindien ancien. In *Entre lacs et montagnes au Méganticois: 12,000 ans d'histoire amérindienne*, (ed. C Chapdelaine), pp. 47-120. Paléo-Québec 32, Recherches amérindiennes au Québec, Montréal.

Chappuis PO, Hamel N, Paradis AJ, Deschenes J, Robidoux A, Potvin C, Cantin J, Tonin P, Ghadirian P, Foulkes WD. 2001. Prevalence of founder BRCA1 and BRCA2 mutations in unselected French Canadian women with breast cancer. *Clin Genet* **59**(6): 418-423.

Charbonneau H, Desjardins B, Guillemette A, Landry Y, Légaré J, Nault F. 1987. *Naissance d'une population. Les Français établis au Canada au XVIIe siècle*. Presses universitaires de France and Presses de l'Université de Montréal, Paris/Montréal.

-. 1993. *The first French Canadians. Pionners in the St. Lawrence Valley*. University of Delaware Press and Associated University Press, Newark, London and Toronto.

Charbonneau H, Desjardins B, Légaré J, Denis H. 2000. The population of the St-Lawrence Valley, 1608-1760. In *A population history of North America*, (ed. MR Haines, RH Steckel), pp. 99-142. Cambridge University Press.

Charlesworth B. 1995. Speciation. Down the bottleneck? *Current biology : CB* **5**(9): 995-996.

-. 2009. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nature reviews Genetics* **10**(3): 195-205.

Chong JX, Ouwenga R, Anderson RL, Waggoner DJ, Ober C. 2012. A population-based study of autosomal-recessive disease-causing mutations in a founder population. *Am J Hum Genet* **91**(4): 608-620.

Clark AG. 1999. The size distribution of homozygous segments in the human genome. *Am J Hum Genet* **65**(6): 1489-1492.

Clegg SM, Degnan SM, Kikkawa J, Moritz C, Estoup A, Owens IP. 2002. Genetic consequences of sequential founder events by an island-colonizing bird. *Proc Natl Acad Sci U S A* **99**(12): 8127-8132.

Colonna V, Nutile T, Astore M, Guardiola O, Antoniol G, Ciullo M, Persico MG. 2007. Campora: a young genetic isolate in South Italy. *Human heredity* **64**(2): 123-135.

Colonna V, Nutile T, Ferrucci RR, Fardella G, Aversano M, Barbujani G, Ciullo M. 2009. Comparing population structure as inferred from genealogical versus genetic information. *Eur J Hum Genet* **17**(12): 1635-1641.

Comeron JM, Williford A, Kliman RM. 2008. The Hill-Robertson effect: evolutionary consequences of weak selection and linkage in finite populations. *Heredity* **100**(1): 19-31.

Cornuet JM, Luikart G. 1996. Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics* **144**(4): 2001-2014.

Côté B, Gosselin C. 2013. Pertinence d'élargir le programme de dépistage néonatal sanguin au Québec (INESS). http://www.inesss.qc.ca/fileadmin/doc/INESSS/Rapports/Genetique/INESSS_Dépistage_néonatal_sanguin.pdf.

Courville S. 1996. *Population et territoire, Atlas historique du Québec*. Presses de l'Université Laval, Sainte-Foy.

-. 2000. *Le Québec. Genèses et mutations du territoire : synthèse de géographie historique, Géographie historique.* Presses de l'Université Laval, Sainte-Foy.

Datta MS, Korolev KS, Cvijovic I, Dudley C, Gore J. 2013. Range expansion promotes cooperation in an experimental microbial metapopulation. *Proc Natl Acad Sci U S A* **110**(18): 7354-7359.

De Braekeleer M. 1990. Homogénéité génétique des Canadiens Français du Québec : mythe ou réalité? *Cahiers québécois de démographie* **19**(1): 29-48.

-. 1991a. Hereditary disorders in Saguenay-Lac-St-Jean (Quebec, Canada). *Human heredity* **41**(3): 141-146.

-. 1991b. Les gènes délétères. In *Histoire d'un génome Population et génétique dans l'est du Québec*, (ed. G Bouchard, M De Braekeleer), pp. 344-364. Presses de l'Université du Québec, Sillery, Québec.

De Braekeleer M, Dao TN. 1994. Hereditary disorders in the French Canadian population of Quebec. II. Contribution of Perche. *Hum Biol* **66**(2): 225-249.

De Braekeleer M, Hechtman P, Andermann E, Kaplan F. 1992. The French Canadian Tay-Sachs disease deletion mutation: identification of probable founders. *Human genetics* **89**(1): 83-87.

DeGiorgio M, Degnan JH, Rosenberg NA. 2011. Coalescence-time distributions in a serial founder model of human evolutionary history. *Genetics* **189**(2): 579-593.

DeGiorgio M, Jakobsson M, Rosenberg NA. 2009. Out of Africa: modern human origins special feature: explaining worldwide patterns of human genetic

variation using a coalescent-based serial founder model of migration outward from Africa. *Proc Natl Acad Sci U S A* **106**(38): 16057-16062.

Depaulis F, Mousset S, Veuille M. 2003. Power of neutrality tests to detect bottlenecks and hitchhiking. *Journal of molecular evolution* **57 Suppl 1**: S190-200.

Deshpande O, Batzoglou S, Feldman MW, Cavalli-Sforza LL. 2009. A serial founder effect model for human settlement out of Africa. *Proceedings Biological sciences / The Royal Society* **276**(1655): 291-300.

Desjardins B. 1998. Le Registre de la population du Québec Ancien. *Ann Demogr Hist* **2**: 215-226.

-. 2008. La contribution différentielle des immigrants français à la souche canadienne-française. *Annales de Normandie* **58**(3-4): 69-79.

Desjardins M, Frenette Y, Bélanger J, Hétu B. 1999. *Histoire de la Gaspésie*. Les Presses de l'Université Laval, Sainte-Foy.

Desy M, Filiatrault F, Henderson R. 2012. Comité d'éthique de santé publique de l'Institut national de santé publique du Québec. Avis sur le Programme québécois de dépistage néonatal sanguin - Volets information, consentement et dévoilement du statut de porteur. http://www.inspq.qc.ca/pdf/publications/1551_AvisPQDNS_VoletsInfoConsentDevoilStatutPorteur.pdf

Devlin B, Roeder K. 1999. Genomic control for association studies. *Biometrics* **55**(4): 997-1004.

Dickinson J. 1994. Les réfugiés acadiens au Canada, 1755-1775. *Études canadiennes/Canadian Studies* **37**: 51-61.

Diouf B. 2013. Vive le Québec métissé sérré! La Presse, Section Débats, 19 octobre 2013.

Dobzhansky T, Pavlovsky O. 1957. An experimental study of interaction between genetic drift and natural selection. *Evolution* **11**: 311-319.

Donnelly KP. 1983. The probability that related individuals share some section of genome identical by descent. *Theoretical population biology* **23**(1): 34-63.

Dubut V, Chollet L, Murail P, Cartault F, Beraud-Colomb E, Serre M, Mogentale-Profizi N. 2004. mtDNA polymorphisms in five French groups: importance of regional sampling. *Eur J Hum Genet* **12**(4): 293-300.

Dupré N, Bouchard J-P, Brais B, Rouleau G. 2006. Hereditary ataxia, spastic paraparesis and neuropathy in the French-Canadian population. *Can J Neurol Sci* **33**: 149-157.

Dupre N, Chrestian N, Thiffault I, Brais B, Rouleau GA, Bouchard JP. 2008. [Hereditary ataxias, spastic parapareses and neuropathies in Eastern Canada]. *Revue neurologique* **164**(1): 12-21.

Edmonds CA, Lillie AS, Cavalli-Sforza LL. 2004. Mutations arising in the wave front of an expanding population. *Proc Natl Acad Sci U S A* **101**(4): 975-979.

Edwards AW. 1992. The structure of the Polar Eskimo genealogy. *Human heredity* **42**(4): 242-252.

Edwards AWF. 1968. Simulation studies of genealogies. *Heredity* **23**: 628.

Estoup A, Baird SJ, Ray N, Currat M, Cornuet JM, Santos F, Beaumont MA, Excoffier L. 2010. Combining genetic, historical and geographical data to reconstruct the dynamics of bioinvasions: application to the cane toad Bufo marinus. *Molecular ecology resources* **10**(5): 886-901.

Excoffier L, Foll M, Petit RJ. 2009. Genetic Consequences of Range Expansions. *Annual Review in Ecology, Evolution, and Systematics* **40**(481-501).

Excoffier L, Laval G, Schneider S. 2005. Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evolutionary bioinformatics online* **1**: 47-50.

Excoffier L, Ray N. 2008. Surfing during population expansions promotes genetic revolutions and structuration. *Trends in ecology & evolution* **23**(7): 347-351.

Falconer DS, McKay TFC. 1996. *Introduction to Quantitatve Genetics*. Addison Wesley Longman, Harlow, Essex, UK.

Fay JC, Wu CI. 1999. A human population bottleneck can account for the discordance between patterns of mitochondrial versus nuclear DNA variation. *Molecular biology and evolution* **16**(7): 1003-1005.

Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N, Gabriel SB, Topol EJ, Smoller JW, Pato CN et al. 2004. Assessing the impact of population stratification on genetic association studies. *Nat Genet* **36**(4): 388-393.

Freire-Maia N. 1968. Inbreeding levels in American and Canadian populations : A comparison with Latin America. *Biodemography and Social Biology* **15**(1): 22-33.

Frenette P. 1996. *Histoire de la Côte-Nord.* Institut québécois de recherche sur la culture et Les Presses de l'Université Laval.

Gagnon A. 2001. Patronymes, numéros de lignée et généalogies : trois méthodes d'estimation de la parenté au Québec ancien (1608-1800). In *Le patronyme: histoire, anthropologie et société*, (ed. G Brunet, P Darlu, G Zei), pp. 333-349. CNRS.

Gagnon A, Charbonneau H, Desjardins B. 1998. Apparentement des conjoints par région au Québec ancien. *Le choix du conjoint : premiers entretiens de la Société de démographie historique* **43**: 133-150.

Gagnon A, Heyer E. 2001. Fragmentation of the Quebec population genetic pool (Canada): evidence from the genetic contribution of founders per region in the 17th and 18th centuries. *Am J Phys Anthropol* **114**(1): 30-41.

Gagnon A, Vézina H, Brais B. 2001. Histoire démographique et génétique du Québec. *Pour la science*(No 287 septembre): 62-69.

Gattepaille LM, Jakobsson M, Blum MG. 2013. Inferring population size changes with sequence and SNP data: lessons from human bottlenecks. *Heredity* **110**(5): 409-419.

Gauvreau D, Guérin M, Hamel M. 1991. De Charlevoix au Saguenay: mesure et caractéristiques du mouvement migratoire avant 1911. In *Histoire d'un génome Population et génétique dans l'Est du Québec*, (ed. G Bouchard, M De Braekeleer), pp. 145-162. Presses de l'Université du Québec, Sillery, Québec.

Gemery H. 2000. The white population of the colonial United States, 1607-1790. In *A population history of North America*, (ed. MR Haines, RH Steckel), pp. 143-190. Cambridge University Press.

Gerbault P. 2006. Régionalisation de l'effet fondateur au Québec. Mémoire de maîtrise présenté à la Faculté des études supérieures de l'Université de Montréal. Août 2006. 107 pages.

Gibson MA, Gurmu E. 2011. Land inheritance establishes sibling competition for marriage and reproduction in rural Ethiopia. *Proceedings of the National Academy of Sciences of the United States of America* **108**(6): 2200-2204.

Glasson EJ, de Klerk NH, Bass AJ, Rosman DL, Palmer LJ, Holman CD. 2008. Cohort profile: The Western Australian Family Connections Genealogical Project. *International journal of epidemiology* **37**(1): 30-35.

Gosselin I, Thiffault I, Tetreault M, Chau V, Dicaire MJ, Loisel L, Emond M, Senderek J, Mathieu J, Dupre N et al. 2008. Founder SH3TC2 mutations are responsible for a CMT4C French-Canadians cluster. *Neuromuscular disorders : NMD* **18**(6): 483-492.

Gradie M, Jorde LB, Bouchard G. 1991. La structure génétique de la population du Saguenay. In *Histoire d'un génome Population et génétique dans l'est du Québec*, (ed. G Bouchard, M De Braekeleer), pp. 254-277. Presses de l'Université du Québec, Sillery, Québec.

Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Yu F, Gibbs RA, Genomes P, Bustamante CD. 2011. Demographic history and rare allele sharing among human populations. *Proc Natl Acad Sci U S A* **108**(29): 11983-11988.

Greeff JM. 2007. Deconstructing Jaco: genetic heritage of an Afrikaner. *Annals of human genetics* **71**(Pt 5): 674-688.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH et al. 2010. A draft sequence of the Neandertal genome. *Science* **328**(5979): 710-722.

Gros-Louis F, Dupre N, Dion P, Fox MA, Laurent S, Verreault S, Sanes JR, Bouchard JP, Rouleau GA. 2007. Mutations in SYNE1 lead to a newly discovered form of autosomal recessive cerebellar ataxia. *Nat Genet* **39**(1): 80-85.

Guillemette A, Légaré J. 1989. The influence of kinship on seventeeth-century immigration to Canada. *Continuity and Change* **4**(1): 79-102.

Gulcher J, Kong A, Stefansson K. 2001. The genealogic approach to human genetics of disease. *Cancer journal* **7**(1): 61-68.

Gusev A, Palamara PF, Aponte G, Zhuang Z, Darvasi A, Gregersen P, Pe'er I. 2012. The architecture of long-range haplotypes shared within and across populations. *Molecular biology and evolution* **29**(2): 473-486.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet* **5**(10): e1000695.

Hall D, Wijsman EM, Roos JL, Gogos JA, Karayiorgou M. 2002. Extended intermarker linkage disequilibrium in the Afrikaners. *Genome research* **12**(6): 956-961.

Hallatschek O. 2011. The noisy edge of traveling waves. *Proc Natl Acad Sci U S A* **108**(5): 1783-1787.

Hallatschek O, Hersen P, Ramanathan S, Nelson DR. 2007. Genetic drift at expanding frontiers promotes gene segregation. *Proc Natl Acad Sci U S A* **104**(50): 19926-19930.

Hallatschek O, Nelson DR. 2008. Gene surfing in expanding populations. *Theoretical population biology* **73**(1): 158-170.

-. 2010. Life at the front of an expanding population. *Evolution* **64**(1): 193-206.

Handley LJ, Manica A, Goudet J, Balloux F. 2007. Going the distance: human population genetics in a clinal world. *Trends in genetics : TIG* **23**(9): 432-439.

Harris EE. 2010. Nonadaptive processes in primate and human evolution. *Am J Phys Anthropol* **143 Suppl 51**: 13-45.

Hastbacka J, de la Chapelle A, Kaitila I, Sistonen P, Weaver A, Lander E. 1992. Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. *Nat Genet* **2**(3): 204-211.

Heath SC, Gut IG, Brennan P, McKay JD, Bencko V, Fabianova E, Foretova L, Georges M, Janout V, Kabesch M et al. 2008. Investigation of the fine structure of European populations with applications to disease association studies. *Eur J Hum Genet* **16**(12): 1413-1429.

Helgason A, Hrafnkelsson B, Gulcher JR, 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. *Am J Hum Genet* **72**(6): 1370-1388.

Helgason A, Palsson G, Pedersen HS, Angulalik E, Gunnarsdottir ED, Yngvadottir B, Stefansson K. 2006. mtDNA variation in Inuit populations of Greenland and Canada: migration history and population structure. *Am J Phys Anthropol* **130**(1): 123-134.

Helgason A, Sigureth ardottir S, Gulcher JR, Ward R, Stefansson K. 2000. mtDNA and the origin of the Icelanders: deciphering signals of recent population history. *Am J Hum Genet* **66**(3): 999-1016.

Helgason A, Yngvadottir B, Hrafnkelsson B, Gulcher J, Stefansson K. 2005. An Icelandic example of the impact of population structure on association studies. *Nat Genet* **37**(1): 90-95.

Henn BM, Cavalli-Sforza LL, Feldman MW. 2012. The great human expansion. *Proc Natl Acad Sci U S A* **109**(44): 17758-17764.

Henripin J, Péron Y. 1973. La transition démographique de la province de Québec. In *La population du Québec: études rétrospectives*, (ed. H Charbonneau), pp. 23-44. Éditions du Boréal Express, Trois-Rivières.

Heutink P, Oostra BA. 2002. Gene finding in genetically isolated populations. *Hum Mol Genet* **11**(20): 2507-2515.

Hewitt GM. 2000. The genetic legacy of the Quaternary ice ages. *Nature* **405**(6789): 907-913.

Heyer E. 1995. Genetic consequences of differential demographic behaviour in the Saguenay region, Quebec. *Am J Phys Anthropol* **98**(1): 1-11.

-. 1999. One founder/one gene hypothesis in a new expanding population: Saguenay (Quebec, Canada). *Hum Biol* **71**(1): 99-109.

Heyer E, Puymirat J, Dieltjes P, Bakker E, de Knijff P. 1997. Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. *Hum Mol Genet* **6**(5): 799-803.

Heyer E, Sibert A, Austerlitz F. 2005. Cultural transmission of fitness: genes take the fast lane. *Trends in genetics : TIG* **21**(4): 234-239.

Heyer E, Tremblay M. 1995. Variability of the genetic contribution of Quebec population founders associated to some deleterious genes. *Am J Hum Genet* **56**(4): 970-978.

Heyer E, Zietkiewicz E, Rochowski A, Yotova V, Puymirat J, Labuda D. 2001. Phylogenetic and familial estimates of mitochondrial substitution rates: study of control region mutations in deep-rooting pedigrees. *Am J Hum Genet* **69**(5): 1113-1126.

Hill J, Griffiths H, Thomas C. 2011. Climate change and evolutionary adaptations at species' range margins. *Annual Review of Entomology* **56**(56): 143-159.

Ho SY, Shapiro B. 2011. Skyline-plot methods for estimating demographic history from nucleotide sequences. *Molecular ecology resources* **11**(3): 423-434.

Holsinger KE, Weir BS. 2009. Genetics in geographically structured populations: defining, estimating and interpreting F(ST). *Nature reviews Genetics* **10**(9): 639-650.

Houdaille J. 1980. Quelques aspects de la démographie ancienne de l'Acadie. *Population* **35**(3): 581-602.

Institut de la statistique du Québec. 2010. Gouvernement du Québec. www.stat.gouv.qc.ca.

International HapMap C Frazer KA Ballinger DG Cox DR Hinds DA Stuve LL Gibbs RA Belmont JW Boudreau A Hardenbol P et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**(7164): 851-861.

Jakkula E, Rehnstrom K, Varilo T, Pietilainen OP, Paunio T, Pedersen NL, deFaire U, Jarvelin MR, Saharinen J, Freimer N et al. 2008. The genome-wide patterns of variation expose significant substructure in a founder population. *Am J Hum Genet* **83**(6): 787-794.

Jakobsson M, Edge MD, Rosenberg NA. 2013. The relationship between F(ST) and the frequency of the most frequent allele. *Genetics* **193**(2): 515-528.

James JW. 1972. Computation of Genetic Contributions from Pedigrees. *Theor Appl Genet* **42**(6): 272-&.

Japanese Archipelago Human Population Genetics C, Jinam T, Nishida N, Hirai M, Kawamura S, Oota H, Umetsu K, Kimura R, Ohashi J, Tajima A et al. 2012. The history of human populations in the Japanese Archipelago inferred from genome-wide SNP data with a special reference to the Ainu and the Ryukyuan populations. *Journal of human genetics* **57**(12): 787-795.

Jarry J, Rioux MF, Bolduc V, Robitaille Y, Khoury V, Thiffault I, Tetreault M, Loisel L, Bouchard JP, Brais B. 2007. A novel autosomal recessive limb-girdle muscular dystrophy with quadriceps atrophy maps to 11p13-p12. *Brain : a journal of neurology* **130**(Pt 2): 368-380.

Jetté R, Gauvreau D, Guérin M. 1991. Aux origines d'une région: le peuplement fondateur de Charlevoix avant 1850. In *Histoire d'un génome Population et génétique dans l'Est du Québec*, (ed. G Bouchard, M De Braekeleer), pp. 75-106. Presses de l'Université du Québec, Sillery, Québec.

Jin W, Wang S, Wang H, Jin L, Xu S. 2012. Exploring population admixture dynamics via empirical and simulated genome-wide distribution of ancestral chromosomal segments. *Am J Hum Genet* **91**(5): 849-862.

Jomphe M. 2011. *Validation des généalogies reconstituées à BALSAC à partir de données génétiques (Document I-C-243)*. Projet BALSAC, Chicoutimi.

Jorde LB, Watkins WS, Kere J, Nyman D, Eriksson AW. 2000. Gene mapping in isolated populations: new roles for old friends? *Human heredity* **50**(1): 57-65.

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C, Eskin E. 2010. Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* **42**(4): 348-354.

Karigl G. 1981. A recursive algorithm for the calculation of identity coefficients. *Annals of human genetics* **45**(Pt 3): 299-305.

Keinan A, Clark AG. 2012. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* **336**(6082): 740-743.

Kere J. 2001. Human population genetics: lessons from Finland. *Annu Rev Genomics Hum Genet* **2**: 103-128.

Kirin M, McQuillan R, Franklin CS, Campbell H, McKeigue PM, Wilson JF. 2010. Genomic runs of homozygosity record population history and consanguinity. *PloS one* **5**(11): e13996.

Klopfstein S, Currat M, Excoffier L. 2006. The fate of mutations surfing on the wave of a range expansion. *Molecular biology and evolution* **23**(3): 482-490.

Kolodny A. 2012. *In search of first contact. The Vikings of Vinland, the people of the Dawnland, and the Anglo-American anxiety of discovery*. Duke University Press, United States of America.

Kristiansson K, Naukkarinen J, Peltonen L. 2008. Isolated populations and complex disease gene identification. *Genome biology* **9**(8): 109.

Kruglyak L. 1999a. Genetic isolates: separate but equal? *Proc Natl Acad Sci U S A* **96**(4): 1170-1172.

-. 1999b. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat Genet* **22**(2): 139-144.

Laberge AM. 2007. La prévalence et la distribution des maladies génétiques au Québec. L'impact du passé sur le présent. *Med Sci (Paris)* **23**: 997-1001.

Laberge AM, Jomphe M, Houde L, Vezina H, Tremblay M, Desjardins B, Labuda D, St-Hilaire M, Macmillan C, Shoubridge EA et al. 2005a. A "Fille du Roy" introduced the T14484C Leber hereditary optic neuropathy mutation in French Canadians. *Am J Hum Genet* **77**(2): 313-317.

Laberge AM, Michaud J, Richter A, Lemyre E, Lambert M, Brais B, Mitchell GA. 2005b. Population history and its impact on medical genetics in Quebec. *Clin Genet* **68**(4): 287-301.

Laberge C. 1967. La consanguinité des Canadiens français. *Population (French Edition)* **22**(5): 861-896.

Labuda D, Lefebvre JF, Nadeau P, Roy-Gagnon MH. 2010. Female-to-male breeding ratio in modern humans-an analysis based on historical recombinations. *Am J Hum Genet* **86**(3): 353-363.

Labuda D, Zietkiewicz E, Labuda M. 1997. The genetic clock and the age of the founder effect in growing populations: a lesson from French Canadians and Ashkenazim. *Am J Hum Genet* **61**(3): 768-771.

Labuda M, Labuda D, Korab-Laskowska M, Cole DE, Zietkiewicz E, Weissenbach J, Popowska E, Pronicka E, Root AW, Glorieux FH. 1996. Linkage disequilibrium analysis in young populations: pseudo-vitamin D-deficiency rickets and the founder effect in French Canadians. *Am J Hum Genet* **59**(3): 633-643.

Lacy R. 1989. Analysis of founder representation in pedigrees: founder equivalents and founder genome equivalents. *Zoo Biology* **8**: 111-123.

Lange K, Sinsheimer JS, Sobel E. 2005. Association testing with Mendel. *Genet Epidemiol* **29**(1): 36-50.

Larmuseau MH, Van Geystelen A, van Oven M, Decorte R. 2013. Genetic genealogy comes of age: perspectives on the use of deep-rooted pedigrees in human population genetics. *Am J Phys Anthropol* **150**(4): 505-511.

Lavoie EM, Tremblay M, Houde L, Vézina H. 2005. Demogenetic study of three populations within a region with strong founder effects. *Community Genet* **8**(3): 152-160.

Lavoie Y. 1972. *L'émigration des Canadiens aux États-Unis avant 1930. Mesure du phénomène.* Presses de l'Université de Montréal.

Lebel RR. 1983. Consanguinity studies in Wisconsin I: Secular trends in consanguineous marriage, 1843-1981. *American Journal of Medical Genetics* **15**: 543-560.

Lesca G, Genin E, Blachier C, Olivieri C, Coulet F, Brunet G, Dupuis-Girod S, Buscarini E, Soubrier F, Calender A et al. 2008. Hereditary hemorrhagic telangiectasia: evidence for regional founder effects of ACVRL1 mutations in French and Italian patients. *Eur J Hum Genet* **16**(6): 742-749.

Leutenegger AL, Prum B, Genin E, Verny C, Lemainque A, Clerget-Darpoux F, Thompson EA. 2003. Estimation of the inbreeding coefficient through use of genomic data. *Am J Hum Genet* **73**(3): 516-523.

Levesque S, Morin C, Guay SP, Villeneuve J, Marquis P, Yik WY, Jiralerspong S, Bouchard L, Steinberg S, Hacia JG et al. 2012. A founder mutation in the PEX6 gene is responsible for increased incidence of Zellweger syndrome in a French Canadian population. *BMC medical genetics* **13**: 72.

Lewontin R. 1965. Selection for colonizing ability. In *The genetics of colonizing species*, (ed. H Baker, G Stebbins), pp. 79–94. Academic Press, New York.

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

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**(5866): 1100-1104.

Livingstone F. 1970. The founder effect and deleterious genes. *Am J Phys Anthropol* **30**: 55-60.

Loewen B, Delmas V. 2012. The Basques in the Gulf of St. Lawrence and Adjacent Shores. *Canadian Journal of Archeology* **36**: 351-404.

Loggia ML, Bushnell MC, Tetreault M, Thiffault I, Bherer C, Mohammed NK, Kuchinad AA, Laferriere A, Dicaire MJ, Loisel L et al. 2009. Carriers of recessive WNK1/HSN2 mutations for hereditary sensory and autonomic neuropathy type 2 (HSAN2) are more sensitive to thermal stimuli. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **29**(7): 2162-2166.

Lohmueller KE, Bustamante CD, Clark AG. 2009. Methods for human demographic inference using haplotype patterns from genomewide single-nucleotide polymorphism data. *Genetics* **182**(1): 217-231.

Lohmueller KE, Indap AR, Schmidt S, Boyko AR, Hernandez RD, Hubisz MJ, Sninsky JJ, White TJ, Sunyaev SR, Nielsen R et al. 2008. Proportionally more deleterious genetic variation in European than in African populations. *Nature* **451**(7181): 994-997.

Luikart G, Allendorf FW, Cornuet JM, Sherwin WB. 1998. Distortion of allele frequency distributions provides a test for recent population bottlenecks. *The Journal of heredity* **89**(3): 238-247.

MacCluer JW, VandeBerg JL, Read B, Ryder OA. 1986. Pedigree analysis by computer simulation. *Zoo Biol* **5**(2): 147-160.

Magnan C, Benoist J. 1969. Les groupes sanguins des Canadiens Français : I. La région du Saguenay Lac Saint-Jean. *L'Anthropologie* **73**(1-2): 49-76.

Malécot G. 1948. *Les mathématiques de l'hérédité*. Masson, Paris.

Marchini J, Cardon LR, Phillips MS, Donnelly P. 2004. The effects of human population structure on large genetic association studies. *Nat Genet* **36**(5): 512-517.

Marjoram P, Tavare S. 2006. Modern computational approaches for analysing molecular genetic variation data. *Nature reviews Genetics* **7**(10): 759-770.

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

Martin AO. 1970. The founder effect in a human isolate: evolutionary implications. *Am J Phys Anthropol* **32**(3): 351-367.

Maruyama T, Fuerst PA. 1984. Population bottlenecks and nonequilibrium models in population genetics. I. Allele numbers when populations evolve from zero variability. *Genetics* **108**(3): 745-763.

-. 1985a. Population bottlenecks and nonequilibrium models in population genetics. II. Number of alleles in a small population that was formed by a recent bottleneck. *Genetics* **111**(3): 675-689.

-. 1985b. Population bottlenecks and nonequilibrium models in population genetics. III. Genic homozygosity in populations which experience periodic bottlenecks. *Genetics* **111**(3): 691-703.

Mayer MF, Boisvert M. 1994. Mortalité infantile et consanguinité dans une ppulaiton endogame du Québec. *Population* **49**(3): 685-724.

Maynard Smith J, Haigh J. 1974. The hitch-hiking efect of a favourable gene. *Genet Res* **23**: 23-35.

Mayr E. 1942. *Systematics and the Origin of Species*. Columbia University Press, New York.

McEvoy BP, Powell JE, Goddard ME, Visscher PM. 2011. Human population dispersal "Out of Africa" estimated from linkage disequilibrium and allele frequencies of SNPs. *Genome research* **21**(6): 821-829.

McGhee R. 1984. Contact between Native North Americans and the Medieval Norse: a review of the evidence. *American Antiquity* **49**(1): 4-26.

McInnis M. 2000. The population of Canada in the nineteenth century. In *A population history of North America*, (ed. MR Haines, RH Steckel), pp. 371-432. Cambridge University Press.

McKusick VA. 2007. Mendelian Inheritance in Man and its online version, OMIM. *Am J Hum Genet* **80**(4): 588-604.

McKusick VA, Hostetler JA, Egeland JA. 1964. Genetic Studies of the Amish, Background and Potentialities. *Bulletin of the Johns Hopkins Hospital* **115**: 203-222.

McQuillan R, Leutenegger AL, Abdel-Rahman R, Franklin CS, Pericic M, Barac-Lauc L, Smolej-Narancic N, Janicijevic B, Polasek O, Tenesa A et al. 2008. Runs of homozygosity in European populations. *Am J Hum Genet* **83**(3): 359-372.

McVean G. 2009. A genealogical interpretation of principal components analysis. *PLoS Genet* **5**(10): e1000686.

McVean GA, Charlesworth B. 2000. The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. *Genetics* **155**(2): 929-944.

Meijer IA, Dupre N, Brais B, Cossette P, St-Onge J, Rioux MF, Benard M, Rouleau GA. 2007. SPG4 founder effect in French Canadians with hereditary spastic paraplegia. *The Canadian journal of neurological sciences Le journal canadien des sciences neurologiques* **34**(2): 211-214.

Mendizabal I, Lao O, Marigorta UM, Wollstein A, Gusmao L, Ferak V, Ioana M, Jordanova A, Kaneva R, Kouvatsi A et al. 2012. Reconstructing the population history of European Romani from genome-wide data. *Current biology : CB* **22**(24): 2342-2349.

Méthot D. 2012. Dossier "Les maladies régionales". L'actualité médicale.

Milot E, Mayer FM, Nussey DH, Boisvert M, Pelletier F, Reale D. 2011. Evidence for evolution in response to natural selection in a contemporary human population. *Proc Natl Acad Sci U S A* **108**(41): 17040-17045.

Molloy M. 1990. Considered affinity: kinship, marriage, and social class in New France, 1640-1729. *Social Science History* **14**(1): 1-26.

Montpetit A, Cote S, Brustein E, Drouin CA, Lapointe L, Boudreau M, Meloche C, Drouin R, Hudson TJ, Drapeau P et al. 2008. Disruption of AP1S1, causing a novel neurocutaneous syndrome, perturbs development of the skin and spinal cord. *PLoS Genet* **4**(12): e1000296.

Moorjani P, Patterson N, Loh PR, Lipson M, Kisfali P, Melegh BI, Bonin M, Kadasi L, Riess O, Berger B et al. 2013. Reconstructing Roma history from genome-wide data. *PloS one* **8**(3): e58633.

Mootha VK, Lepage P, Miller K, Bunkenborg J, Reich M, Hjerrild M, Delmonte T, Villeneuve A, Sladek R, Xu F et al. 2003. Identification of a gene causing human cytochrome c oxidase deficiency by integrative genomics. *Proc Natl Acad Sci U S A* **100**(2): 605-610.

Moreau C, Bhérer C, Vézina H, Jomphe M, Labuda D, Excoffier L. 2011a. Deep human genealogies reveal a selective advantage to be on an expanding wave front. *Science* **334**(6059): 1148-1150.

Moreau C, Lefebvre JF, Jomphe M, Bhérer C, Ruiz-Linares A, Vezina H, Roy-Gagnon MH, Labuda D. 2013. Native American admixture in the Quebec founder population. *PloS one* **8**(6): e65507.

Moreau C, Vézina H, Jomphe M, Lavoie EM, Roy-Gagnon MH, Labuda D. 2011b. When genetics and genealogies tell different stories-maternal lineages in Gaspesia. *Annals of human genetics* **75**(2): 247-254.

Moreau C, Vézina H, Labuda D. 2007. Effet fondateur et variabilité génétique au Québec [Founder effects and genetic variability in Quebec]. *Med Sci (Paris)* **23**(11): 1008-1013.

Moreau C, Vézina H, Yotova V, Hamon R, de Knijff P, Sinnett D, Labuda D. 2009. Genetic heterogeneity in regional populations of Quebec-Parental lineages in the Gaspe Peninsula. *Am J Phys Anthropol*.

Morera B, Barrantes R. 2004. Is the Central Valley of Costa Rica a genetic isolate? *Revista de biologia tropical* **52**(3): 629-644.

Motulsky AG. 1995. Jewish diseases and origins. *Nat Genet* **9**(2): 99-101.

Nance WE, Kearsey MJ. 2004. Relevance of connexin deafness (DFNB1) to human evolution. *Am J Hum Genet* **74**(6): 1081-1087.

Neale RE, Stiller CA, Bunch KJ, Milne E, Mineau GP, Murphy MF. 2013. Familial aggregation of childhood and adult cancer in the Utah genealogy. *International journal of cancer Journal international du cancer*.

Nei M. 1973. Analysis of gene diversity in subdivided populations. *Proc Natl Acad Sci U S A* **70**(12): 3321-3323.

Nei M, Maruyama T, Chakraborty R. 1975. The bottleneck effect and genetic variability in populations. *Evolution* **29**: 1-10.

Newman DL, Hoffjan S, Bourgain C, Abney M, Nicolae RI, Profits ET, Grow MA, Walker K, Steiner L, Parry R et al. 2004. Are common disease susceptibility alleles the same in outbred and founder populations? *Eur J Hum Genet* **12**(7): 584-590.

Nordborg M, Tavare S. 2002. Linkage disequilibrium: what history has to tell us. *Trends in genetics : TIG* **18**(2): 83-90.

Noreau A, Dupré N, Bouchard J-P, Dion PA, Rouleau GA. 2013. Autosomal recessive cerebellar ataxias. In *Handbook of the Cerebellum and Cerebellar Disorders*, (ed. M Manto, JD Gruol, N Schmahmann, N Koibuchi, F Rossi), pp. 2177-2191. Springer Science + Business Media, Dordrecht.

Norio R. 2003a. Finnish Disease Heritage I: characteristics, causes, background. *Human genetics* **112**(5-6): 441-456.

-. 2003b. The Finnish Disease Heritage III: the individual diseases. *Human genetics* **112**(5-6): 470-526.

O'Brien E, Kerber RA, Jorde LB, Rogers AR. 1994. Founder effect: assessment of variation in genetic contributions among founders. *Hum Biol* **66**(2): 185-204.

Ober C, Abney M, McPeek MS. 2001. The genetic dissection of complex traits in a founder population. *Am J Hum Genet* **69**(5): 1068-1079.

Olson MV. 2012. Human genetic individuality. *Annu Rev Genomics Hum Genet* **13**: 1-27.

Orton NC, Innes AM, Chudley AE, Bech-Hansen NT. 2008. Unique disease heritage of the Dutch-German Mennonite population. *American journal of medical genetics Part A* **146A**(8): 1072-1087.

Ostrer H. 2001. A genetic profile of contemporary Jewish populations. *Nature reviews Genetics* **2**(11): 891-898.

Ostrer H, Skorecki K. 2013. The population genetics of the Jewish people. *Human genetics* **132**(2): 119-127.

Otto SP, Whitlock MC. 1997. The probability of fixation in populations of changing size. *Genetics* **146**(2): 723-733.

Ouellet F. 1983. L'accroissement naturel de la population catholique québécoise avant 1850 : aperçus historiographiques et quantitatifs. *L'actualité économique* **59**(3): 402-422.

Pardo LM, MacKay I, Oostra B, van Duijn CM, Aulchenko YS. 2005. The effect of genetic drift in a young genetically isolated population. *Annals of human genetics* **69**(Pt 3): 288-295.

Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. *PLoS Genet* **2**(12): e190.

Peltonen L. 1997. Molecular background of the Finnish disease heritage. *Annals of medicine* **29**(6): 553-556.

Peltonen L, Jalanko A, Varilo T. 1999. Molecular genetics of the Finnish disease heritage. *Hum Mol Genet* **8**(10): 1913-1923.

Peltonen L, Palotie A, Lange K. 2000. Use of population isolates for mapping complex traits. *Nature reviews Genetics* **1**(3): 182-190.

Phillips BL, Brown GP, Shine R. 2010. Life-history evolution in range-shifting populations. *Ecology* **91**(6): 1617-1627.

Piché V. 2003. Un siècle d'immigration au Québec : de la peur à l'ouverture. In *La démographie québécoise Enjeux du XXIe siècle*, (ed. V Piché, C Le Bourdais), pp. 225-263. Les Presses de l'Université de Montréal.

-. 2005. Les vagues migratoires et leur impact : le cas du Québec. *Santé, Société et Solidarité* **4**(1): 19-29.

Plante M, Claveau S, Lepage P, Lavoie EM, Brunet S, Roquis D, Morin C, Vezina H, Laprise C. 2008. Mucolipidosis II: a single causal mutation in the N-acetylglucosamine-1-phosphotransferase gene (GNPTAB) in a French Canadian founder population. *Clin Genet* **73**(3): 236-244.

Pluzhnikov A, Nolan DK, Tan Z, McPeek MS, Ober C. 2007. Correlation of intergenerational family sizes suggests a genetic component of reproductive fitness. *Am J Hum Genet* **81**(1): 165-169.

Pool JE, Hellmann I, Jensen JD, Nielsen R. 2010. Population genetic inference from genomic sequence variation. *Genome research* **20**(3): 291-300.

Poon AH, Laprise C, Lemire M, Montpetit A, Sinnett D, Schurr E, Hudson TJ. 2004. Association of vitamin D receptor genetic variants with susceptibility to asthma and atopy. *American journal of respiratory and critical care medicine* **170**(9): 967-973.

Pouyez C, Lavoie Y. 1983. *Les Saguenayens. Introduction à l'histoire des populations du Saguenay, XVIe-XXe siècles*. Presses de l'Université du Québec, Sillery (Québec) Canada.

Price AL, Helgason A, Palsson S, Stefansson H, St Clair D, Andreassen OA, Reich D, Kong A, Stefansson K. 2009. The impact of divergence time on the nature of population structure: an example from Iceland. *PLoS Genet* **5**(6): e1000505.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**(8): 904-909.

Price AL, Zaitlen NA, Reich D, Patterson N. 2010. New approaches to population stratification in genome-wide association studies. *Nature reviews Genetics* **11**(7): 459-463.

Pritchard JK, Przeworski M. 2001. Linkage disequilibrium in humans: models and data. *Am J Hum Genet* **69**(1): 1-14.

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. 2000. Association mapping in structured populations. *Am J Hum Genet* **67**(1): 170-181.

Provine WB. 2004. Ernst Mayr: Genetics and speciation. *Genetics* **167**(3): 1041-1046.

Prugnolle F, Manica A, Balloux F. 2005a. Geography predicts neutral genetic diversity of human populations. *Current biology : CB* **15**(5): R159-160.

Prugnolle F, Manica A, Charpentier M, Guegan JF, Guernier V, Balloux F. 2005b. Pathogen-driven selection and worldwide HLA class I diversity. *Current biology : CB* **15**(11): 1022-1027.

Puffenberger EG. 2003. Genetic heritage of the Old Order Mennonites of southeastern Pennsylvania. *American journal of medical genetics Part C, Seminars in medical genetics* **121C**(1): 18-31.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**(3): 559-575.

Pysek P, Hulme PE. 2005. Spatio-temporal dynamics of plant invasions : Linking pattern to process. *Ecoscience* **12**(3): 302-315.

Raelson JV, Little RD, Ruether A, Fournier H, Paquin B, Van Eerdewegh P, Bradley WE, Croteau P, Nguyen-Huu Q, Segal J et al. 2007. Genome-wide association study for Crohn's disease in the Quebec Founder Population

identifies multiple validated disease loci. *Proc Natl Acad Sci U S A* **104**(37): 14747-14752.

Rahman P, Jones A, Curtis J, Bartlett S, Peddle L, Fernandez BA, Freimer NB. 2003. The Newfoundland population: a unique resource for genetic investigation of complex diseases. *Hum Mol Genet* **12 Spec No 2**: R167-172.

Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc Natl Acad Sci U S A* **102**(44): 15942-15947.

Ray N, Excoffier L. 2009. Inferring past demography using spatially explicit population genetic models. *Hum Biol* **81**(2-3): 141-157.

Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PL et al. 2010a. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**(7327): 1053-1060.

Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PLF et al. 2010b. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**(7327): 1053-1060.

Reich D, Patterson N, Campbell D, Tandon A, Mazieres S, Ray N, Parra MV, Rojas W, Duque C, Mesa N et al. 2012. Reconstructing Native American population history. *Nature* **488**(7411): 370-374.

Reich D, Thangaraj K, Patterson N, Price AL, Singh L. 2009. Reconstructing Indian population history. *Nature* **461**(7263): 489-494.

Reynolds J, Weir BS, Cockerham CC. 1983. Estimation of the coancestry coefficient: basis for a short-term genetic distance. *Genetics* **105**(3): 767-779.

Richard C, Pennarun E, Kivisild T, Tambets K, Tolk HV, Metspalu E, Reidla M, Chevalier S, Giraudet S, Lauc LB et al. 2007. An mtDNA perspective of French genetic variation. *Ann Hum Biol* **34**(1): 68-79.

Risch N, Tang H, Katzenstein H, Ekstein J. 2003. Geographic distribution of disease mutations in the Ashkenazi Jewish population supports genetic drift over selection. *Am J Hum Genet* **72**(4): 812-822.

Rius M, Darling JA. 2014. How important is intraspecific genetic admixture to the success of colonising populations? *Trends in ecology & evolution*.

Roberts DF. 1968. Genetic effects of population size reduction. *Nature* **220**(172): 1084-1088.

Robertson A. 1960. A theory of limits in artificial selection. *Proc R Soc Lond B* **153**(951): 234-249.

Roddier K, Thomas T, Marleau G, Gagnon AM, Dicaire MJ, St-Denis A, Gosselin I, Sarrazin AM, Larbrisseau A, Lambert M et al. 2005. Two mutations in the HSN2 gene explain the high prevalence of HSAN2 in French Canadians. *Neurology* **64**(10): 1762-1767.

Roman J, Darling JA. 2007. Paradox lost: genetic diversity and the success of aquatic invasions. *Trends in ecology & evolution* **22**(9): 454-464.

Rosenberg NA, Nordborg M. 2002. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature reviews Genetics* **3**(5): 380-390.

Rossignol E, Mathieu J, Thiffault I, Tetreault M, Dicaire MJ, Chrestian N, Dupre N, Puymirat J, Brais B. 2007. A novel founder SCN4A mutation causes painful cold-induced myotonia in French-Canadians. *Neurology* **69**(20): 1937-1941.

Roy R. 1975. La croissance démographique en Acadie de 1671 à 1763. Mémoire de maîtrise (démographie). Université de Montréal. 152 pages.

Roy R, Bouchard G, Declos M. 1988. La première génération de Saguenayens: provenance, apparentement, enracinement. *Cahiers québécois de démographie* **17**: 113-134.

Roy-Gagnon MH, Moreau C, Bhérer C, St-Onge P, Sinnett D, Laprise C, Vezina H, Labuda D. 2011. Genomic and genealogical investigation of the French Canadian founder population structure. *Human genetics* **129**(5): 521-531.

Sabatti C, Service SK, Hartikainen AL, Pouta A, Ripatti S, Brodsky J, Jones CG, Zaitlen NA, Varilo T, Kaakinen M et al. 2009. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat Genet* **41**(1): 35-46.

Saillard J, Forster P, Lynnerup N, Bandelt HJ, Norby S. 2000. mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion. *Am J Hum Genet* **67**(3): 718-726.

Salas A, Richards M, Lareu MV, Sobrino B, Silva S, Matamoros M, Macaulay V, Carracedo A. 2005. Shipwrecks and founder effects: divergent demographic histories reflected in Caribbean mtDNA. *Am J Phys Anthropol* **128**(4): 855-860.

Samuels ME, Majewski J, Alirezaie N, Fernandez I, Casals F, Patey N, Decaluwe H, Gosselin I, Haddad E, Hodgkinson A et al. 2013. Exome sequencing

identifies mutations in the gene TTC7A in French-Canadian cases with hereditary multiple intestinal atresia. *Journal of medical genetics* **50**(5): 324-329.

Scriver CR. 2001. Human genetics: lessons from Quebec populations. *Annu Rev Genomics Hum Genet* **2**: 69-101.

Seldin MF, Shigeta R, Villoslada P, Selmi C, Tuomilehto J, Silva G, Belmont JW, Klareskog L, Gregersen PK. 2006. European population substructure: clustering of northern and southern populations. *PLoS Genet* **2**(9): e143.

Service SK, Ophoff RA, Freimer NB. 2001. The genome-wide distribution of background linkage disequilibrium in a population isolate. *Hum Mol Genet* **10**(5): 545-551.

Sheffield VC, Stone EM, Carmi R. 1998. Use of isolated inbred human populations for identification of disease genes. *Trends in genetics : TIG* **14**(10): 391-396.

Shifman S, Darvasi A. 2001. The value of isolated populations. *Nat Genet* **28**(4): 309-310.

Sibert A, Austerlitz F, Heyer E. 2002. Wright-Fisher revisited: the case of fertility correlation. *Theoretical population biology* **62**(2): 181-197.

Simons YB, Turchin MC, Pritchard JK, Sella G. 2014. The deleterious mutation load is insensitive to recent population history. *Nat Genet* **46**(3): 220-224.

Slatkin M. 1987. Gene flow and the geographic structure of natural populations. *Science* **236**(4803): 787-792.

-. 1995. A measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139**(1): 457-462.

-. 1996. In defense of founder-flush theories of speciation. *The American Naturalist* **147**(4): 493-505.

-. 2004. A population-genetic test of founder effects and implications for Ashkenazi Jewish diseases. *Am J Hum Genet* **75**(2): 282-293.

-. 2008. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nature reviews Genetics* **9**(6): 477-485.

Slatkin M, Excoffier L. 2012. Serial founder effects during range expansion: a spatial analog of genetic drift. *Genetics* **191**(1): 171-181.

Srour M, Bolduc V, Guergueltcheva V, Lochmuller H, Gendron D, Shevell MI, Poulin C, Mathieu J, Bouchard JP, Brais B. 2010. DOK7 mutations presenting as a

proximal myopathy in French Canadians. *Neuromuscular disorders : NMD* **20**(7): 453-457.

Srour M, Hamdan FF, Schwartzentruber JA, Patry L, Ospina LH, Shevell MI, Desilets V, Dobrzeniecka S, Mathonnet G, Lemyre E et al. 2012. Mutations in TMEM231 cause Joubert syndrome in French Canadians. *Journal of medical genetics* **49**(10): 636-641.

Stankovich J, Bahlo M, Rubio JP, Wilkinson CR, Thomson R, Banks A, Ring M, Foote SJ, Speed TP. 2005. Identifying nineteenth century genealogical links from genotypes. *Human genetics* **117**(2-3): 188-199.

Stenico M, Nigro L, Bertorelle G, Calafell F, Capitanio M, Corrain C, Barbujani G. 1996. High mitochondrial sequence diversity in linguistic isolates of the Alps. *Am J Hum Genet* **59**(6): 1363-1375.

Strauss KA, Puffenberger EG. 2009. Genetics, medicine, and the Plain people. *Annu Rev Genomics Hum Genet* **10**: 513-536.

Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**(2): 437-460.

-. 1989. The effect of change in population size on DNA polymorphism. *Genetics* **123**(3): 597-601.

Templeton AR. 2008. The reality and importance of founder speciation in evolution. *BioEssays : news and reviews in molecular, cellular and developmental biology* **30**(5): 470-479.

Tetreault M, Srour M, Allyson J, Thiffault I, Loisel L, Robitaille Y, Bouchard JP, Brais B. 2011. Founder mutation for alpha-sarcoglycan-LGMD2D in a Magdalen Islands Acadian cluster. *The Canadian journal of neurological sciences Le journal canadien des sciences neurologiques* **38**(5): 747-752.

Thiffault I, Rioux MF, Tetreault M, Jarry J, Loiselle L, Poirier J, Gros-Louis F, Mathieu J, Vanasse M, Rouleau GA et al. 2006. A new autosomal recessive spastic ataxia associated with frequent white matter changes maps to 2q33-34. *Brain : a journal of neurology* **129**(Pt 9): 2332-2340.

Thompson EA. 1986. *Pedigree analysis in human genetics*. Johns Hopkins University Press, Baltimore.

Thompson EA, Neel JV. 1978. Probability of founder effect in a tribal population. *Proc Natl Acad Sci U S A* **75**(3): 1442-1445.

-. 1997. Allelic disequilibrium and allele frequency distribution as a function of social and demographic history. *Am J Hum Genet* **60**(1): 197-204.

Thompson EA, Roberts DF. 1980. Kinship structure and heterozygosity on Tristan da Cunha. *Am J Hum Genet* **32**(3): 445-452.

Tonin PN, Perret C, Lambert JA, Paradis AJ, Kantemiroff T, Benoit MH, Martin G, Foulkes WD, Ghadirian P. 2001. Founder BRCA1 and BRCA2 mutations in early-onset French Canadian breast cancer cases unselected for family history. *International journal of cancer Journal international du cancer* **95**(3): 189-193.

Travis JM, Munkemuller T, Burton OJ, Best A, Dytham C, Johst K. 2007. Deleterious mutations can surf to high densities on the wave front of an expanding population. *Molecular biology and evolution* **24**(10): 2334-2343.

Tremblay M. 1997. Transmission intergénérationnelle de la reproduction au Saguenay à la fin du XIXe siècle. *Cahiers québécois de démographie* **26**(1): 129-145.

-. 2010. La contribution des immigrants d'origine germanique au peuplement des régions de Lanaudière, de la Mauricie, de la Montérégie, de Chaudière-Appalaches et du Bas-Saint-Laurent. *Cahiers québécois de démographie* **39**(2): 179-200.

Tremblay M, Arsenault J, Heyer E. 2003. The transmission probabilities of founder genes in five regional populations of Quebec. *Population* **58**(3): 361-380.

Tremblay M, Letendre M, Houde L, Vézina H. 2009. The contribution of Irish Immigrants to the Quebec (Canada) gene pool : An estimation using data from deep-rooted genealogies. *Eur J Population* **25**: 215-233.

Tremblay M, Vézina H. 2000. New estimates of intergenerational time intervals for the calculation of age and origins of mutations. *Am J Hum Genet* **66**(2): 651-658.

Tremblay M, Vézina H, Desjardins B, Houde L. 2008. Distant kinship and founder effects in the Quebec population. *Kinship and Demographic Behavior in the Past International Studies in Population* **7**.

Van Dyken JD, Muller MJ, Mack KM, Desai MM. 2013. Spatial population expansion promotes the evolution of cooperation in an experimental Prisoner's Dilemma. *Current biology : CB* **23**(10): 919-923.

Vézina H. 1996. Démographie génétique et maladies héréditaires au Québec: l'état des recherches. *Cahiers québécois de démographie* **25**(2): 293-322.

Vézina H. 2013. *Projet BALSAC - Rapport annuel 2012-2013*, Chicoutimi (Québec), Canada.

Vézina H, Durocher F, Dumont M, Houde L, Szabo C, Tranchant M, Chiquette J, Plante M, Laframboise R, Lepine J et al. 2005a. Molecular and genealogical characterization of the R1443X BRCA1 mutation in high-risk French-Canadian breast/ovarian cancer families. *Human genetics* **117**(2-3): 119-132.

Vézina H, Jomphe M, Lavoie EM, Moreau C, Labuda D. 2012. L'apport des données génétiques à la mesure généalogique des origines amérindiennes des Canadiens français. *Cahiers québécois de démographie* **41**(1): 87-105.

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

Vézina H, Tremblay M, Houde L. 2004. Mesures de l'apparentement biologique au Saguenay-Lac-St-Jean (Québec, Canada) à partir de reconstitutions généalogiques. *Annales de démographie historique* **2**: 67-84.

Vicoso B, Charlesworth B. 2009. Effective population size and the faster-X effect: an extended model. *Evolution* **63**(9): 2413-2426.

Villanueva P, Jara L, Palomino H. 2010. Association of D16S515 microsatellite with specific language impairment on Robinson Crusoe Island, an isolated Chilean population: a possible key to understanding language development. *Hum Biol* **82**(4): 395-408.

Voight BF, Pritchard JK. 2005. Confounding from cryptic relatedness in case-control association studies. *PLoS Genet* **1**(3): e32.

Wakeley J, King L, Low BS, Ramachandran S. 2012. Gene genealogies within a fixed pedigree, and the robustness of Kingman's coalescent. *Genetics* **190**(4): 1433-1445.

Wallace Linderoth B. 1990. L'Anse aux Meadows: Gateway to Vinland. *Acta Archaeologica* **61**: 166-197.

Walsh LS. 2000. The African American population of colonial United States. In *A population history of North America*, (ed. MR Haines, RH Steckel), pp. 191-240. Cambridge University Press.

Wang S, Lewis C, Jakobsson M, Ramachandran S, Ray N, Bedoya G, Rojas W, Parra MV, Molina J, Gallo C et al. 2007. Genetic Variation and Population Structure in Native Americans. *PLoS Genetics* **3**(11).

Wang S, Ray N, Rojas W, Parra MV, Bedoya G, Gallo C, Poletti G, Mazzotti G, Hill K, Hurtado AM et al. 2008. Geographic patterns of genome admixture in Latin American Mestizos. *PLoS Genet* **4**(3): e1000037.

Waters JM, Fraser CI, Hewitt GM. 2013. Founder takes all: density-dependent processes structure biodiversity. *Trends in ecology & evolution* **28**(2): 78-85.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. *Theoretical population biology* **7**(2): 256-276.

-. 1984. Allele frequencies after a bottleneck. *Theoretical population biology* **26**(3): 387-407.

Wigginton JE, Cutler DJ, Abecasis GR. 2005. A note on exact tests of Hardy-Weinberg equilibrium. *Am J Hum Genet* **76**(5): 887-893.

Wilhelmy J-P. 1984. *Les mercenaires allemands au Québec du XVIIIe siècle et leur apport à la population*. Maison des mots, Beloeil (Québec).

Williams JT. 1993. Origin and population structure of the Icelanders. *Hum Biol* **65**(2): 167-191.

Winney B, Boumertit A, Day T, Davison D, Echeta C, Evseeva I, Hutnik K, Leslie S, Nicodemus K, Royrvik EC et al. 2012. People of the British Isles: preliminary analysis of genotypes and surnames in a UK-control population. *Eur J Hum Genet* **20**(2): 203-210.

Wright S. 1922. Coefficient of inbreeding and relationship. *American Naturalist* **56**: 330-338.

Wright S. 1931. Evolution in Mendelian Populations. *Genetics* **16**(2): 97-159.

-. 1933. Inbreeding and Homozygosis. *Proc Natl Acad Sci U S A* **19**(4): 411-420.

-. 1943. Isolation by Distance. *Genetics* **28**(2): 114-138.

Yamaguchi-Kabata Y, Nakazono K, Takahashi A, Saito S, Hosono N, Kubo M, Nakamura Y, Kamatani N. 2008. Japanese population structure, based on SNP genotypes from 7003 individuals compared to other ethnic groups: effects on population-based association studies. *Am J Hum Genet* **83**(4): 445-456.

Yotova V, Labuda D, Zietkiewicz E, Gehl D, Lovell A, Lefebvre JF, Bourgeois S, Lemieux-Blanchard E, Labuda M, Vezina H et al. 2005. Anatomy of a founder

effect: myotonic dystrophy in Northeastern Quebec. *Human genetics* **117**(2-3): 177-187.

Yotova V, Lefebvre JF, Moreau C, Gbeha E, Hovhannesyan K, Bourgeois S, Bedarida S, Azevedo L, Amorim A, Sarkisian T et al. 2011. An X-linked haplotype of Neandertal origin is present among all non-African populations. *Molecular biology and evolution* **28**(7): 1957-1962.

Zapala MA, Schork NJ. 2006. Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. *Proc Natl Acad Sci U S A* **103**(51): 19430-19435.

Zhang Z, Ersoz E, Lai CQ, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu J, Arnett DK, Ordovas JM et al. 2010. Mixed linear model approach adapted for genome-wide association studies. *Nat Genet* **42**(4): 355-360.

Zlotogora J. 2007. Multiple mutations responsible for frequent genetic diseases in isolated populations. *Eur J Hum Genet* **15**(3): 272-278.

Zlotogora J, Hujerat Y, Barges S, Shalev SA, Chakravarti A. 2007. The fate of 12 recessive mutations in a single village. *Annals of human genetics* **71**(Pt 2): 202-208.