Hum Genet
DOI 10.1007/s00439-016-1702-6

ORIGINAL INVESTIGATION

Whole-genome sequencing in French Canadians from Quebec

Cécile Low-Kam[1,2] · David Rhainds[1] · Ken Sin Lo[1] · Sylvie Provost[1] · Ian Mongrain[1] · Anick Dubois[1] · Sylvie Perreault[3] · John F. Robinson[4] · Robert A. Hegele[4,5] · Marie-Pierre Dubé[1,2] · Jean-Claude Tardif[1,2] · Guillaume Lettre[1,2]

Received: 12 April 2016 / Accepted: 21 June 2016

Abstract Genome-wide association studies (GWAS) have had a tremendous success in the identification of common DNA sequence variants associated with complex human diseases and traits. However, because of their design, GWAS are largely inappropriate to characterize the role of rare and low-frequency DNA variants on human phenotypic variation. Rarer genetic variation is geographically more restricted, supporting the need for local whole-genome sequencing (WGS) efforts to study these variants in specific populations. Here, we present the first large-scale low-pass WGS of the French-Canadian population. Specifically, we sequenced at ~5.6× coverage the whole genome of 1970 French Canadians recruited by the Montreal Heart Institute Biobank and identified 29 million bi-allelic variants (51 % novel), including 19 million variants with a minor allele frequency (MAF) <0.5 %. Genotypes from the WGS data are highly concordant with genotypes obtained by exome array on the same individuals (99.8 %), even when restricting this analysis to rare variants (MAF <0.5, 99.9 %) or heterozygous sites (98.9 %). To further validate our data set, we showed that we can effectively use it to replicate several genetic associations with myocardial infarction risk and blood lipid levels. Furthermore, we analyze the utility of our WGS data set to generate a French-Canadian-specific imputation reference panel and to infer population structure in the Province of Quebec. Our results illustrate the value of low-pass WGS to study the genetics of human diseases in the founder French-Canadian population.

Introduction

The recent revolution in human complex trait genetics comes from the development of genome-wide genotyping platforms. These, when combined with genotype imputation using reference haplotypes from the HapMap or 1000 Genomes Project, have led to the identification of thousands of robust genetic associations between complex human phenotypes and single nucleotide polymorphisms (SNPs) (1000 Genomes Project Consortium 2010; 1000 Genomes Project Consortium 2012; Altshuler et al. 2005, 2010; Frazer et al. 2007; Welter et al. 2014). Under the genome-wide association study (GWAS) framework, genotype imputation is key: it improves power to find markers that are not directly genotyped and facilitate meta-analyses between studies that are genotyped on different arrays. Notwithstanding its success, the GWAS approach has suboptimal ability to capture low frequency and rare DNA sequence variants because of limited or absent linkage

J.-C. Tardif and G. Lettre contributed equally.

Electronic supplementary material The online version of this article (doi:10.1007/s00439-016-1702-6) contains supplementary material, which is available to authorized users.

✉ Jean-Claude Tardif
Jean-Claude.Tardif@icm-mhi.org

✉ Guillaume Lettre
guillaume.lettre@umontreal.ca

[1] Montreal Heart Institute, 5000 Bélanger Street, Montréal, QC H1T 1C8, Canada

[2] Faculté de Médecine, Université de Montréal, 2900 Boul. Édouard-Montpetit, Montréal, QC H3T 1J4, Canada

[3] Faculté de Pharmacie, Université de Montréal, 2940 chemin de la Polytechnique, Montréal, QC H3T 1J4, Canada

[4] Robarts Research Institute, Schulich School of Medicine and Dentistry, Western University, London, ON, Canada

[5] Department of Biochemistry, Schulich School of Medicine and Dentistry, Western University, London, ON, Canada

Published online: 04 July 2016

disequilibrium (LD) between markers. The genetic community is currently shifting part of its interest from common to rare genetic variants to study the aetiology of complex human diseases. To this end, it is seeking strategies to more comprehensively characterize rare genetic variation in large and diverse human populations. In particular, because rare genetic variants tend to be population specific, such comprehensive surveys need to be done by direct DNA sequencing in populations of interest.

Although feasible, high-coverage (30×) whole-genome sequencing (WGS) remains expensive, especially if we consider the large sample size required for rare-variant genetic association studies. One alternative strategy is to sequence at low coverage (4–6×) a large number of participants, and then uses imputation methodology to recover missing genotypes within the sequenced individuals (Li et al. 2011). This method can also yield high-quality haplotypes that are useful to impute individuals from the same population with genome-wide genotype data available. The 1000 Genomes Project implemented successfully this approach in individuals from several populations from across the globe (1000 Genomes Project Consortium 2010; 1000 Genomes Project Consortium 2012). The SardiNIA and Genome of the Netherlands consortia also used low-to-intermediate coverage WGS to find common and rare population-specific variants (Danjou et al. 2015; Genome of the Netherlands 2014; Sidore et al. 2015; Zoledziewska et al. 2015). Their analyses confirmed that low-pass WGS data sets are appropriate for genetic association studies, to create imputation panels, or to ask population genetics questions that are specific to these populations.

The population of the Province of Quebec in Canada now includes 8.1 million individuals, ~78 % of whom are French speaking. This French-Canadian population largely originates from the immigration of 8500 mostly French settlers starting in 1608 and ending with the British Conquest of 1759. Following these events, the French-Canadian population of Quebec expanded rapidly, and an intense population growth was paralleled with the colonization of new geographical regions, such as Saguenay-Lac-St-Jean in the North-East and Gaspesia in the East (Supplementary Fig. 1). This rapid expansion and migration gave rise to regional populations with genetic features that can be explained by founder effects (Bherer et al. 2011; Gauvin et al. 2014; Moreau et al. 2011a; Roy-Gagnon et al. 2011). Geneticists have taken advantage of this genetic homogeneity in the French-Canadian founder population of Quebec to map genes involved in rare and common human diseases (Chami et al. 2014; Chetaille et al. 2014; Laprise 2014; Scriver 2001). To date, however, no WGS project has explored the pattern of rare genetic variation in French Canadians from Quebec, as well as its contribution to disease aetiology.

In this study, we sequenced the whole genome of 1970 French Canadians at 5.6× coverage. We identified ~29 million high-quality bi-allelic variants, of which 51 % were not previously reported in public databases (dbSNP 147). When stratifying by minor allele frequency (MAF), we found 6 million common (MAF > 5 %), 4 million low-frequency (0.5 % < MAF ≤ 5 %), and 19 million rare (MAF ≤ 0.5 %) markers. The majority of the bi-allelic variants are intergenic, while 44 % are intronic and 3 % are exonic. The WGS data show 99.8 % concordance when compared to ExomeChip genotype data available in the same participants, suggesting very high quality. We further demonstrate the utility of this WGS data set for genetic association studies, to build a French-Canadian haplotype reference panel for imputation, and to begin to explore the genetic history of this specific population.

Results

Genetic variation in French Canadians

We sequenced at 5.6× coverage 1970 French Canadians recruited by the Montreal Heart Institute (MHI) Biobank (Supplementary Fig. 2). Demographic and clinical information for the participants is available in Supplementary Table 1. After quality-control steps, we identified 29 M single nucleotide variants (SNVs) and small insertion deletions (indels). Nearly, 51 % of these genetic variants are not present in public databases (dbSNP 147); this number is consistent with results from a low-pass WGS project in European Americans (Morrison et al. 2013). We annotated all bi-allelic SNVs and indels identified (Table 1). As expected, likely detrimental mutations (nonsense, splice site, frameshift indel) are enriched at the rarer end of the allele frequency spectrum, consistent with purifying selection. For instance, whereas 17 % of synonymous SNVs are common, 5 % of nonsense and 11 % of frameshift indel variants have a frequency >5 % (Fig. 1).

The transition-to-transversion ratio, a metric used to assess the quality of DNA sequencing data, is 2.13, very close to the theoretical 2.1 ratio expected for WGS (DePristo et al. 2011). To further evaluate the quality of the WGS data, we calculated concordance with genotypes from an independent exome array experiment available for 1967 individuals. Across 83,273 bi-allelic SNPs shared between the WGS and ExomeChip data sets, we calculated an overall concordance rate of 99.8 %. When restricting this analysis to heterozygous genotype calls, we found concordance rate of 99.3, 95.2, and 85.3 % for common, low-frequency, and rare SNVs, respectively (Supplementary Tables 2–5). The rediscovery rate (percentage of bi-allelic markers from the ExomeChip data set detected by WGS)

Table 1 Number of bi-allelic variants per annotation type and minor allele frequency (MAF) class: common (MAF > 5 %), low frequency (0.5 % < MAF ≤ 5 %), and rare (MAF ≤ 0.5 %)

Annotation class	Definition (EPACTS)	All variants (%)	Common (% novel)	Low-frequency (% novel)	Rare (% novel)
Exon	Codon gain, codon loss, exon, frameshift, nonsynonymous, start gain, start loss, stop gain, stop loss, synonymous, 5′UTR, 3′UTR	795,781 (2.70 %)	126,678 (0.2)	95,707 (5.7)	573,396 (66.2)
Frameshift indel	Frameshift	3633 (0.01 %)	390 (25.1)	572 (68.4)	2671 (94.0)
Intergenic	Intergenic, downstream, upstream	14,436,195 (49.01 %)	3,161,456 (3.8)	2,005,936 (13.6)	9,268,803 ((73.6)
Intron	Intron, essential splice site, normal splice site	12,943,152 (43.94 %)	2,512,344 (0.08)	1,634,909 (4.1)	8,795,899 (73.2)
Insertion Deletion	Insertion, deletion	1,278,660 (4.34 %)	379,998 (31.4)	335,883 (65.7)	562,779 (88.0)
Missense	Nonsynonymous, stop loss, start gain	165,934 (0.6 %)	16,395 (0.03)	16,549 (0.07)	132,990 (53.9)
Nonsense	Stop gain, start loss	4364 (0.01 %)	204 (0)	273 (0.4)	3887 (66.0)
Regulatory region	Downstream (50 bp), upstream (50 bp), 5′UTR, 3′UTR	354,953 (1.21 %)	60,511 (0.09)	43,773 (5.6)	250,669 (74.2)
Splice site	Essential splice site	3136 (0.01 %)	392 (0)	279 (2.2)	2465 (70.5)
Synonymous	Synonymous	106,533 (0.36 %)	18,197 (0.07)	13,318 (1.0)	75,018 (50.5)

We used functional annotations from the software EPACTS to annotate variants

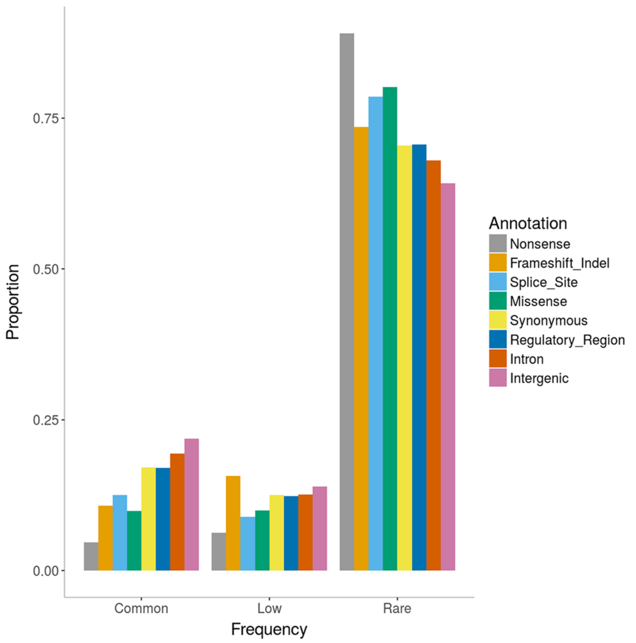

Fig. 1 Proportion of annotated variants in three minor allele frequency (MAF) classes: common (MAF > 5 %), low frequency (0.5 % < MAF ≤ 5 %), and rare (MAF ≤ 0.5 %). The definition of each functional annotation is provided in Table 1. We note an enrichment of missense, nonsense, and splice site variants among rare variants when compared to synonymous DNA sequence changes

was 98, 97, and 76 % for common, low-frequency and rare SNVs, respectively.

Replication of GWAS findings

This WGS experiment is part of a larger research program that focuses on the genetics of coronary artery disease in the French-Canadian population. For this project, we explicitly selected 984 myocardial infarction (MI) patients and 980 MI-free controls; additional clinical information on MI risk factors is also available for these participants (Supplementary Table 1). Although our sample size (and thus statistical power) is modest when compared to recent meta-analyses of GWAS results, we tested the association between known SNPs and MI status as well as high-density lipoprotein (HDL) cholesterol, low-density lipoprotein (LDL) cholesterol, total cholesterol (TC), and triglyceride (TG) levels to assess the quality of our data set to perform association testing.

For MI, 35 of the 44 known variants available in the WGS data set have an effect in the right direction (binomial $P = 1 \times 10^{-4}$) and five are nominally significant (Table 2 and Supplementary Table 6) (CARDIoGRAMplusC4D Consortium 2013; Coronary Artery Disease (C4D) Genetics Consortium 2011; Myocardial Infarction Genetics Consortium 2009; Schunkert et al 2011). This includes an intronic marker

Table 2 Replication of a subset of markers associated with myocardial infarction, high-density lipoprotein (HDL) cholesterol, low-density lipoprotein (LDL) cholesterol, total cholesterol, and triglyceride levels

SNP	Chr	Position (hg19)	Effect allele	Other allele	Odds-ratio/beta (odds-ratio/beta from the literature)	*P* value	Locus
Myocardial infarction							
rs6725887	2	203745885	C	T	1.245 (1.278)	2.40×10^{-2}	*WDR12*
rs12526453	6	12927544	G	C	0.734 (0.788)	1.64×10^{-5}	*PHACTR1*
rs4977574	9	22098574	G	A	1.222 (1.087)	3.42×10^{-3}	*CDKN2A/CDKN2B*
rs974819	11	103660567	C	T	0.789 (0.860)	8.67×10^{-4}	*PDGFD*
rs46522	17	46988597	T	C	1.214 (1.238)	4.10×10^{-3}	*UBE2Z*
High-density lipoprotein (HDL)-cholesterol							
rs2290547	3	47061183	A	G	−0.074 (−0.030)	4.36×10^{-2}	*SETD2*
rs13326165	3	52532118	G	A	−0.074 (−0.029)	3.90×10^{-2}	*STAB 1*
rs9987289	8	9183358	G	A	0.140 (0.082)	1.43×10^{-2}	*PPP1R3B*
rs12678919	8	19844222	G	A	0.167 (0.155)	1.31×10^{-3}	*LPL*
rs1883025	9	107664301	T	C	−0.104 (−0.070)	2.29×10^{-3}	*ABCA1*
rs174546	11	61569830	T	C	−0.065 (−0.039)	4.28×10^{-2}	*FADS1-2-3*
rs7134594	12	110000193	T	C	0.062 (0.035)	3.69×10^{-2}	*MVK*
rs1532085	15	58683366	G	A	−0.122 (−0.107)	7.13×10^{-5}	*LIPC*
rs3764261	16	56993324	A	C	0.170 (0.241)	7.06×10^{-8}	*CETP*
rs11869286	17	37813856	C	G	0.082 (0.032)	7.07×10^{-3}	*STARD3*
rs7241918	18	47160953	T	G	0.098 (0.090)	1.59×10^{-2}	*LIPG*
Low-density lipoprotein (LDL)-cholesterol							
rs2131925	1	63025942	T	G	0.091 (0.049)	1.65×10^{-3}	*ANGPTL3*
rs629301	1	109818306	T	G	0.093 (0.167)	5.77×10^{-3}	*SORT1*
rs2072183	7	44579180	C	G	0.095 (0.039)	4.25×10^{-3}	*NPC1L1*
rs10102164	8	55421614	A	G	0.073 (0.032)	3.48×10^{-2}	*SOX17*
rs3780181	9	2640759	G	A	−0.142 (−0.044)	6.39×10^{-3}	*VLDLR*
rs964184	11	116648917	C	G	−0.170 (−0.086)	8.06×10^{-6}	*APOA1*
rs2000999	16	72108093	A	G	0.073 (0.065)	2.93×10^{-2}	*HPR*
rs6511720	19	11202306	T	G	−0.113 (−0.221)	5.99×10^{-3}	*LDLR*
rs4420638	19	45422946	G	A	0.167 (0.225)	7.47×10^{-6}	*APOE*
Total cholesterol (TC)							
rs1077514	1	23766233	T	C	0.074 (0.030)	4.59×10^{-2}	*ASAP3*
rs2131925	1	63025942	T	G	0.084 (0.075)	3.25×10^{-3}	*ANGPTL3*
rs629301	1	109818306	T	G	0.097 (0.134)	3.38×10^{-3}	*SORT1*
rs1260326	2	27730940	C	T	−0.084 (−0.051)	1.44×10^{-3}	*GCKR*
rs2072183	7	44579180	C	G	0.077 (0.036)	1.78×10^{-2}	*NPC1L1*
rs3780181	9	2640759	G	A	−0.165 (−0.044)	1.24×10^{-3}	*VLDLR*
rs964184	11	116648917	C	G	−0.187 (−0.121)	5.27×10^{-7}	*APOA1*
rs1532085	15	58683366	G	A	−0.065 (−0.054)	1.63×10^{-2}	*LIPC*
rs3764261	16	56993324	A	C	0.096 (0.050)	5.72×10^{-4}	*CETP*
rs2000999	16	72108093	A	G	0.085 (0.062)	8.93×10^{-3}	*HPR*
rs4420638	19	45422946	G	A	0.131 (0.197)	3.47×10^{-4}	*APOE*
Triglycerides							
rs2131925	1	63025942	T	G	0.071 (0.066)	3.68×10^{-2}	*ANGPTL3*
rs1260326	2	27730940	C	T	−0.104 (−0.115)	1.05×10^{-3}	*GCKR*
rs6882076	5	156390297	C	T	0.070 (0.029)	3.36×10^{-2}	*TIMD4*
rs1495741	8	18272881	A	G	−0.110 (−0.040)	3.01×10^{-3}	*NAT2*
rs12678919	8	19844222	G	A	−0.164 (−0.170)	2.87×10^{-3}	*LPL*
rs2954029	8	126490972	T	A	−0.090 (−0.076)	3.73×10^{-3}	*TRIB1*

Table 2 continued

SNP	Chr	Position (hg19)	Effect allele	Other allele	Odds-ratio/beta (odds-ratio/beta from the literature)	*P* value	Locus
rs174546	11	61569830	T	C	0.083 (0.045)	1.37×10^{-2}	*FADS1-2-3*
rs964184	11	116648917	C	G	−0.304 (−0.234)	8.29×10^{-12}	*APOA1*

We present only nominally significant results ($P \leq 0.05$) with phenotypic effect in the expected direction based on the literature. All replication results are available in Supplementary Tables 6–10. For myocardial infarction, we provide the effect size as odds ratio. For blood lipids levels, effect sizes are in the standard deviation units. We also report in parentheses the effect sizes and odds ratios from the literature (Beaudoin et al. 2015; Global Lipids Genetics et al. 2013)

at the *PHACTR1* locus, which we had previously shown to be associated with MI in the MHI Biobank using a different genomic technology for DNA genotyping (Beaudoin et al. 2015). For HDL-C, LDL-C, TC, and TG levels, respectively, 51 of the 69 (binomial $P = 9 \times 10^{-5}$), 42 of the 57 (binomial $P = 5 \times 10^{-4}$), 56 of the 73 (binomial $P = 5 \times 10^{-6}$), and 26 of the 40 (binomial $P = 0.08$) variants tested have effect alleles that modulate their respective phenotype in the right direction based on the literature (Supplementary Tables 7–10) (Global Lipids Genetics et al. 2013). We found eleven, nine, eleven, and eight genetic associations with HDL-C, LDL-C, TC, and TG levels that are in the correct direction and nominally significant at $P < 0.05$ (Table 2). Overall, these results suggest that our WGS data set represents a robust resource to investigate a novel genetic association with coronary artery disease or its risk factor in the French-Canadian population.

Imputation

Motivated by recent reports that showed improvements in imputation quality when the target samples and reference haplotypes are from individuals of the same population (Genome of the Netherlands 2014; Pistis et al. 2015), we tested the utility of our WGS-derived haplotypes to impute genotypes in French Canadians. For this analysis, we imputed genotypes in 731 independent French Canadians previously genotyped on the Illumina Omni2.5 M array. To this end, we used three reference panels: (1) 5008 haplotypes from all the populations from the 1000 Genomes Project phase 3, (2) 3940 French-Canadian haplotypes from our low-pass WGS experiment, and (3) the combined set of 1000 Genomes Project and French-Canadian haplotypes. We limited our analyses to 892,893 high-quality markers (HWE *P* value in WGS controls $>1 \times 10^{-5}$) on chromosome 1 that are shared between the 1000 Genomes Project and our WGS data set. To estimate imputation quality, we had each known genotype in turn and computed the squared correlation (r^2) between the imputed dosages and the masked true genotypes. Overall, the three reference panels supported excellent imputation quality, with mean imputation r^2 of 0.926, 0.942, and 0.944 for the 1000 Genomes Project, French-Canadian, and combined haplotypes, respectively (Fig. 2). We noted a slightly better imputation quality for rare variants using the French Canadian rather than the 1000 Genomes Project haplotypes, consistent with the previous report (mean r^2 of 0.76 vs. 0.71) (Fig. 2). The estimated imputation accuracy, calculated on both known and imputed genotypes, behaves in a similar way (Supplementary Fig. 3). Practically, investigators often filter at an imputation quality score ≥0.3 to select imputed markers for association testing. Using this threshold, 59 % of the rare variants imputed with the French-Canadian haplotypes would be selected instead of 55 % of the rare variants imputed with the 1000 Genomes Project haplotypes only.

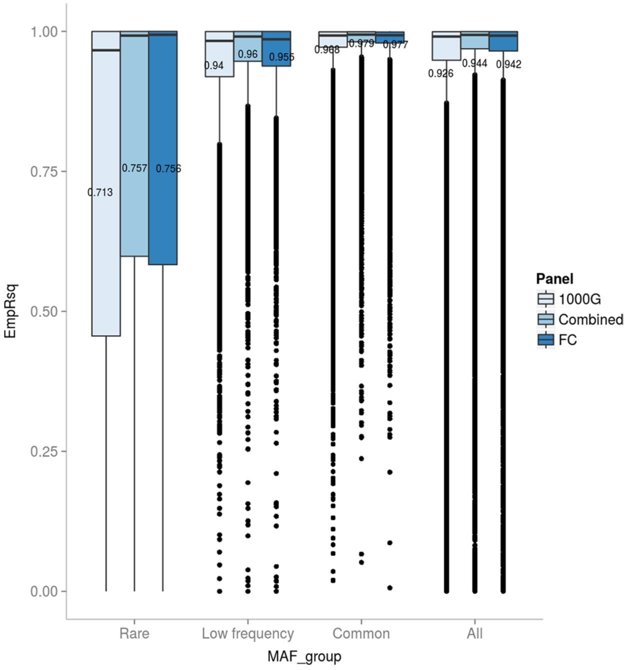

Fig. 2 Boxplots and mean of the squared correlation between masked genotypes and imputed dosages. For each of three haplotype reference panels, we report the imputation quality for ~10 K rare (MAF ≤ 0.5 %), ~20 K low-frequency (0.5 % < MAF ≤ 5 %), ~46 K common (MAF > 5 %), and all ~76 K shared polymorphic variants on chromosome 1. The *number on each box* corresponds to the mean EmpRsq quality score. 1000G, 5008 haplotypes from phase 3 of the 1000 Genomes Project; FC, 3940 haplotypes from this whole-genome DNA re-sequencing project in French Canadians; and combined, combination of the 1000G and FC haplotypes

The recently developed Haplotype Reference Consortium (HRC, http://www.haplotype-reference-consortium.org/) has combined 20 cohorts to obtain over 64,000 haplotypes that cover ~39 M variants. We compared the imputation performance of our French-Canadian panel to haplotypes from HRC, restricting this comparison to 808,746 shared variants on chromosome 1. Overall, the HRC panel performs slightly better than the French-Canadian panel (mean r^2 of 0.96 vs. 0.94), and that is also true for rare variants (mean r^2 of 0.79 vs. 0.76) (Supplementary Figs. 4 and 5). Thus, for variants present in HRC, the large number of haplotypes outweighs the population specificity of the French-Canadian panel. However, this result does not question the relevance of our French-Canadian WGS, since nearly 50 % of the variants found in our project are absent from public databases and could not be imputed with HRC.

Population structure in Québec

Although all MHI Biobank participants were recruited in the cosmopolitan Montreal area, we have information on the geographical origin of their four grandparents. Using this data, we could assign each participant to the 17 administrative regions in Quebec (Supplementary Fig. 1). For the subsequent analyses, we focus our genetic analyses to four regions: Montreal, Quebec City, and the Saguenay-Lac-St-Jean regions that are characterized by a strong founder effect and the Gaspesia region, which harbours higher genetic heterogeneity, caused by an influx of populations, namely, Acadians and Loyalists (Moreau et al. 2011b). For each region, the participants receive a score (0–4) that corresponds to the number of grandparents from this region. Thus, a participant with two grandparents from Montreal and two from Saguenay would get scores of 2 for these regions and scores of 0 for Quebec City and Gaspesia (Supplementary Table 11).

We used multidimensional scaling (MDS) on partially independent bi-allelic variants ($r^2 < 0.5$) with MAF > 1 % to project the data on the ten dimensions that preserve most of the genetic distance. Then, we determined how genetic variation captured by low-pass WGS reflects regional ancestry. The first two dimensions from MDS model are the regional structure of our population within Quebec. For instance, dimension C1 highlights individuals from Saguenay-Lac-St-Jean—and to a lesser extent participants from Quebec City—from the rest of the sequenced participants (Fig. 3). Thus, C1 appears to capture the serial migration events that led to the peopling of the Northern regions of Quebec in the 19th century, consistent with the previous observations (Moreau et al. 2011a; Roy-Gagnon et al. 2011). The second MDS dimension (C2) isolates individuals from Gaspesia (Fig. 3). Since a small number of the French Canadians in our data set ($N = 28$, Supplementary Table 11) do not originate from the province of Quebec, we derive similar scores for the other Canadian provinces. Interestingly, C2 also captures ancestry from Canada's Maritime provinces (Nova Scotia, New Brunswick, and Prince Edward Island) that are historically and geographically linked to Gaspesia (Supplementary Fig. 6). Supplementary Figure 7 shows the same projection, but restricted to 199 individuals, whose four grandparents are from the same region (Montreal, Quebec City, Saguenay, and Gaspesia). Dimensions C1 and C2 capture the same patterns of population structure. Finally, we combined our French-Canadian data, restricted to variants with MAF > 5 %, with data from the 1000 Genomes Project to create a data set of 5 million shared genetic variants among 4474 individuals (2504 individuals from the 1000 Genomes Project). We then performed a new MDS analysis using only 429,800 partially independent variants ($r^2 < 0.5$). On C1 and C2, the French Canadians overlap mostly with individuals of northern European ancestry (CEU) (Supplementary Figs. 8 and 9).

Discussion

In this article, we present the first low-pass WGS effort to characterize the French-Canadian population. Our work builds on the previous studies that targeted specific subpopulations to discover population-specific genetic variation. The resulting high-quality WGS data set reflects the regional structure of the French-Canadian population and is highly concordant with genotype data available for the same samples but obtained using a different technology. Despite the relatively small sample size of our data set, we replicated genetic associations with myocardial infarction and lipid traits. Furthermore, the corresponding population-specific haplotypes outperformed the 1000 Genomes Project panel for the imputation of rare variants in French-Canadian samples. All these results illustrate the depth, quality, and future usefulness of this WGS resource to investigate the genetics of complex human diseases and other quantitative traits in this specific population.

Although the French-Canadian haplotypes were more efficient than the 1000 Genomes Project haplotypes to impute rare variants in French Canadians, the gain remains modest. This is particularly noticeable when comparing with the gain in imputation quality of rare variants observed in other low-pass WGS projects, such as the Genome of the Netherlands and SardiNIA (Genome of the Netherlands 2014; Pistis et al. 2015). One major difference between these efforts and our WGS project is the absence of related individuals among the French Canadians that were sequenced. Related individuals would simplify the phasing of the genotypes and improve the quality of the haplotypes. An alternative approach to improve the quality

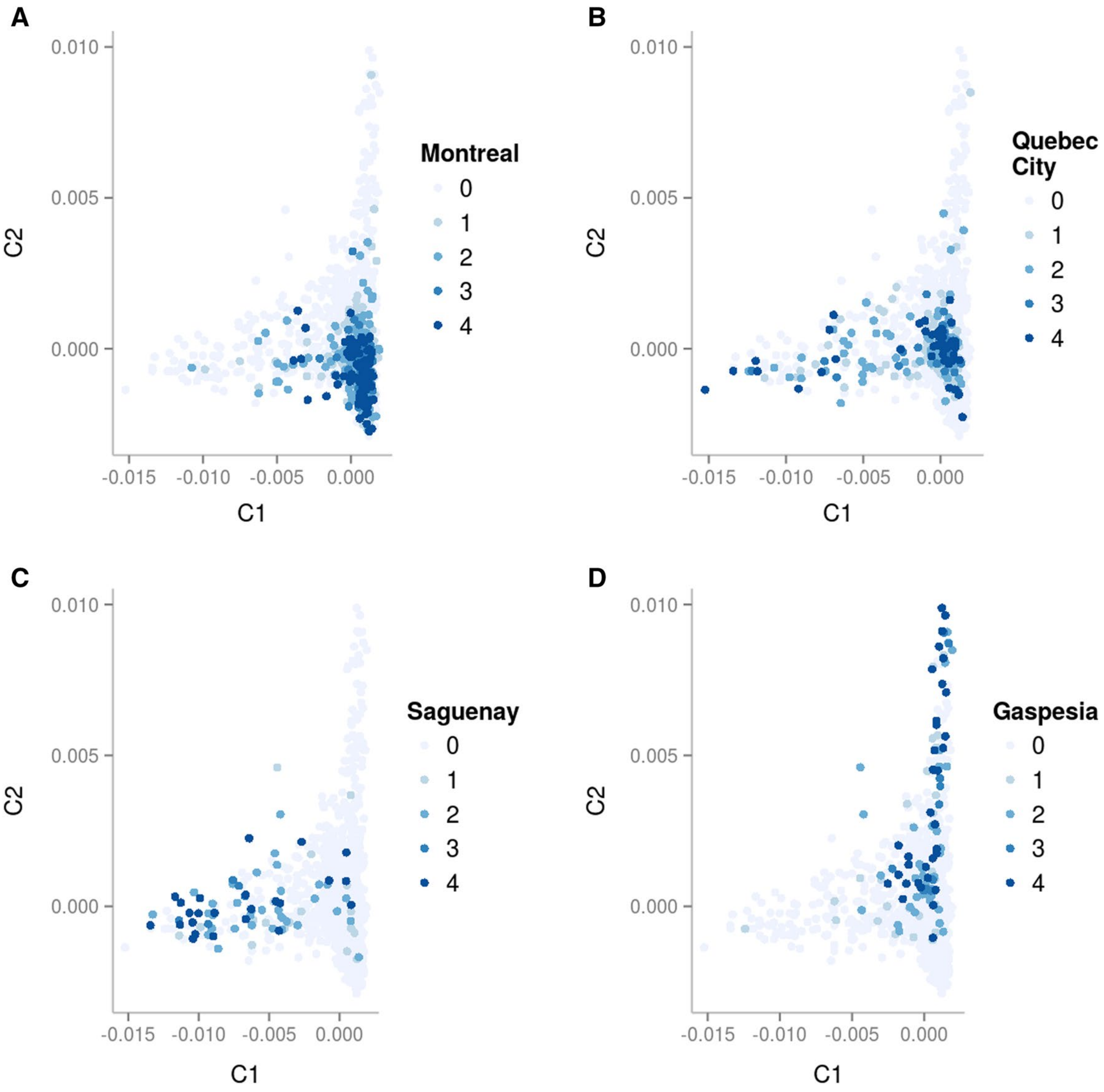

Fig. 3 Projection of 1970 French-Canadian individuals on dimensions C1 and C2 from multidimensional scaling (MDS). Individuals are color-coded according to the number of their grandparents from Montreal (**a**), Quebec City (**b**), Saguenay-Lac-St-Jean (**c**), and Gaspesia (**d**). C1 isolates individuals with ancestry in Saguenay-Lac-St-Jean and Quebec City, consistent with the historical migration wave in the Province of Quebec. C2 opposes individuals from Gaspesia to the rest of the population in Quebec

of our French-Canadian haplotypes would be to use independent dense genotyping data (e.g., GWAS-type array) as a so-called “scaffold” to constrain the search space for imputation (Delaneau et al. 2014).

The development of very large imputation panels, such as the 64,976 haplotypes panel from HRC, has called into question the value of WGS to identify genetic risk factors for complex human diseases. In particular, we compared the performance of HRC and haplotypes from our WGS project to impute rare variants in French Canadians, and noted slightly better imputation quality for the larger HRC panel (e.g., mean r^2 of 0.79 vs. 0.76 for rare variants). As expected, this suggests that for variants common to both data sets, the number of haplotypes is the main predictor of imputation quality. However, this result does not undermine the importance to carry out population-specific WGS projects, since 51 % of the variants identified in French Canadians are not currently available in public databases (and would not have been imputed using HRC alone). This is consistent with the discovery by low-pass WGS of population-specific variants of large effect size on human complex phenotypes (Danjou et al. 2015; Sidore et al. 2015; Zoledziewska et al. 2015). WGS data also provide the opportunity to explore other type of DNA sequence polymorphisms, such as structural and highly repetitive variants. As sequencing costs continue to drop, the sample size of WGS projects will increase, allowing powerful association analyses between the entire human genome and complex diseases.

Online methods

Sample collection

984 cases and 986 controls for myocardial infarction (MI) (Supplementary Table 1) were recruited from the Montreal Heart Institute Biobank (Beaudoin et al. 2012). 980 controls were selected to be free of percutaneous coronary intervention, coronary artery bypass surgery, transient ischemic attack or stroke, peripheral vascular disease, congestive heart failure, and angina. Cases and controls were matched on sex, hypertension, diabetes, and dyslipidemia. All samples have French-Canadian ancestry, and the origin (birthplace) of their grandparents is available. Clinical information (MI status, age, sex, statin use, nuclear magnetic resonance-based HDL cholesterol and triglyceride levels, biochemistry-measured LDL-cholesterol and total cholesterol) is available for all participants.

Whole-genome DNA sequencing

Samples were sequenced with Illumina HiSeq technology at the McGill and Genome Quebec Innovation Centre. The resulting reads were then aligned on the UCSC reference genome b37 and filtered using a bioinformatics pipeline combining the Burrows–Wheeler Aligner (BWA) (Li and Durbin 2009), SAMtools (http://samtools.sourceforge.net), Picard (http://picard.sourceforge.net), and the Genome Analysis Tool Kit (GATK) v1.8 (McKenna et al. 2010) (Supplementary Note). Variant calling was performed across samples using GATK's UnifiedGenotyper. We then applied GATK's Variant QualityScore Recalibration to filter the resulting vcf file (Supplementary Note). GATK's VariantEvaluation was used to compute statistics on the filtered variants. Sequences were phased, and sporadic missing genotypes were imputed with Beagle v4 (Browning and Browning 2007). For each sample, the algorithm took the genotype likelihoods as input and used the phased remaining sequences as a reference panel (Supplementary Note). We used PLINK 1.9 (Purcell et al. 2007) and vcftools 0.1.12 to control the quality of our sequence data (Supplementary Note). The resulting quality-controlled bi-allelic variants were annotated using build hg19 of the reference human genome and the Efficient and Parallelizable Association Container Toolbox (EPACTS, http://genome.sph.umich.edu/wiki/EPACTS).

Concordance with the ExomeChip

All samples have been genotyped on the Illumina ExomeChip. Concordance on individuals and positions shared in common between the whole-genome sequence data set and the ExomeChip genotype data was calculated with PLINK using the option diff. We then recoded both files in additive components with PLINK and built the contingency table with R (Supplementary Note).

Genetic association analyses

The association analyses were performed using the Wald test from the software rvtest (http://genome.sph.umich.edu/wiki/RvTests). The covariates for all phenotypes included age-squared, sex, statin usage, and the first ten dimensions from multidimensional scaling (MDS). The MI case–control status was also a covariate for the inverse-normal transformed lipids phenotypes. We accounted for experimental batch effects for LDL-C and TC phenotypes.

Genotype imputation in French Canadians

We evaluated the imputation performance of our French-Canadian reference panel on 751 French-Canadian samples genotyped on the Illumina Omni2.5 array. We excluded 20 samples already present in the reference panel based on an identity-by-descent analysis with independent SNPs at MAF > 1 %. The following analyses were performed on chromosome 1 only: the remaining 731 samples were filtered with PLINKseq and checkvcf, and phased with shapeit (Delaneau et al. 2012). The resulting haplotypes were imputed with minimac3 (http://genome.sph.umich.edu/wiki/Minimac3) using four reference panels: the Haplotype Reference Consortium (HRC), 1000 Genomes Project version 3 (1000G), our French-Canadian reference panel (FC), excluding variants out of Hardy–Weinberg equilibrium ($P < 1 \times 10^{-5}$), and a panel combining 1000G and FC haplotypes. Imputation quality was measured with the metrics EmpRsq and Rsq produced by minimac3.

Population structure analysis

Multidimensional scaling (MDS) was performed with PLINK on partially independent ($r^2 < 0.5$) variants with MAF > 1 %. The origin of the four grandparents of each patient is heterogeneous: it might be a city, an administrative region of Quebec, the global province of Quebec, or even other Canadian provinces or countries. To assign regional ancestry scores to individuals, we mapped their grandparents' origin to the 17 Quebec administrative regions based on electoral data available at http://www.electionsquebec.qc.ca/francais/provincial/carte-electorale/municipalites-et-entites-administratives-2011.php.

Acknowledgments We thank all participants and staff of the André and France Desmarais Montreal Heart Institute (MHI) Biobank.

Sequencing of the MHI Biobank samples was performed at the McGill University and Génome Québec Innovation Centre. This work is funded by Génome Québec and Génome Canada, the MHI Foundation, and the Canada Research Chair program. RAH is funded by a Foundation Award from the Canadian Institutes of Health Research and by the Jacob J. Wolfe Distinguished Medical Research Chair and Edith Schulich Vinet Medical Research Chair.

References

1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467:1061–1073

1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491:56–65

Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, Donnelly P (2005) A haplotype map of the human genome Nature 437:1299–1320

Altshuler DM et al (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467:52–58

Beaudoin M et al (2012) Pooled DNA resequencing of 68 myocardial infarction candidate genes in French Canadians. Circ Cardiovasc 5:547–554

Beaudoin M et al (2015) Myocardial infarction-associated SNP at 6p24 interferes with MEF2 binding and associates with PHACTR1 expression levels in human coronary arteries. Arterioscler Thromb Vasc Biol 35:1472–1479

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

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

CARDIoGRAMplusC4D Consortium (2013) Large-scale association analysis identifies new risk loci for coronary artery disease. Nat Genet 45(1):25–33

Chami N et al (2014) Nonsense mutations in BAG3 are associated with early-onset dilated cardiomyopathy in French Canadians. Can J Cardiol 30:1655–1661

Chetaille P et al (2014) Mutations in SGOL1 cause a novel cohesinopathy affecting heart and gut rhythm. Nat Genet 46:1245–1249

Coronary Artery Disease (C4D) Genetics Consortium (2011) A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease. Nat Genet 43(4):339–344

Danjou F et al (2015) Genome-wide association analyses based on whole-genome sequencing in Sardinia provide insights into regulation of hemoglobin levels. Nat Genet 47:1264–1271

Delaneau O, Marchini J, Zagury JF (2012) A linear complexity phasing method for thousands of genomes. Nat Methods 9:179–181

Delaneau O, Marchini J, Genomes Project C, Genomes Project C (2014) Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. Nat Commun 5:3934

DePristo MA et al (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genet 43:491–498

Frazer KA et al (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861

Gauvin H, Moreau C, Lefebvre JF, Laprise C, Vezina H, Labuda D, Roy-Gagnon MH (2014) Genome-wide patterns of identity-by-descent sharing in the French Canadian founder population. Eur J Hum Genet 22:814–821

Genome of the Netherlands (2014) Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat Genet 46:818–825

Global Lipids Genetics C et al (2013) Discovery and refinement of loci associated with lipid levels. Nat Genet 45:1274–1283

Laprise C (2014) The Saguenay-Lac-Saint-Jean asthma familial collection: the genetics of asthma in a young founder population. Genes Immun 15:247–255

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760

Li Y, Sidore C, Kang HM, Boehnke M, Abecasis GR (2011) Low-coverage sequencing: implications for design of complex trait association studies. Genome Res 21:940–951

McKenna A et al (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297–1303

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

Moreau C, Vezina H, Jomphe M, Lavoie EM, Roy-Gagnon MH, Labuda D (2011b) When genetics and genealogies tell different stories-maternal lineages in Gaspesia. Ann Hum Genet 75:247–254

Morrison AC et al (2013) Whole-genome sequence-based analysis of high-density lipoprotein cholesterol. Nat Genet 45:899–901

Myocardial Infarction Genetics Consortium (2009) Genome-wide association of early-onset myocardial infarction with common single nucleotide polymorphisms, common copy number. Nat Genet 41(3):334–341

Pistis G et al (2015) Rare variant genotype imputation with thousands of study-specific whole-genome sequences: implications for cost-effective study designs. Eur J Hum Genet 23:975–983

Purcell S et al (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559–575

Roy-Gagnon MH et al (2011) Genomic and genealogical investigation of the French Canadian founder population structure. Hum Genet 129:521–531

Schunkert H et al (2011) Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. Nat Genet 43(4):333–338

Scriver CR (2001) Human genetics: lessons from Quebec populations. Annu Rev Genomics Hum Genet 2:69–101

Sidore C et al (2015) Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. Nat Genet 47:1272–1281

Welter D et al (2014) The NHGRI GWAS catalog, a curated resource of SNP-trait associations. Nucleic Acids Res 42:D1001–D1006

Zoledziewska M et al (2015) Height-reducing variants and selection for short stature in Sardinia. Nat Genet 47:1352–1356