Hum Genet (2011) 129:521–531
DOI 10.1007/s00439-010-0945-x

ORIGINAL INVESTIGATION

# Genomic and genealogical investigation of the French Canadian founder population structure

**Marie-Hélène Roy-Gagnon · Claudia Moreau · Claude Bherer · Pascal St-Onge · Daniel Sinnett · Catherine Laprise · Hélène Vézina · Damian Labuda**

Received: 9 November 2010 / Accepted: 30 December 2010 / Published online: 15 January 2011

**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 (Reich et al. 2009) populations, 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,

**Electronic supplementary material** The online version of this article (doi:10.1007/s00439-010-0945-x) contains supplementary material, which is available to authorized users.

M.-H. Roy-Gagnon (✉)
Department of Social and Preventive Medicine,
Université de Montréal, Montréal, QC, Canada
e-mail: marie-helene.roy-gagnon@umontreal.ca

M.-H. Roy-Gagnon · C. Moreau · C. Bherer · P. St-Onge ·
D. Sinnett · D. Labuda
CHU Ste-Justine Research Center, Montréal, QC, Canada

D. Sinnett · D. Labuda (✉)
Department of Pediatrics, Université de Montréal,
Montréal, QC, Canada
e-mail: damian.labuda@umontreal.ca

C. Laprise
Department of Fundamental Sciences, Université du Québec
à Chicoutimi, Chicoutimi, QC, Canada

C. Laprise · H. Vézina
Interdisciplinary Research Group on Demography and Genetic
Epidemiology (GRIG), Université du Québec à Chicoutimi,
Chicoutimi, QC, Canada

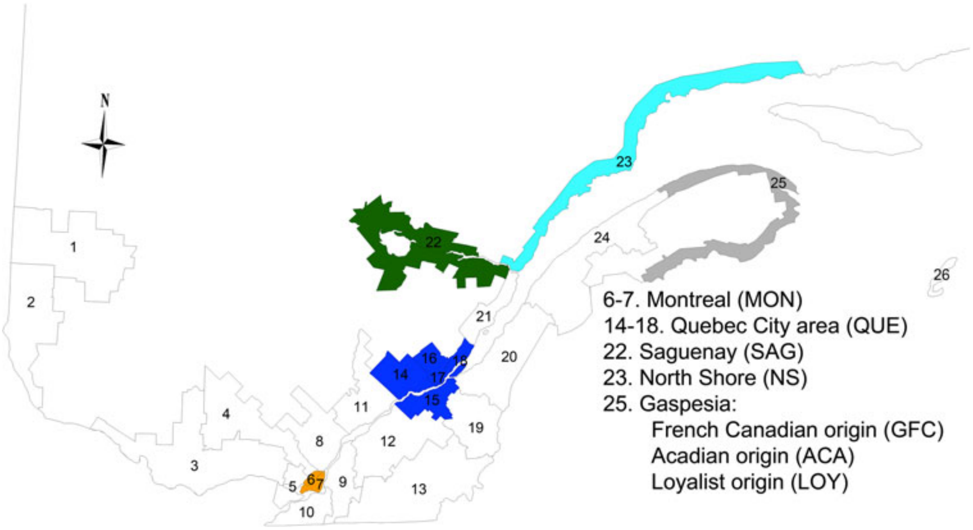
H. Vézina
Department of Human Sciences, Université du Québec
à Chicoutimi, Chicoutimi, QC, Canada

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 (Marchini et al. 2004; Freedman 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 nineteenth and twentieth centuries, immigrants of various origins mixed into the French Canadian population with a very limited genetic impact (Vézina et al. 2005; Bherer et al. 2010). 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 (Fig. 1).

Permanent European settlement began in Gaspesia during the second half of the eighteenth 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 nineteenth 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, Gouvernement 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, Gouvernement du Québec 2010). We also included in our study French Canadians

**Fig. 1** Map of Quebec regions. The regions where participants were recruited for each regional or ethno-cultural sample are indicated



from Montreal and Quebec City. The French-Canadian population of these two cities is not only composed of descendants of the first European settlers but also the migrants from rural regions who moved to urban areas in the context of urbanization and industrialization processes in the nineteenth and twentieth centuries.

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 nineteenth and twentieth centuries (mostly marriage records) (Bouchard and Vezina 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 into 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 (Fig. 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 (The International HapMap Consortium 2007) samples 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 first 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 (second 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 (The International HapMap Consortium 2007) project 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 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; Fig. 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), $p > 0.001$], 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 5,000 kb containing a minimum of 50 SNPs. A maximum of one heterozygote and of five 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 (S-PLUS 8.0. Copyright 1988, 2007 Insightful Corp) 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 Fig. 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 (R Development Core Team 2010) 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 (16–22 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 $\geq$ 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; Fig. S1), consistent with similar distributions of common single nucleotide variations in these three populations (Fig. S2). Correlation of common (MAF $\geq$ 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; Fig. S3).

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 (Fig. S4). However, finer structure could be detected when PCA was performed for Quebec only (Fig. 2a, b). This analysis identified five sub-populations. Eigenvectors 1 and 2 (Fig. 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 (Fig. 2b; Table S3). As observed with PCA from other founder populations (Price et al. 2009; Jakkula et al. 2008), 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 $p < 5.4\mathrm{E}{-25}$; Table S4). A large number of SNPs across the genome contributed to these principal components as opposed to a few highly differentiated SNPs (Fig. S5).

**Table 1** Pairwise Fst statistics for the seven 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 | |

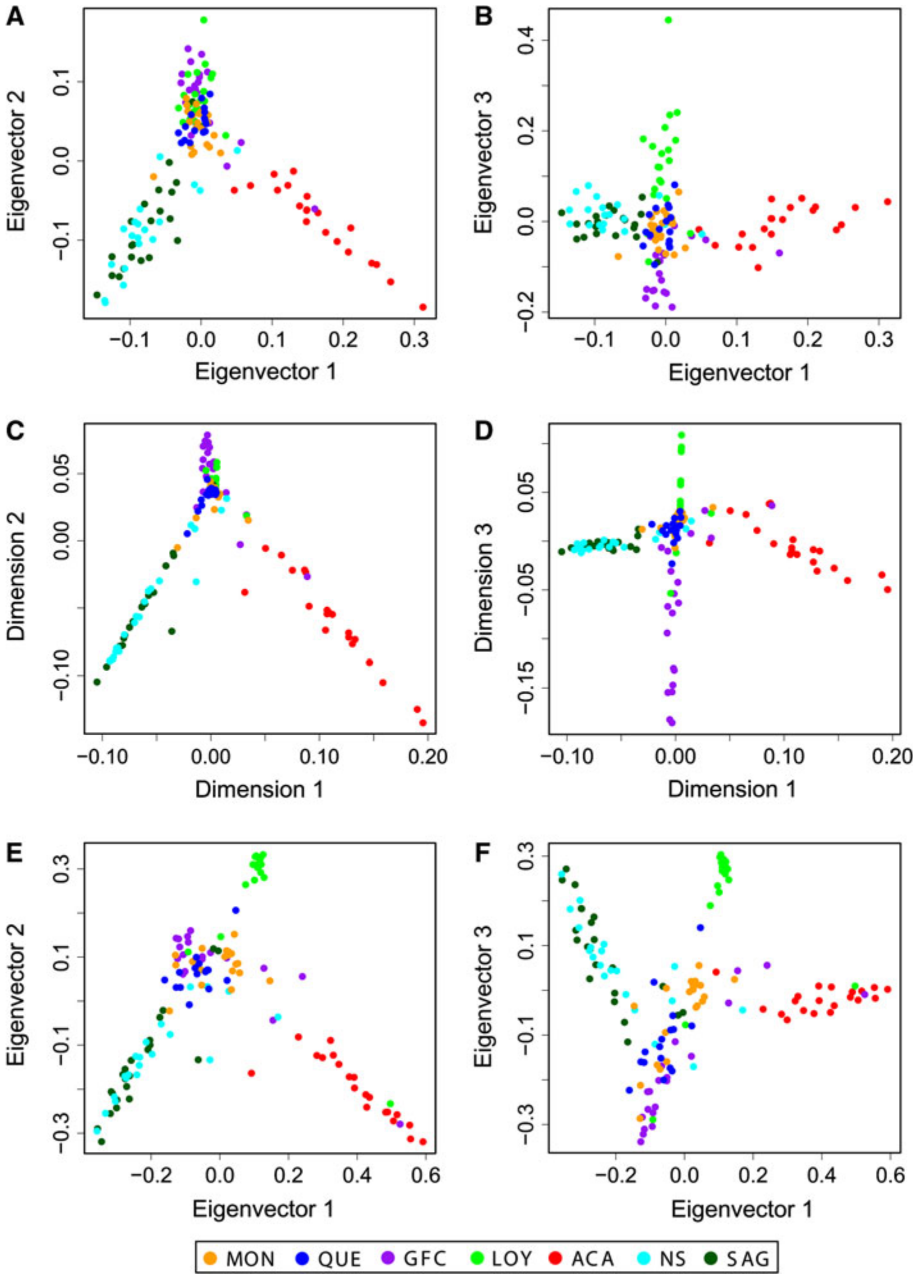
All $p$ values = 0 except MON-QUE $p$ value = 0.153, based on 110 permutations

We 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.

### 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 fifth generation and over 80% up to the ninth generation (Fig. 3). The Gaspesian Loyalists had lower completeness mainly because they arrived later in Quebec and to a lesser extent because

**Fig. 2** Quebec population structure captured by genomic and genealogical 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 genealogical 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



Protestant records were far less complete and well kept than Catholic records (which cover French Canadians and Acadians). At the tenth 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 Figs. 4 and S6. At the tenth generation, estimated average kinship was ≤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 (Fig. 4), indicating some population structure. Kinship also varied between sub-populations, with Loyalists showing the lowest level of relatedness with the others (Fig. 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.

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

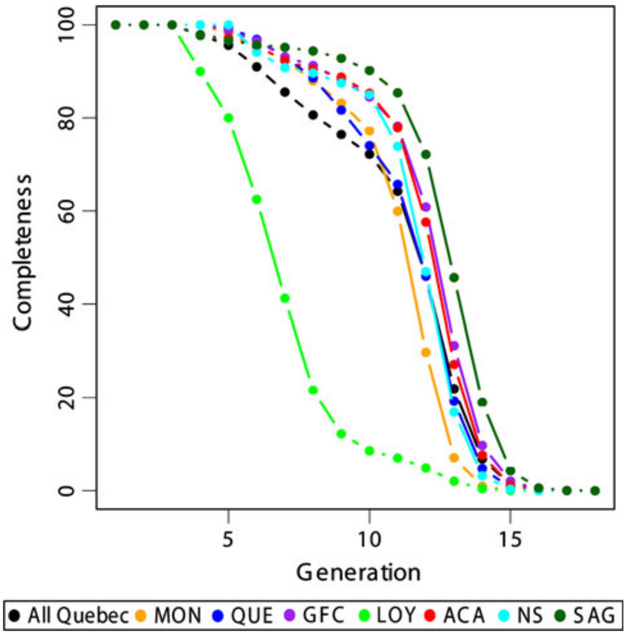


**Fig. 3** Completeness of the genealogical data. *MON* Montreal, *QUE* Quebec City area, *GFC* Gaspesian French Canadians, *LOY* Loyalists, *ACA* Acadians, *NS* North Shore, *SAG* Saguenay

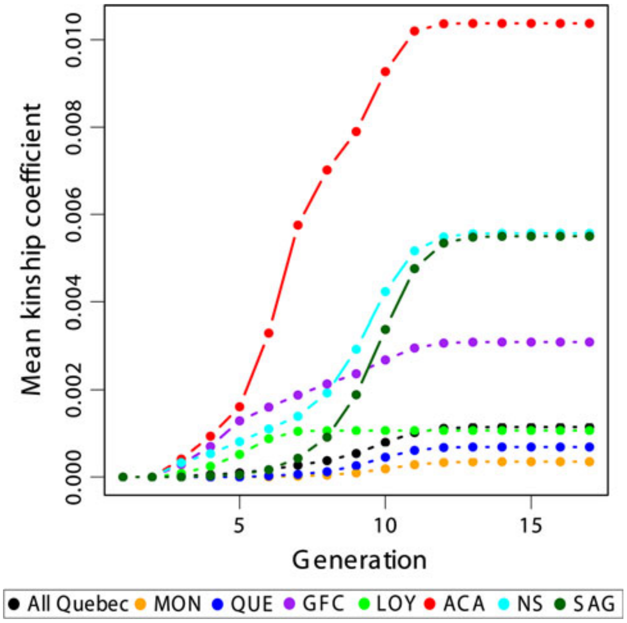


**Fig. 4** Average kinship estimated from genealogical data. *MON* Montreal, *QUE* Quebec City area, *GFC* Gaspesian French Canadians, *LOY* Loyalists, *ACA* Acadians, *NS* North Shore, *SAG* Saguenay

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).

### Linkage disequilibrium and extended runs of homozygosity

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 Fig. 5. LD was slightly higher in Acadian, North Shore, and Saguenay individuals, especially for long-range LD (Fig. 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). Finally, we described extended ROHs within Quebec and compared it to the two reference populations. As shown in Figs. 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; Fig. 6d).

## 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 intermarry. 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 with variation in allele sharing among individuals and further support that the structure observed is consistent with the known founder events.

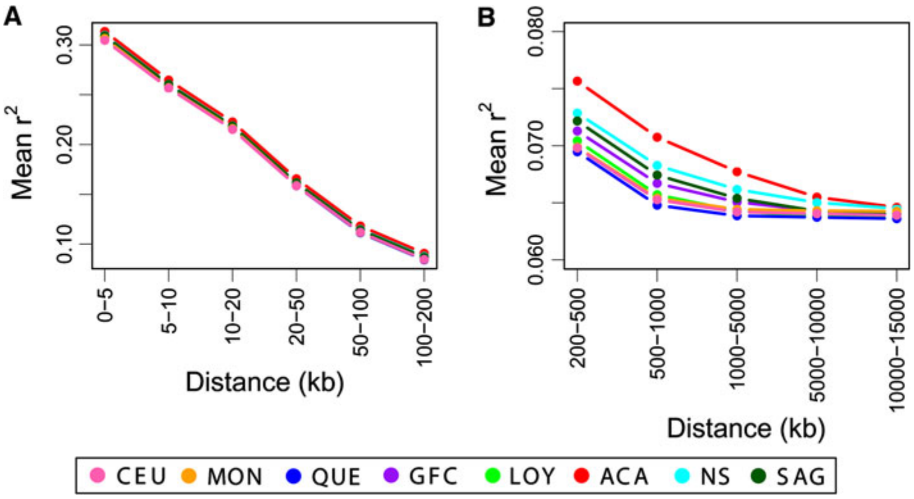


**Fig. 5** Average LD over the genome in Quebec and HapMap CEU. Average $r^2$ estimates shown were obtained from 16 randomly selected individuals from each Quebec regional or ethno-cultural population and HapMap CEU. **a** SNPs located <200 kb apart. **b** SNPs located >200 kb apart. *MON* Montreal, *QUE* Quebec City area, *GFC* Gaspesian French Canadians, *LOY* Loyalists, *ACA* Acadians, *NS* North Shore, *SAG* Saguenay

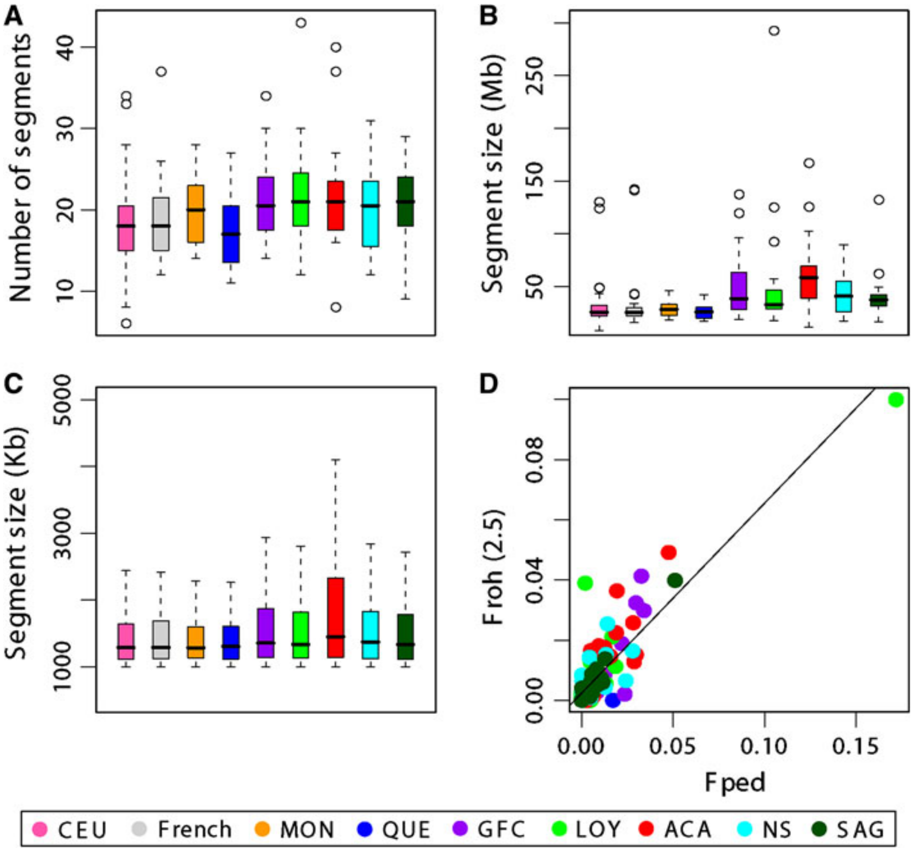


**Fig. 6** Distribution of extended runs of homozygosity (ROHs) for Quebec, HapMap CEU, and HGDP French. **a** Number of ROHs longer than 1 Mb per individual. **b** Total length covered by ROHs per individual. **c** Length of the ROHs that are longer than 1 Mb, outliers excluded for clarity. **d** Correlation between genealogical 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

We have previously shown that population structure could be identified in Quebec from a larger sample of genealogical data only (Bherer et al. 2010). 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 $\geq$ 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 seventeenth 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. 2010). 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; Bherer et al. 2010) as well as from regional partition of hereditary disorders (Scriver 2001) ascribed to local founder effects (Yotova et al. 2005; Labuda et al. 1996).

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 $\lambda$ 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 (Price et al. 2010; Kang et al. 2010; Zhang et al. 2010). More traditional mixed models from the classical polygenic theory (Boerwinkle et al. 1986; Lange et al. 2005; Ober et al. 2001) 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 suggests 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 which makes it an interesting model for these studies.

## Web resources

International HapMap project, http://hapmap.ncbi.nlm.nih.gov/

Human Genome Diversity Panel, ftp://ftp.cephb.fr/hgdp_supp1/

UCSC Genome Browser, http://genome.ucsc.edu/

Online computer program for the Multivariate Distance Matrix Regression, http://polymorphism.scripps.edu/programs.html

PLINK, http://pngu.mgh.harvard.edu/~purcell/plink/

Haploview, http://www.broadinstitute.org/haploview

EIGENSOFT, http://genepath.med.harvard.edu/~reich/Software.htm

Arlequin, http://cmpg.unibe.ch/software/arlequin3/

GenLib, http://www.uqac.ca/grig/

R, http://www.r-project.org/

**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.

## References

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

Bherer C, Labuda D, Roy-Gagnon MH, Houde L, Tremblay M, Vézina H (2010) Admixed ancestry and stratification of Quebec regional populations. Am J Phys Anthropol (in press). doi: 10.1002/ajpa.21424

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. Ann Hum Genet 50(Pt 2):181–194

Bouchard G, Vezina H (2009) Projet BALSAC—Rapport annuel 2008–2009. Université du Québec à Chicoutimi. http://www.uqac.ca/balsac/pdf/ra0708.pdf

Cann HM, de Toma C, Cazes L, Legrand M-F, Morel V, Piouffre L, Bodmer J, Bodmer WF, Bonne-Tamir B, Cambon-Thomsen A, Chen Z, Chu J, Carcassi C, Contu L, Du R, Excoffier L, Friedlaender JS, Groot H, Gurwitz D, Herrera RJ, Huang X, Kidd J, Kidd KK, Langaney A, Lin AA, Mehdi SQ, Parham P, Piazza A, Pistillo MP, Qian Y, Shu Q, Xu J, Zhu S, Weber JL, Greely HT, Feldman MW, Thomas G, Dausset J, Cavalli-Sforza LL (2002) A human genome diversity cell line panel. Science 296(5566):261–262

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

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

Desjardins B (1998) Le Registre de la population du Québec ancien. Ann Demogr Hist 2:215–226

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

Devlin BB, Roeder KK (1999) Genomic control for association studies. Biometrics 55(4):997–1004

Excoffier L, Laval G, Schneider S (2005) Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evol Bioinform Online 1:47–50

Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N, Gabriel SB, Topol EJ, Smoller JW, Pato CN, Pato MT, Petryshen TL, Kolonel LN, Lander ES, Sklar P, Henderson B, Hirschhorn JN, Altshuler D (2004) Assessing the impact of population stratification on genetic association studies. Nat Genet 36(4):388–393

Frenette P (1996) Histoire de la Côte-Nord. Collection Les Régions du Québec. Institut québécois de recherche sur la culture et Presses de l'Université Laval

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

Heath SC, Gut IG, Brennan P, McKay JD, Bencko V, Fabianova E, Foretova L, Georges M, Janout V, Kabesch M, Krokan HE, Elvestad MB, Lissowska J, Mates D, Rudnai P, Skorpen F, Schreiber S, Soria JM, Syvanen AC, Meneton P, Hercberg S, Galan P, Szeszenia-Dabrowska N, Zaridze D, Genin E, Cardon LR, Lathrop M (2008) Investigation of the fine structure of European populations with applications to disease association studies. Eur J Hum Genet 16(12):1413–1429

Institut de la statistique du Québec, Gouvernement du Québec (2010). http://www.stat.gouv.qc.ca/

Jakkula E, Rehnström K, Varilo T, Pietiläinen OPH, Paunio T, Pedersen NL, deFaire U, Järvelin M-R, Saharinen J, Freimer N, Ripatti S, Purcell S, Collins A, Daly MJ, Palotie A, Peltonen L (2008) The genome-wide patterns of variation expose significant substructure in a founder population. Am J Hum Genet 83(6):787–794

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. Ann Hum Genet 45(Pt 3):299–305

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

Lange K, Sinsheimer JS, Sobel E (2005) Association testing with Mendel. Genet Epidemiol 29(1):36–50

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

McQuillan R, Leutenegger AL, Abdel-Rahman R, Franklin CS, Pericic M, Barac-Lauc L, Smolej-Narancic N, Janicijevic B, Polasek O, Tenesa A, Macleod AK, Farrington SM, Rudan P, Hayward C, Vitart V, Rudan I, Wild SH, Dunlop MG, Wright AF, Campbell H, Wilson JF (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

Moreau C, Vezina 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 139(4):512–522

Moreau C, Vezina H, Jomphe M, Lavoie EM, Roy-Gagnon MH, Labuda D (2010) When genetics and genealogies tell different stories—maternal lineages in Gaspesia. Ann Hum Genet (in press). doi:10.1111/j.1469-1809.2010.00617.x

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

Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genet 2(12):e190

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. Am J Respir Crit Care Med 170(9):967–973

Pouyez C, Lavoie Y (1983) Les Saguenayens. Introduction à l'histoire des populations du Saguenay. Presses de l'Université du Québec

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, 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, Zaitlen NA, Reich D, Patterson N (2010) New approaches to population stratification in genome-wide association studies. Nat Rev Genet 11(7):459–463

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000) Association mapping in structured populations. Am J Hum Genet 67(1):170–181

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81(3):559–575

R Development Core Team (2010) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. http://www.R-project.org

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

Scriver CR (2001) Human genetics: lessons from Quebec populations. Annu Rev Genomics Hum Genet 2(1):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):1339

Slatkin M (1995) A measure of population subdivision based on microsatellite allele frequencies. Genetics 139(1):457–462

The International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449(7164):851–861

Vézina H, Tremblay M, Desjardins B, Houde L (2005) Origines et contributions génétiques des fondatrices et des fondateurs de la population québécoise. Cah Que Demogr 34(2):235–258

Voight BF, Pritchard JK (2005) Confounding from cryptic relatedness in case-control association studies. PLoS Genet 1(3):e32

Wigginton JE, Cutler DJ, Abecasis GR (2005) A note on exact tests of Hardy-Weinberg equilibrium. Am J Hum Genet 76(5):887–883

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, Houde L, Tremblay M, Toupance B, Heyer E, Hudson TJ, Laberge C (2005) Anatomy of a founder effect: myotonic dystrophy in Northeastern Quebec. Hum Genet 117(2–3):177–187

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 USA 103(51):19430–19435

Zhang Z, Ersoz E, Lai C-Q, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu J, Arnett DK, Ordovas JM, Buckler ES (2010) Mixed linear model approach adapted for genome-wide association studies. Nat Genet 42(4):355–360