European Journal of Human Genetics (2014) 22, 814–821
© 2014 Macmillan Publishers Limited All rights reserved 1018-4813/14
www.nature.com/ejhg

EJHG Open

ARTICLE

Genome-wide patterns of identity-by-descent sharing in the French Canadian founder population

Héloïse Gauvin[1,2], Claudia Moreau[2], Jean-François Lefebvre[2], Catherine Laprise[3], Hélène Vézina[4], Damian Labuda[2,5] and Marie-Hélène Roy-Gagnon*[,2,6]

In genetics the ability to accurately describe the familial relationships among a group of individuals can be very useful. Recent statistical tools succeeded in assessing the degree of relatedness up to 6–7 generations with good power using dense genome-wide single-nucleotide polymorphism data to estimate the extent of identity-by-descent (IBD) sharing. It is therefore important to describe genome-wide patterns of IBD sharing for more remote and complex relatedness between individuals, such as that observed in a founder population like Quebec, Canada. Taking advantage of the extended genealogical records of the French Canadian founder population, we first compared different tools to identify regions of IBD in order to best describe genome-wide IBD sharing and its correlation with genealogical characteristics. Results showed that the extent of IBD sharing identified with FastIBD correlates best with relatedness measured using genealogical data. Total length of IBD sharing explained 85% of the genealogical kinship's variance. In addition, we observed significantly higher sharing in pairs of individuals with at least one inbred ancestor compared with those without any. Furthermore, patterns of IBD sharing and average sharing were different across regional populations, consistent with the settlement history of Quebec. Our results suggest that, as expected, the complex relatedness present in founder populations is reflected in patterns of IBD sharing. Using these patterns, it is thus possible to gain insight on the types of distant relationships in a sample from a founder population like Quebec.

European Journal of Human Genetics (2014) **22,** 814–821; doi:10.1038/ejhg.2013.227; published online 16 October 2013

Keywords: identity-by-descent; relatedness; founder population; population structure; genealogical data; SNP data

INTRODUCTION

In genetics research, the ability to accurately describe the familial relationships among a group of individuals can be very useful. For example, genome-wide association studies generally assume that studied subjects are independent and this assumption can be assessed easily if the list of their recent ancestors is known and error-free. Almost everybody can identify their parents and generally also their grand-parents or even great-grand-parents. However, most people do not know about their ancestors more remote than two or three generations unless extensive genealogical records are available for the population studied, as in the cases of the Hutterites,[1] Icelanders[2] or Amish,[3] for example.

Another way to describe relationships among individuals in a data set is to look directly at their genome. Recent statistical tools succeeded in assessing the degree of relatedness up to 6–7 generations with good power using identity-by-descent (IBD) sharing.[4] IBD sharing, estimated with genome-wide single-nucleotide polymorphism (SNP) data, is defined as segments of the genome shared identically between two individuals. These chromosome segments are identical-by-state (IBS) and descend from a common ancestor without occurrence of any recombination event.[5] A segment IBD is always IBS but the reverse is not necessarily true unless the time scale is unlimited. In practice, IBD detection from SNPs captures relatively recent ancestry since the resolution of IBD segment detection in a specific data set limits the time scale that can be considered.[6]

Following the important technological innovations that made large amounts of genome-wide SNP data available at reasonable costs, several methods to detect IBD sharing between individuals have been developed. Approaches are generally based on the likelihood that a genetic sequence is IBD, which is measured with a probabilistic model detailing the whole IBD process or using the frequency of haplotypes, where low frequencies of a shared haplotype is an indication of highly probable IBD, or by setting a segment length threshold as a sequence is more likely to be IBD as it is spanning a large chromosomal segment. For example, GERMLINE is a method using a length threshold that builds up a dictionary with chunks of haplotypes and IBD segments are spotted in accordance with a minimal length and with some flexibility as genotyping errors might be present.[7] The most flexible method is the hidden Markov model (HMM) that provides a basic framework to which probabilities for genotyping error and a linkage disequilibrium (LD) model can be added.[8–11] Haplotypes or genotypes can be used and some inference methods also use IBD detection to improve or to perform phasing.[12–14] Simulations studies have shown that more complex models had lower false-discovery rates and higher sensitivity, in particular higher power to detect small segments, resulting in greater accuracy of IBD segment detection.[7–9,12]

[1]Département de médecine sociale et préventive, Université de Montréal, Montréal, Québec, Canada; [2]Centre de recherche, Centre hospitalier universitaire Sainte-Justine, Université de Montréal, Montréal, Québec, Canada; [3]Département des sciences fondamentales, Université du Québec à Chicoutimi, Chicoutimi, Québec, Canada; [4]Département des sciences humaines, Université du Québec à Chicoutimi, Chicoutimi, Québec, Canada; [5]Département de pédiatrie, Université de Montréal, Montréal, Québec, Canada; [6]Department of Epidemiology and Community Medicine, University of Ottawa, Ottawa, Ontario, Canada
*Correspondence: Dr M-H Roy-Gagnon, Department of Epidemiology and Community Medicine, University of Ottawa, Ottawa, Ontario, Canada K1H 8M5. Tel: +1 613 562 5800, extn. 8098; Fax: +1 613 562 5465; E-mail: marie.roy-gagnon@uottawa.ca

Received 25 March 2013; revised 7 August 2013; accepted 4 September 2013; published online 16 October 2013

Most comparisons of IBD inference methods have been conducted in homogeneous, unstructured populations and in simulation frameworks. In fact, to our knowledge, IBD inference methods have not been compared in a real-data setting with extensively documented genealogical records, and genome-wide patterns of IBD sharing have not been described for remote and complex relatedness such as that observed in the French Canadian founder population of the province of Quebec, Canada. The history of the French Canadian founder population begins with French settlers arriving at the beginning of 17th century.[15] Immigration from France ceased with the British Conquest in 1759. From 1755, Acadians, who were descendants of French pioneers who settled in Acadia (located in areas of present-day Nova Scotia, New Brunswick and Prince-Edward Island), started to move to several regions of Quebec, escaping the deportation led by the British.[16] In the last part of the 18th century, American Loyalists, who wanted to stay under the British rule, also moved to Quebec. Meanwhile, the French Canadian population expanded rapidly in relative isolation caused by linguistic, religious and geographic barriers, which amplified the founder effect.[17] As population size grew, settlers colonized new regions of Quebec, including remote and isolated regions, which resulted in population structure.[18,19]

In this study, we focused on three regions: the Saguenay-Lac-St-Jean, the western part of the North Shore and the Gaspe peninsula, as well as the two main cities of the province, Montreal and Quebec City (Figure 1 in Roy-Gagnon *et al*[18]). In Saguenay, French Canadian settlement started around 1840 with the arrival of inhabitants from the neighboring region of Charlevoix. Between 1840 and 1910, 75% of the 30 000 immigrants to Saguenay came from that region.[20] The region of the North Shore was mainly colonized by people from the Charlevoix and Bas-St-Laurent regions between 1840 and 1920.[21] On the other side of the St Laurence River, in Gaspesia, permanent European settlement began some decades earlier. In the second half of the 18th century the Gaspe Peninsula first greeted Acadians. Soon after, Loyalists joined them. Lastly, French Canadians attracted by developing fishing, naval and lumber industries also moved to Gaspesia.[22] These three groups then evolved quite separately as they married mostly among themselves.[22]

During the 19th and 20th century, immigration from various origins mixed into the French Canadian population with a very limited genetic impact and it has been shown that early founders have a greater contribution to the current gene pool.[19,23] Today, about 80% of the 8 million inhabitants of the province is French speaking.[24]

The availability of genealogical data is a major advantage for genetic research in Quebec. Two important population registers exist: the BALSAC population register and the Early Quebec Population Register. The information contained in these databases comes primarily from vital statistics (births, marriages and deaths). As of November 2012, the BALSAC population register contained over 3 million records, which have been computerized and linked to cover the whole province for the 19th and 20th centuries (mostly marriage records).[25] The Early Quebec Population Register contains all records from the beginning of settlement (1608) to 1800 for a total of 700 000 records.[26] Using these population registers, it is possible to reconstruct ascending genealogies of subjects from the present-day population going back over four centuries.

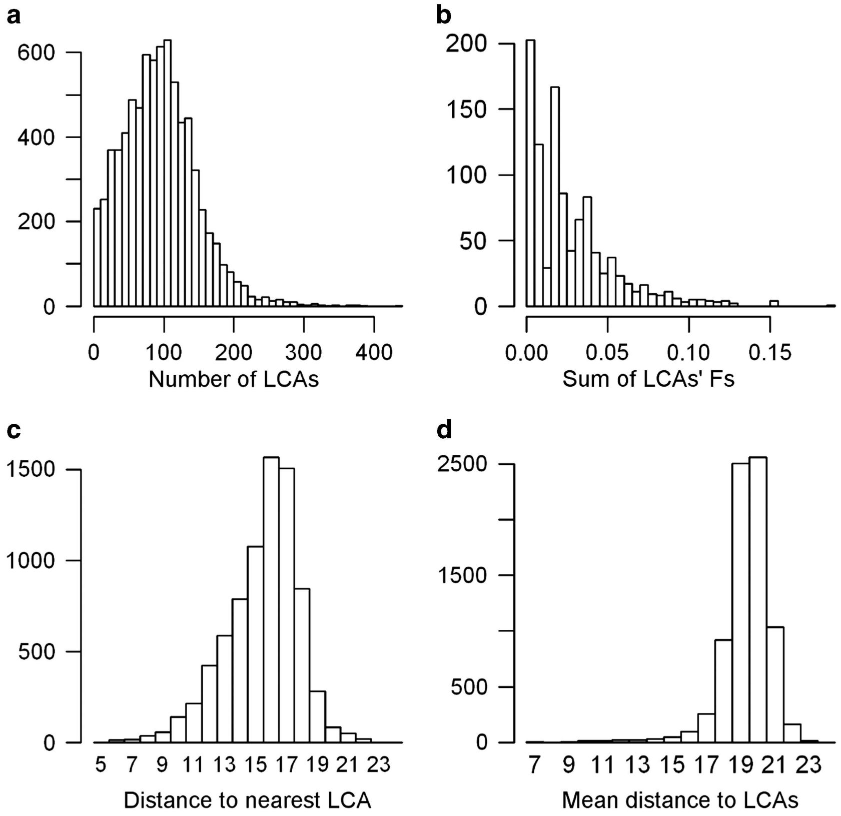

Figure 1 Distributions of genealogical characteristics. Histograms of genealogical characteristics calculated for each of the 7704 related pairs (ie, genealogical kinship >0). (**a**) Number of LCAs; (**b**) sum of LCAs' inbreeding coefficients (Fs) among pairs having at least one inbred LCA (n=1034); (**c**) distance to nearest LCA; and (**d**) mean distance to LCAs.

In this study, we used extensive genealogical data from Quebec in combination with genome-wide SNP data to first compare inference of IBD sharing provided by different methods in order to best describe genome-wide IBD sharing and its correlation with genealogical characteristics. IBD sharing detection was performed on a sample including seven populations of Quebec: French Canadians, Acadians and Loyalists from Gaspesia as well as French Canadians from Saguenay-Lac-St-Jean, North Shore, Quebec City and Montreal. Our analyses showed a good correlation between total length of IBD sharing and genealogical kinship coefficients for most methods with FastIBD yielding the best correlation overall. Using IBD results from FastIBD, we found differences in genome-wide IBD sharing patterns across sub-populations, which reflect genealogical characteristics. This information suggests that IBD sharing can reveal, at least in part, the complex relatedness present in a sample from a founder population like Quebec.

MATERIAL AND METHODS

Study population

The data consist of 143 individuals from a previously reported sample from seven sub-populations of Quebec.[18] Recruitment criteria focused on the geographical origin of participants, and as much as possible, we recruited participants with at least one parent born in the region before 1960 or who were themselves born in the region before 1960. For all individuals, using the BALSAC population register and the early Quebec Population Register, genealogies were reconstructed as far back as possible and confirmed the absence of closely related individuals (first cousins and closer) in the sample. All participants gave their informed consent and the CHU Sainte-Justine Ethics Committee approved the study protocol.

For comparison purposes, we downloaded the original CEU sample (II + III) from the International HapMap project.[27] We excluded two highly related individuals,[28] leading to a set of 109 individuals (more distantly related than cousins) with North–Western European origin (see Supplementary Table S1).

Genotyping and quality control

Sample from Quebec was genotyped on Illumina HumanHap650Y arrays at the McGill University and Genome Quebec Innovation Center. Quality control procedures were the same as in the first publication using this sample.[18] Briefly, quality check was performed to retain individuals and SNPs with at least 90% genotypes and to select only common autosomal SNPs (MAF$>$5%) in Hardy–Weinberg equilibrium (exact test,[29] $P>0.001$). These restrictions yielded 140 individuals (20 Gaspesian French Canadian, 20 Acadians, 20 Loyalists, 22 from Saguenay-Lac-St-Jean, 20 from the North Shore, 16 from Quebec City and 22 from Montreal) and 539 742 SNPs. The same quality control criteria were applied to HapMap CEU, yielding 538 776 SNPs. All genomic positions are according to NCBI build 37.

Genealogical data and associated measures

The completeness of the genealogical data is measured by the proportion of ancestors observed (i.e., ancestors for whom information is available) in the data at a given generation divided by the expected number of ancestors. The completeness of the genealogical data of our sample of 140 individuals is over 90% up to the 5th generation and over 80% up to the 9th generation, except for the Gaspesian Loyalists (see Roy-Gagnon *et al*[18] for a more detailed description of completeness in these data). The lower amount of genealogical information available for the Loyalists sample is mainly due to their later arrival in Quebec and, to a lesser extent, to the fact that Protestant records were less complete and less well kept than Catholic records (which cover French Canadians and Acadians).

To describe the sample, kinship and inbreeding coefficients were calculated using the S-Plus 8.0 (S-PLUS 8.0. Copyright 1988, 2007 Insightful Corp) function library GenLib. This library implements the algorithm of Karigl to calculate kinship coefficients.[30] We also used PedHunter software[31] to get the set of lowest common ancestors (LCAs) for each pair of individuals. LCAs are the most recent ancestors shared by a pair of individuals. A pair can have more than one LCA as long as no ancestor in the set of LCAs shares a descendant who is also an ancestor of the pair of individuals. We also obtained, using PedHunter, the length of the shortest paths from one member of the pair to the other member through their LCAs, named hereafter distances to LCAs. Once each set of LCAs was obtained, we calculated the inbreeding coefficients of these LCAs. We used the sum of these inbreeding coefficients to measure the total amount of inbreeding present among the LCAs.

Genomic IBD sharing

We selected five different methods to perform the detection of IBD segments, all using a probabilistic framework except GERMLINE. GERMLINE is a computationally efficient software implementing a method that builds a dictionary of haplotypes to find matches between individuals. These matches are then extended to identify long shared segments, while allowing some flexibility by assuming an error rate per SNP in order to avoid too many false negatives caused by genotyping inaccuracies.[7] Other methods are largely based on hidden Markov models (HMM). PLINK is the simplest method as it does not allow genotyping error and assumes that SNPs are in approximate linkage equilibrium.[11] IBDLD incorporates potential genotyping errors and missing data and has an extension for LD.[9] The FastIBD method also includes a LD model when estimating IBD. The inference is conducted on sampled haplotypes for which an IBD score is calculated using shared haplotype frequency. Detected tracts are then extended and identified as being IBD according to a threshold set on score values.[8] The last method that we considered, SLRP, also uses a HMM to approximate the IBD process while considering a genotyping error rate.[12]

For all methods default parameters were used and some data manipulations were performed when necessary (Supplementary Table S2). For PLINK, which does not include LD, we did SNP pruning (pairwise $r^2<0.2$ in sliding windows of size 50 shifting every 5 SNPs) leading to a subset of 65 959 SNPs. For GERMLINE, we phased data with two different methods; Beagle version 3.3.1[32] and ShapeIT version 1.378.[33] For all analyses, we kept only segments greater than or equal to 2 cM, corresponding to the expected length of segments for common ancestors up to 25 generations ago.[34,35] This length ensures a good sensitivity and limits the false-discovery rate.[7,8,34,36]

Statistical analysis

We first examined the correlation between IBD sharing identified with the different methods and genealogical kinship coefficients. We used the total length of all segments shared IBD and calculated Pearson's correlation coefficients. Assuming that genealogical kinship is the true expected kinship, we selected the method providing the best correlation as the best method for our population and retained this method for further analyses. We also examined the distribution of the lengths of the IBD segments identified by each method and we considered computation time.

We then examined the relationships between genomic IBD sharing and genealogical characteristics using simple linear regression models. We also looked at genomic sharing in pairs of individuals with or without at least one inbred LCA. Lastly, we investigated differences in IBD among the sub-populations. We plotted the average number of segments of a certain size shared per pair of individuals and also the proportion of pairs of individuals having IBD sharing at each position on the genome.

RESULTS

Genealogical description

Levels of relatedness among individuals within the different sub-populations, as measured by the kinship coefficients estimated from the genealogical data, vary greatly (Supplementary Figure S1). As described in Roy-Gagnon *et al*,[18] people from Saguenay and North Shore as well as Acadians had higher levels of kinship, while populations from Montreal and Quebec City areas were less related. These observations are consistent with the settlement history of the province of Quebec and are also supported by previous findings based on genealogical data that emphasized a West–East decreasing gradient of diversity among regional populations as well as a stratification of regional populations.[19]

Figure 1a presents the distributions of the number of LCAs per pair of individuals excluding unrelated pairs according to the genealogical kinship coefficients (ie, pairs of individuals with kinship = 0). Pairs of individuals with kinship value equals to zero are pairs unrelated relatively to the time scale considered or related but without enough genealogical information available to support the relationship. In the whole sample, the average number of LCAs per pair of individuals was 74, ranging from 2–433. Average numbers of LCAs for the sub-populations ranged from 2.3 (Loyalists) to 152.6 (Montreal area), and the distributions were significantly different among sub-populations (all Kolmogorov–Smirnov test *P*-values <0.007). For each related pair, we also looked at the distance to the most recent LCA and the mean distance to LCAs (Figure 1c and d), which were on average 15.5 (ranging from 5–24) and 19.8 (ranging from 7–24), respectively. These distributions were also significantly different across populations (*P*-values <0.02) except for minimal distance to LCA for Loyalists compared with Acadians and Gaspesian French Canadians compared with North Shore.

We also described inbreeding among LCAs. Only 13% of pairs of related individuals had one inbred LCA or more but this percentage varied greatly from one population to another. The proportions of pairs with at least one inbred LCA was more than half for the Saguenay, North Shore and Acadian populations, 31% for Gaspesian French Canadians and $<8\%$ for the other populations. The number of inbred LCAs for a pair of individuals with LCAs ranged from 0–13 and inbreeding coefficients ranged from 0.00006–0.06, which are approximately equivalent to individuals with parents that are seventh-degree relatives and first cousins, respectively. Figure 1b shows the sum of all LCAs' inbreeding coefficients, which is the measure that we chose to summarize the inbreeding information. This sum ranged from 0.0001–0.2 for pairs of individuals with at least one inbred LCA. Overall, distributions of genealogical characteristics reflect the diversity and complexity of the relationships present in the structured founder population of Quebec.

Comparison of different IBD sharing detection methods

Before comparing results from selected methods, we looked at results from the only method using phased data, GERMLINE, for which we used two different phasing methods (ShapeIT and Beagle). Haplotypes obtained with different phasing methods are not consistent and this might impact IBD inference. Indeed, data phased with ShapeIT provided IBD results that were more strongly correlated with the genealogical information than those phased with Beagle. The correlation between total length of IBD segments and genealogical kinship coefficients for results from GERMLINE was 0.92 for genotype phased with ShapeIT and 0.72 for genotype phased with Beagle. Hence, we retained GERMLINE's results with ShapeIT phasing for further analyses.

In the whole sample from the Province of Quebec, we observed Pearson's correlation coefficients ranging from 0.69–0.92 for the total length of IBD sharing identified with the different methods against the genealogical kinship coefficient (Table 1, Supplementary Figure S2). Three methods (GERMLINE, FastIBD and IBDLD) stand out with correlation coefficients of 0.92. IBD sharing identified by PLINK and SLRP was less concordant with genealogical information.

To get a better idea of which method provided the most appropriate results for our data, we further examined the correlation between IBD sharing and kinship in each sub-population separately. Correlations varied across populations (Table 1). We noted the low correlation of IBD sharing inferred by some methods in the Saguenay region with kinship coefficient despite the presence of a noteworthy degree of relatedness among individuals in this region. Less surprisingly, populations with lower expected relatedness, such as Montreal and Quebec City areas, had lower correlations with values ranging from 0.02–0.46. We also noted that correlations found for the Loyalists were either very good (0.84–0.86) or very weak (−0.03 or 0.01).

Table 1 Pearson's correlation coefficients between total length of IBD sharing and kinship coefficients for each population and each method

	Methods				
Population	*PLINK*	*GERMLINE*	*FastIBD*	*IBDLD*	*SLRP*
ACA	0.87	0.88	0.89	0.89	0.85
GFC	0.92	0.91	0.92	0.92	0.88
LOY	−0.03	0.84	0.86	0.86	0.01
MON	0.10	0.39	0.46	0.45	−0.02
NS	0.85	0.88	0.90	0.88	0.83
QUE	0.09	0.31	0.45	0.42	0.15
SAG	0.15	0.82	0.84	0.83	0.13
PQ	0.77	0.92	0.92	0.92	0.69

Abbreviations: ACA, Acadians; GFC, Gaspesian French Canadians; LOY, Loyalists; NS, North Shore; MON, Montreal; QUE, Quebec City area; SAG, Saguenay; PQ, whole sample from the Province of Quebec.

Results from FastIBD were retained for further analyses. Assuming that genealogical kinship is the true expected kinship, FastIBD was among the fastest (see Supplementary Table S3 for detailed information on computation time) and best reflected the relatedness described by our genealogical data, as evaluated by the correlation between total length of IBD sharing and genealogical kinship coefficient.

Genealogical measures *versus* inferred IBD sharing

Before looking at the relationship between IBD sharing and different genealogical variables, we examined the impact of genealogical completeness on the correlation between total length of IBD sharing and the genealogical kinship coefficient. We recalculated the correlation coefficients with pairs of individuals having $>50\%$ of their genealogical information complete at the 5th generation and also with the same completeness at the 10th generation. Almost no change was observed at the 5th generation, while at the 10th changes in correlation coefficients were small (0–0.10, all within one s.d. of the estimates) except for the Loyalists that did not have enough complete pairs at the 10th generation to recalculate the correlation. We chose to keep all pairs in our sample.

As IBD sharing was highly correlated with genealogical kinship coefficient, a simple linear regression fits the data well. Hence, the overall degree of relatedness is well captured by overall IBD sharing with 85% of the variance in kinship coefficients explained by total length of IBD sharing (Figure 2a). Total length of IBD sharing also reflected characteristics of relatedness, such as shorter distance to LCA or having an inbred ancestor. Total length of IBD sharing explained 26% of the variance in the mean distance to LCAs, 39% of the variance in the distance to the nearest LCA and 31% of the variance in the sum of LCAs' inbreeding coefficients (Figure 2b–d). As pairs of individuals sharing an inbred common ancestor seemed to be a distinct group we separated the whole sample based on this criterion to assess the impact on IBD sharing. Comparing the two groups obtained, we observed significantly more IBD sharing for pairs having

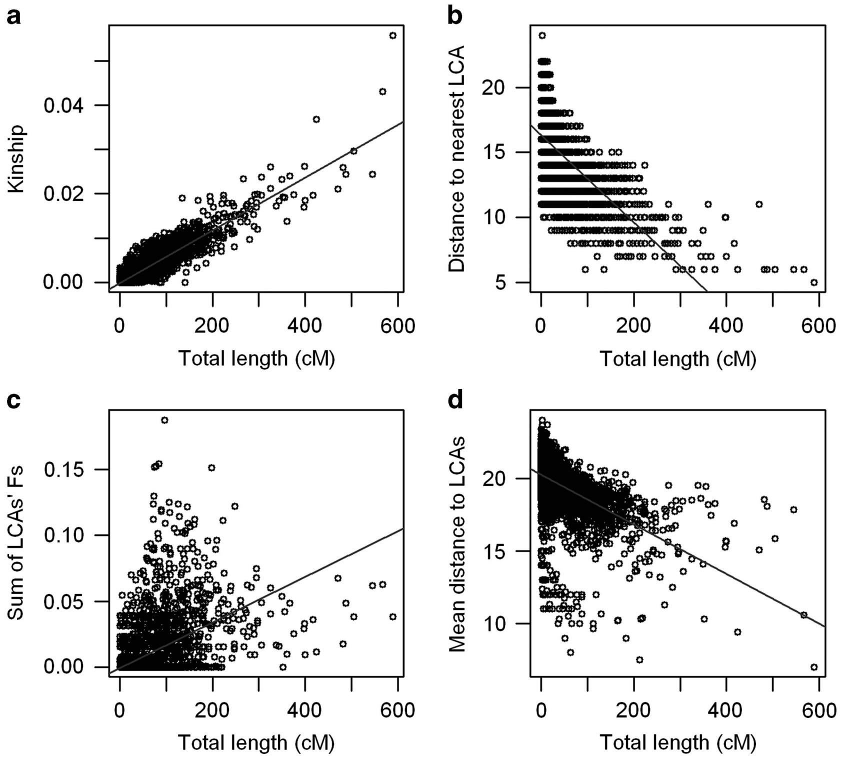

Figure 2 IBD sharing and genealogical characteristics. Scatter plots of total length of IBD sharing *versus* genealogical characteristics for each pair ($n=9730$): (**a**) kinship coefficients; (**b**) distance to nearest LCA; (**c**) sum of LCAs' inbreeding coefficients (Fs); and (**d**) mean distance to LCAs. A simple linear regression line is plotted in gray on each graph.

at least one inbred LCA (Supplementary Figure S3). These pairs have, on average, 7.2 times more total length of IBD sharing and 4.5 times more IBD segments.

IBD sharing in populations

The amount of IBD sharing per population is shown on Figure 3. Each dot represents the mean number of segments shared per pair for specific length ranging from 2–15 cM. The number of segments and their length vary with the degree of relationships, yielding distinct curves for the different levels of kinship present in the populations. The Acadians, which have the highest levels of kinship, have a curve well above the other populations. The Saguenay and North Shore curves overlap, reflecting similar kinship levels in these two populations. Montreal and Quebec City show lower and more variable levels of IBD sharing. We also observed a clear difference between our whole sample from Quebec and the HapMap CEU sample. On average pairs of individuals from Quebec shared 3.8 IBD segments and have 21.3 cM of IBD sharing, while those from HapMap CEU share 2.7 IBD segments and have 8.0 cM of IBD sharing. Thus, pairs of individuals from Quebec also shared longer segments, with segments smaller than 5 cM representing 63 and 96% of segments for Quebec and HapMap CEU, respectively.

Whole-genome IBD sharing

Figure 4 shows the proportions of pairs of individuals having IBD sharing at specific chromosomal positions across the whole genome. Patterns across populations are different and, as in Figure 3, we can see that average sharing differs among populations, with the CEU sharing less than the Quebec population. Some IBD sharing seems consistent across populations, for example, around the HLA region on chromosome 6 where a peak can be observed for the whole Quebec sample and CEU sample.

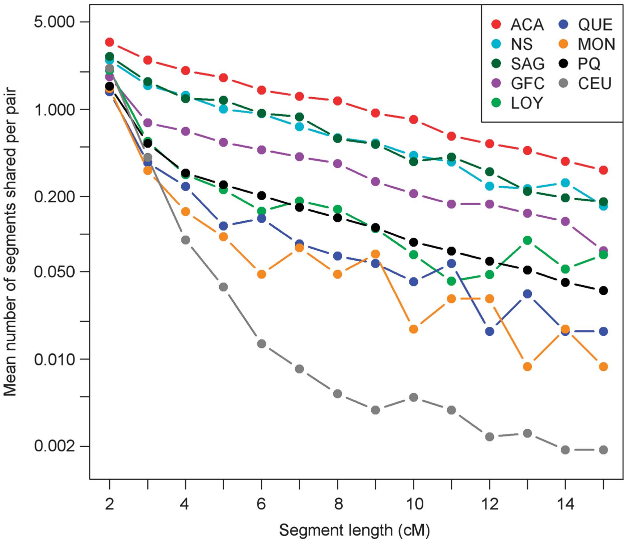

Figure 3 Pairwise IBD sharing in each population. The mean number of segments shared is shown (*y* axis, log-scale) per pair for specific 1 cM class length ranging from 2–15 cM. ACA, Acadians, GFC, Gaspesian French Canadians; LOY, Loyalists; NS, North Shore; MON, Montreal; QUE, Quebec City area; SAG, Saguenay, PQ whole sample from the Province of Quebec, CEU HapMap.

DISCUSSION

In this study, we first compared IBD inference provided by five different methods by correlating total length of IBD sharing with

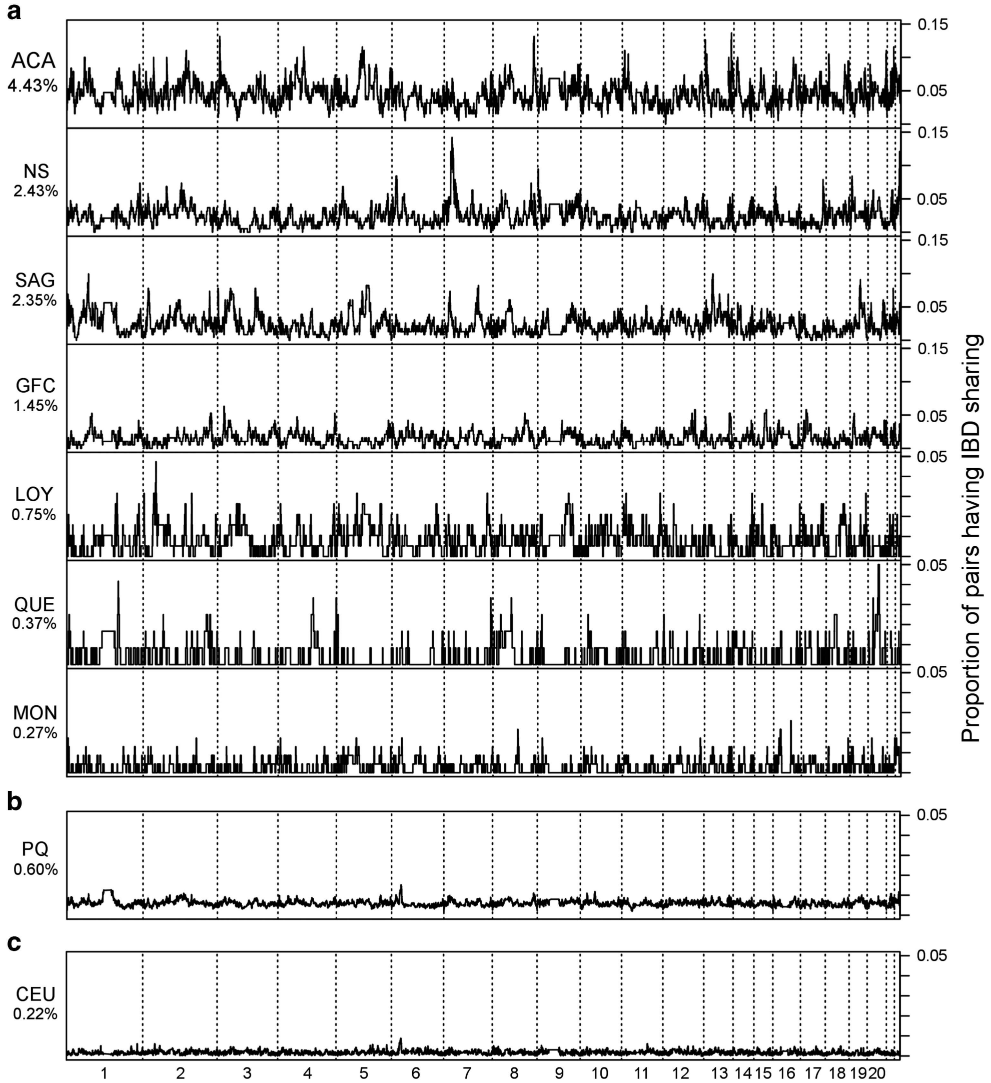

Figure 4 Genome-wide patterns of IBD sharing in each population. The proportion of IBD sharing across the genome is shown for each population with vertical dashed lines to separate chromosomes and mean proportion of sharing indicated on the left. The proportion was calculated at each position that was genotyped in our data. Note that the scale is different for the first four graphs in panel (**a**). (**a**) Quebec populations: ACA, Acadians; GFC, Gaspesian French Canadians; LOY, Loyalists; NS, North Shore; MON, Montreal; QUE Quebec City area; SAG, Saguenay; (**b**) PQ whole sample from the Province of Quebec (**c**) CEU, HapMap.

genealogical kinship. To our knowledge, our study is the first to provide a comparison of the performance of different methods in a complex data set from a founder population. It is difficult to evaluate the performance of methods in a real-data setting as we do not know the truth. Because of the availability of extended genealogies in our population, we could evaluate, at least in part, the performance of IBD detection methods by comparing them with genealogical information. Our results confirmed the importance of a well-defined and flexible model or algorithm for IBD inference and identified FastIBD, GERMLINE and IBDLD as the best-performing methods based on the high correlations between total length of IBD segments shared by pairs of individuals and their genealogical kinship coefficient. As noted in previous studies, simple models that do not consider genotyping errors and LD, such as that implemented in PLINK, yield lower resolution of IBD detection.[7,36] In our sample, the smallest segment detected with PLINK was 3.8 cM long, almost twice as our threshold of 2 cM. With respect to genotyping errors, a modification of PLINK has been proposed in order to include genotyping confidence scores into the IBD inference process, which could improve IBD inference.[37] SLRP has previously been shown to yield a high accuracy of IBD detection compared with GERMLINE and FastIBD in simulated data.[12] However, in our population, SLRP identified more IBD sharing than the other methods, while yielding the lowest correlation coefficients with genealogical kinship overall.

We selected FastIBD for further analyses, because IBD sharing within populations was more associated with genealogical information with this method and it was fast to run. We recognize that we did not optimize the parameters selected for each method but simply used the ones recommended by the authors for their methods. Parameter optimization could have affected our comparison and improved our results. However, our results are consistent with most simulations reported in the literature comparing different methods. We also restricted our study to five methods as other existing methods were more difficult to use or not implemented in a software.[10,34,38,39]

Using results from FastIBD, we then related IBD sharing to the different genealogical measures. Total length of IBD sharing explained a large portion of the variance in kinship coefficients. Our results highlight the variability in realized IBD sharing for a variety of pairs of remotely related individuals with known kinship. Not surprisingly, total length of IBD sharing also explained more of the variance in the distance to nearest LCAs than of the variance in mean distance to LCA as the most recent ancestors have a higher impact on IBD sharing. We aggregated inbreeding coefficients from LCAs into a unique sum and found that IBD sharing also explained a noteworthy part of its variance. However, we are conscious that a pair of individuals could have no inbred LCA identified but still share a more distant inbred ancestor. This occurred for only 20 pairs of individuals. Some shared inbred LCAs may also not be identified because of lower genealogical completeness. This might explain a few pairs of individuals without inbred LCAs (shown as outliers on Supplementary Figure S3) that had an important amount of IBD sharing compared with their group average and that had inbred ancestors that were not shared according to the information available.

Length of segments identified in the different populations was also a good way to identify population differences. The odds of sharing more segments as well as longer segments were higher in population with more relatedness, as expected. Furthermore, IBD sharing in the whole sample was very high and, as expected, mean length of segments inferred (data not shown) was higher than in any other HapMap or Ashkenazi Jewish populations considered in Gusev *et al*[40] except for one sample in which many pairs were closely related (closer than 1st degree cousin according to IBD inference). The high IBD sharing and increased proportion of longer segments is explained by founder events that occurred and population expansions following them.[41] The fact that the Saguenay population size underwent a 25-fold increase in only a century, from 1861–1961, while the whole Quebec population increased about five times, is in good conjunction with our results.

Despite important differences in mean proportion of IBD sharing between Quebec and HapMap CEU, we noted the presence of a common peak of IBD on chromosome 6 covering the HLA region. Increased IBD sharing has been reported for this region in several populations and could be the results of selection.[40,42] As pointed by Browning and Browning,[35] IBD inference will be facilitated in region with high LD but LD may also lead to overestimating the true IBD sharing relative to recent common ancestor. Knowing that important LD normally arises in presence of natural selection, the relevance of an excess of IBD sharing in the HLA region should be investigated more deeply.

IBD detection methods are useful in many contexts such as identifying phasing errors or polymorphic deletions, estimating heritability,[7,43] inferring kinship[4,38,44–46] and mapping diseases in association studies.[47,48] In cases, where genealogical information is not available we now know that IBD is an alternative to account for unknown relatedness.[49] Even in samples that are widely used such as those coming from CEPH, precautions are necessary as important consanguinity as been identified recently.[46] Our study is an additional example putting forward the importance of considering relatedness in a sample before studying it. The high correlation that we observed between genealogical information and IBD sharing, over the wide range of remote relatedness present in our study population, further demonstrates the usefulness of genomic IBD detection to capture even complex relatedness involving inbreeding and our findings can guide the interpretation of results in other population without genealogical data. Our study highlights the great variety in types of relatedness present in the French Canadian founder population and how this complex relatedness is reflected in patterns of IBD sharing. Using these patterns, it is thus possible to gain insight on the types of distant relatedness in a sample from a founder population like Quebec, leading to better genetic study design and analysis.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We are grateful to all participants who generously shared their DNA and information required to reconstruct their genealogies. This work was supported by a Grant from the Fonds de la Recherche en Santé du Québec (FRSQ) to MHRG. Support from the FRSQ Réseau de Médecine Génétique Appliquée (RMGA) and the Canadian Institutes of Health Research (CIHR; to DL and HV) is also gratefully acknowledged. HG is the recipient of a scholarship from the RMGA.

Ethical standards The study complies with the current laws of Canada.

1 Abney M, McPeek MS, Ober C: Estimation of variance components of quantitative traits in inbred populations. *Am J Hum Genet* 2000; **66**: 629–650.
2 Tulinius H: Multigenerational information: the example of the Icelandic Genealogy Database. In: Dillner J (ed.) *Methods in Biobanking*. New York: Humana Press, 2011; Vol 675, pp 221–229.
3 Khoury MJ, Cohen BH, Diamond EL, Chase GA, McKusick VA: Inbreeding and prereproductive mortality in the Old Order Amish. I. Genealogic epidemiology of inbreeding. *Am J Epidemiol* 1987; **125**: 453–461.
4 Huff CD, Witherspoon DJ, Simonson TS *et al*: Maximum-likelihood estimation of recent shared ancestry (ERSA). *Genome Res* 2011; **21**: 768–774.
5 Powell JE, Visscher PM, Goddard ME: Reconciling the analysis of IBD and IBS in complex trait studies. *Nat Rev Genet* 2010; **11**: 800–805.
6 Browning SR, Browning BL: Identity by descent between distant relatives: detection and applications. *Annu Rev Genet* 2012; **46**: 617–633.
7 Gusev A, Lowe JK, Stoffel M *et al*: Whole population, genome-wide mapping of hidden relatedness. *Genome Res* 2009; **19**: 318–326.
8 Browning BL, Browning SR: A fast, powerful method for detecting identity by descent. *Am J Hum Genet* 2011; **88**: 173–182.
9 Han L, Abney M: Identity by descent estimation with dense genome-wide genotype data. *Genet Epidemiol* 2011; **35**: 557–567.
10 Albrechtsen A, Sand Korneliussen T, Moltke I, van Overseem Hansen T, Nielsen FC, Nielsen R: Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. *Genet Epidemiol* 2009; **33**: 266–274.
11 Purcell S, Neale B, Todd-Brown K *et al*: PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; **81**: 559–575.
12 Palin K, Campbell H, Wright AF, Wilson JF, Durbin R: Identity-by-descent-based phasing and imputation in founder populations using graphical models. *Genet Epidemiol* 2011; **35**: 853–860.
13 Genovese G, Leibon G, Pollak M, Rockmore D: Improved IBD detection using incomplete haplotype information. *BMC Genet* 2010; **11**: 58.
14 Kong A, Masson G, Frigge ML *et al*: Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat Genet* 2008; **40**: 1068–1075.
15 Charbonneau H, Desjardins B, Légaré J, Denis H: The population of the St-Lawrence Valley, 1608-1760; in: Haines MR, Steckel RH (eds) *A population history of North America*. New York: Cambridge University Press, 2000; pp 99–142.
16 Bergeron J, Vézina H, Houde L, Tremblay M: La contribution des Acadiens au peuplement des régions du Québec. *Cahiers québécois de démographie* 2008; **37**: 181–204.
17 Bouchard G, De Braekeleer M: Homogénéité ou diversité? L'histoire de la population du Québec revue à travers ses gènes. *Social History/Histoire Sociale* 1990; **23**: 325–361.
18 Roy-Gagnon M-H, Moreau C, Bherer C *et al*: Genomic and genealogical investigation of the French Canadian founder population structure. *Hum Genet* 2011; **129**: 521–531.
19 Bherer C, Labuda D, Roy-Gagnon M-H, Houde L, Tremblay M, Vézina H: Admixed ancestry and stratification of Quebec regional populations. *Am J Phys Anthropol* 2011; **144**: 432–441.
20 Pouyez C, Lavoie Y, Bouchard G: *Les Saguenayens: introduction à l'histoire des populations du Saguenay, XVIe-XXe siècles*. Sillery, Québec: Presses de l'Université du Québec 1983.
21 Frenette P: *Histoire de la Côte-Nord. Sainte-Foy, Québec: Institut québécois de recherche sur la culture* 1996.
22 Desjardins M, Bélanger J: *Histoire de la Gaspésie. [Sainte-Foy, Québec]: Institut québécois de recherche sur la culture* 1999.
23 Heyer E, Tremblay M, Desjardins B: Seventeenth-century European origins of hereditary diseases in the Saguenay population (Quebec, Canada). *Hum Biol* 1997; **69**: 209–225.
24 Statistics Canada. 2011 Census of Population. http://www.statcan.gc.ca, 2012
25 BALSAC: Rapport annuel 2011–2012. http://balsac.uqac.ca, 2012

26 Desjardins B: Le Registre de la population du Québec ancien. *Annales de démographie historique* 1998; **2**: 215–226.

27 The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007; **449**: 851–861.

28 Pemberton TJ, Wang C, Li JZ, Rosenberg NA: Inference of unexpected genetic relatedness among individuals in HapMap Phase III. *Am J Hum Genet* 2010; **87**: 457–464.

29 Wigginton JE, Cutler DJ, Abecasis GR: A note on exact tests of Hardy–Weinberg equilibrium. *Am J Hum Genet* 2005; **76**: 887–893.

30 Karigl G: A recursive algorithm for the calculation of identity coefficients. *Ann Hum Genet* 1981; **45**: 299–305.

31 Lee W-J, Pollin T, O'Connell J, Agarwala R, Schaffer A: PedHunter 2.0 and its usage to characterize the founder structure of the old order amish of lancaster county. *BMC Med Genet* 2010; **11**: 68.

32 Browning SR, Browning BL: Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* 2007; **81**: 1084–1097.

33 Delaneau O, Marchini J, Zagury J-F: A linear complexity phasing method for thousands of genomes. *Nat Meth* 2012; **9**: 179–181.

34 Brown MD, Glazner CG, Zheng C, Thompson EA: Inferring coancestry in population samples in the presence of linkage disequilibrium. *Genetics* 2012; **190**: 1447–1460.

35 Browning SR, Browning BL: Identity by descent between distant relatives: detection and applications. *Annu Rev Genet* 2012; **46**: 615–631.

36 Browning SR, Browning BL: High-resolution detection of identity by descent in unrelated individuals. *Am J Hum Genet* 2010; **86**: 526–539.

37 Markus B, Birk OS, Geiger D: Integration of SNP genotyping confidence scores in IBD inference. *Bioinformatics* 2011; **27**: 2880–2887.

38 Stevens EL, Heckenberg G, Roberson EDO, Baugher JD, Downey TJ, Pevsner J: Inference of relationships in population data using identity-by-descent and identity-by-state. *PLoS Genet* 2011; **7**: e1002287.

39 Moltke I, Albrechtsen A, TvO Hansen, Nielsen FC, Nielsen R: A method for detecting IBD regions simultaneously in multiple individuals—with applications to disease genetics. *Genome Res* 2011; **21**: 1168–1180.

40 Gusev A, Palamara PF, Aponte G *et al*: The architecture of long-range haplotypes shared within and across populations. *Mol Biol Evol* 2012; **29**: 473–486.

41 Palamara PF, Lencz T, Darvasi A, Pe'er I: Length distributions of identity by descent reveal fine-scale demographic history. *Am J Hum Genet* 2012; **91**: 809–822.

42 Albrechtsen A, Moltke I, Nielsen R: Natural selection and the distribution of identity-by-descent in the human genome. *Genetics* 2010; **186**: 295–308.

43 Price AL, Helgason A, Thorleifsson G, McCarroll SA, Kong A, Stefansson K: Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. *PLoS Genet* 2011; **7**: e1001317.

44 Henn BM, Hon L, Macpherson JM *et al*: Cryptic distant relatives are common in both isolated and cosmopolitan genetic samples. *PLoS One* 2012; **7**: e34267.

45 Thornton T, Tang H, Hoffmann Thomas J, Ochs-Balcom Heather M, Caan Bette J, Risch N: Estimating kinship in admixed populations. *Am J Hum Genet* 2012; **91**: 122–138.

46 Stevens EL, Heckenberg G, Baugher JD, Roberson EDO, Downey TJ, Pevsner J: Consanguinity in Centre d'Etude du Polymorphisme Humain (CEPH) pedigrees. *Eur J Hum Genet* 2012; **20**: 657–667.

47 Browning SR, Thompson EA: Detecting rare variant associations by identity-by-descent mapping in case-control studies. *Genetics* 2012; **190**: 1521–1531.

48 Gusev A, Kenny EE, Lowe JK *et al*: DASH: a method for identical-by-descent haplotype mapping uncovers association with recent variation. *Am J Hum Genet* 2011; **88**: 706–717.

49 Newman DL, Abney M, McPeek MS, Ober C, Cox NJ: The importance of genealogy in determining genetic associations with complex traits. *Am J Hum Genet* 2001; **69**: 1146–1148.

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/3.0/

Supplementary Information accompanies this paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)