Original Article

Understanding the population structure of North American patients with cystic fibrosis


It is generally presumed that the cystic fibrosis (CF) population is relatively homogeneous, and predominantly of European origin. The complex ethnic make-up observed in the CF patients collected by the North American CF Modifier Gene Consortium has brought this assumption into question, and suggested the potential for population substructure in the three CF study samples collected from North America. It is well appreciated that population substructure can result in spurious genetic associations. To understand the ethnic composition of the North American CF population, and to assess the need for population structure adjustment in genetic association studies with North American CF patients, genome-wide single-nucleotide polymorphisms on 3076 unrelated North American CF patients were used to perform population structure analyses. We compared self-reported ethnicity to genotype-inferred ancestry, and also examined whether geographic distribution and cystic fibrosis transmembrane regulator (CFTR) mutation type could explain the population structure observed. Although largely Caucasian, our analyses identified a considerable number of CF patients with admixed African-Caucasian, Mexican-Caucasian and Indian-Caucasian ancestries. Population substructure was present and comparable across the three studies of the consortium. Neither geographic distribution nor CFTR mutation type explained the population structure. Given the ethnic diversity of the North American CF population, it is essential to carefully detect, estimate and adjust for population substructure to guard against potential spurious findings in CF genetic association studies. Other Mendelian diseases that are presumed to predominantly affect single ethnic groups may also benefit from careful analysis of population structure.

Conflict of interest

We declare no conflict of interests.

Key words: ethnicity – population substructure – principal component analysis – population stratification

Corresponding author: Dr. Lisa J Strug, Child Health Evaluative Sciences Program, Hospital for Sick Children, Toronto, M5G 1X8, ON, Canada, bDalla Lana School of Public Health, University of Toronto, Toronto, M5T 3M7, ON, Canada, cDepartment of Statistics, University of Toronto, Toronto, M5S 3G3, ON, Canada, dDepartment of Biostatistics, School of Public Health, University of North Carolina at Chapel Hill, Chapel Hill, 27599, NC, USA, eDivision of Pediatric Endocrinology, Johns Hopkins University School of Medicine, Baltimore, 21287, MD, USA, fAdult Cystic Fibrosis Program, and Keenan Research Centre of the Li Ka Shing Knowledge Institute, St. Michael’s Hospital, Toronto, M5B 1W8, ON, Canada, gThe James Hogg iCAPTURE Centre, University of British Columbia, St Paul’s Hospital, Vancouver, V6Z 1Y6, BC, Canada, hGenetics and Genome Biology, Hospital for Sick Children, Toronto, M5G 1X8, ON, Canada, iDepartments of Pediatrics and Genetics, Case Western Reserve University, Cleveland, 44106, OH, USA, jPulmonary Research and Treatment Center, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, 27599, NC, USA, and kProgram in Physiology and Experimental Medicine, Hospital for Sick Children, Toronto, M5G 1X8, ON, Canada

© 2010 John Wiley & Sons A/S

doi: 10.1111/j.1399-0004.2010.01502.x
Cystic fibrosis (CF) is a recessive monogenic disease caused by mutations in the cystic fibrosis transmembrane regulator (CFTR) gene. Despite this simple Mendelian inheritance pattern, there is substantial variability among CF patients in their severity of lung disease and other phenotypes, even among CF patients carrying the same CFTR mutations (1); thus, environmental and non-CFTR genetic factors must contribute to overall disease variability. Genetic factors account for 54–86% of the overall variability in lung disease (2), pointing to a substantial role for genetic modifiers in CF lung disease, independent of CFTR genotype.

To identify genetic modifiers that contribute to CF disease variability, the North American CF Modifier Gene Consortium was established to conduct a genome-wide association study (GWAS) genotyping 610,000 single-nucleotide polymorphisms (SNPs) and copy number variation (CNV) probes among 3076 CF patients and 665 parents.

Genome-wide association studies are a powerful tool to identify common variants associated with disease. They assume that a common founder has given rise to a disease-associated mutation such that these methods can capitalize on linkage disequilibrium (LD) between the mutation and nearby SNPs for localization. However, in a sample of ethnically diverse individuals whose disease severity may not be accounted for by a common founder, the specific pattern of marker alleles surrounding and in LD with a causal mutation may vary, hindering the potential for identification. In addition, population structure can lead to spurious association results due to differential allele frequencies and phenotypic differences in subpopulations in the sample. Therefore, care must be taken during analysis of GWAS data to understand ethnic diversity in a sample; if heterogeneity is present it must be considered in an association analysis.

As CF is considered to be a disease predominantly of European origin, it has been presumed that the CF population is relatively homogenous. However, the diverse, complex and dynamic ethnic make-up observed in the subjects collected by the North American CF Modifier Gene Consortium has required that we take a closer look at this assumption. Here we aim to (i) use our genome-wide genotype data on 3076 North American CF patients to better understand the ethnic composition of the North American CF population and (ii) assess the need for population structure adjustment in association studies of North American CF patients.

A variety of statistical software packages have been developed to detect and correct for population structure, and they are generally based on one of three methods: (i) structured association (3), (ii) genomic control (4) and (iii) principal component analysis (5). Here we use the SMARTPCA program in the EIGENSTRAT package (5, 6), which implements a principal component approach to analyze population structure in the North American CF population.

Materials and methods

Subjects

Subjects for this study were collected by the North American CF Modifier Gene Consortium, which consists of three separate CF research groups (Table 1): (i) the Canadian Cystic Fibrosis Genetic Modifier Study (Canadian population study), a population-based sample consisting of 75% of the total Canadian CF population, (ii) the University of North Carolina/Case Western Reserve University Case-Control Study, a sample ascertained on the phenotype of extremes of lung disease (7) (US extremes of phenotype study), and (iii) the US CF Twin and Sibling Study (2), a family-based study with CF twins, siblings and parents (US sibling study).

Canadian population study

In this study, 1437 unrelated Canadian pancreatic insufficient (PI) CF patients were recruited from 33 specialized CF clinics across Canada. Exocrine pancreatic function status was determined by genetic mutations (8). PI status was assigned if
Table 1. Summary of samples collected by the Canadian population study, the US extremes of phenotype study, and the US sibling study

<table>
<thead>
<tr>
<th>Study design</th>
<th>Genotyped subjects</th>
<th>Subjects included in the population structure analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canadian Population Study</td>
<td>Population-based, unrelated individuals</td>
<td>1505</td>
</tr>
<tr>
<td>US extremes of phenotype study</td>
<td>Unrelated individuals (a) with extreme values of lung function, (b) homozygous for F508del</td>
<td>1203</td>
</tr>
<tr>
<td>US sibling study</td>
<td>Families with CF twins or siblings</td>
<td>1131 CF children, 665 parents</td>
</tr>
</tbody>
</table>

Table 2. Geographic distribution of subjects included in the population structure analysis

(A) Subjects from the Canadian population study

<table>
<thead>
<tr>
<th>Province</th>
<th>AB</th>
<th>BC</th>
<th>MB</th>
<th>NB</th>
<th>NL</th>
<th>NS</th>
<th>ON</th>
<th>QC</th>
<th>SK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>209</td>
<td>178</td>
<td>46</td>
<td>13</td>
<td>40</td>
<td>110</td>
<td>491</td>
<td>314</td>
<td>36</td>
</tr>
</tbody>
</table>

(B) Subjects from the US extremes of phenotype study

<table>
<thead>
<tr>
<th>State</th>
<th>AL</th>
<th>AZ</th>
<th>CA</th>
<th>CO</th>
<th>FL</th>
<th>GA</th>
<th>IL</th>
<th>IN</th>
<th>IA</th>
<th>MD</th>
<th>MA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>18</td>
<td>3</td>
<td>99</td>
<td>31</td>
<td>26</td>
<td>47</td>
<td>9</td>
<td>32</td>
<td>15</td>
<td>30</td>
<td>24</td>
</tr>
<tr>
<td>MI</td>
<td>MN</td>
<td>MS</td>
<td>MO</td>
<td>NE</td>
<td>NH</td>
<td>NJ</td>
<td>NM</td>
<td>NY</td>
<td>NC</td>
<td>OH</td>
<td></td>
</tr>
<tr>
<td>49</td>
<td>48</td>
<td>6</td>
<td>42</td>
<td>12</td>
<td>15</td>
<td>16</td>
<td>9</td>
<td>55</td>
<td>179</td>
<td>165</td>
<td></td>
</tr>
<tr>
<td>PA</td>
<td>RI</td>
<td>SC</td>
<td>TN</td>
<td>TX</td>
<td>UT</td>
<td>VA</td>
<td>WA</td>
<td>WV</td>
<td>WI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>72</td>
<td>13</td>
<td>22</td>
<td>38</td>
<td>17</td>
<td>23</td>
<td>20</td>
<td>42</td>
<td>22</td>
<td>4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(C) Subjects from the US sibling study

<table>
<thead>
<tr>
<th>State</th>
<th>AK</th>
<th>AL</th>
<th>AR</th>
<th>AZ</th>
<th>CA</th>
<th>CO</th>
<th>CT</th>
<th>DC</th>
<th>DE</th>
<th>FL</th>
<th>FO</th>
<th>GA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>4</td>
<td>27</td>
<td>8</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>9</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>IA</td>
<td>ID</td>
<td>IL</td>
<td>IN</td>
<td>KS</td>
<td>KY</td>
<td>LA</td>
<td>MA</td>
<td>MD</td>
<td>ME</td>
<td>MI</td>
<td>MN</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>6</td>
<td>22</td>
<td>22</td>
<td>10</td>
<td>4</td>
<td>2</td>
<td>14</td>
<td>13</td>
<td>2</td>
<td>12</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>MO</td>
<td>MS</td>
<td>MT</td>
<td>NC</td>
<td>ND</td>
<td>NE</td>
<td>NH</td>
<td>NJ</td>
<td>NM</td>
<td>NY</td>
<td>OH</td>
<td>OK</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>5</td>
<td>11</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>11</td>
<td>5</td>
<td>34</td>
<td>30</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>OR</td>
<td>PA</td>
<td>RI</td>
<td>SC</td>
<td>TX</td>
<td>UT</td>
<td>VA</td>
<td>VT</td>
<td>WA</td>
<td>WI</td>
<td>WV</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>26</td>
<td>1</td>
<td>8</td>
<td>16</td>
<td>15</td>
<td>17</td>
<td>2</td>
<td>7</td>
<td>10</td>
<td>5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

both CFTR mutations were known to be associated with pancreatic insufficiency (9); otherwise, the subject was categorized as pancreatic sufficient and excluded from analysis. The study sample was representative of the overall Canadian PI CF population (Table 2A for a summary of the geographic distribution of the subjects). Ethnicity data collected was broadly categorized as ‘Caucasian’, ‘Asian’, ‘African’, ‘Other’ and ‘Mixed’, and was based on self-report.

US extremes of phenotype study

The extremes of phenotype study collected and genotyped 1203 unrelated CF patients who (i) were homozygous for F508del, the most common CFTR mutation and (ii) had lung function at the extremes of the population distribution. With respect to self-reported ethnicity, 1162 (97%) out of the 1203 subjects categorized themselves as ‘Caucasian’; whereas 41 subjects reported they were ‘Hispanic’ (n = 18), ‘African American’ (n = 12), ‘Asian’ (n = 1) and of ‘Other’ ethnicity (n = 10) (specific composition usually defined by patients). The geographic distribution of these subjects is shown in Table 2B.

US sibling study

The sibling study consisted of 1796 genotyped CF children and parents from 559 unique families. Selection of twins or siblings resulted in a milder and younger sample. We included 436 unrelated PI CF children from this sample in our population structure analysis. The sibling study collected more detailed information on self-reported ethnicity, including the categories of ‘Caucasian’, ‘Hispanic’, ‘African American’, ‘Asian’ and ‘Middle Eastern’. For admixed subjects, the ethnic backgrounds of the parents were recorded in a composite manner as follows: ‘African, Caucasian’, ‘Aleut, Caucasian’, ‘Asian, Caucasian’, ‘Hispanic, Caucasian’, ‘Middle Eastern, Caucasian’ and ‘Native American, Caucasian’. The geographic
Understanding the population structure of North American patients with cystic fibrosis

Genotype and quality control

To ensure data quality and comparability, the three study samples within the North American CF Modifier Gene Consortium were genotyped simultaneously on the same platform, using the Illumina Infinium 610K array (610,000 SNPs and CNV probes) at McGill University and the Genome Quebec Innovation Centre.

Quality control procedures were carried out simultaneously at a single centre using a common set of rules and thresholds on all three study samples. Subjects with high missing genotype rate (>10%) were removed, and identity by descent (IBD) statistics were calculated using PLINK (10) to detect and remove cryptically related subjects. We excluded SNPs with high missing rate (>10%), and SNPs that departed from Hardy-Weinberg equilibrium (p < 0.001), or had low minor allele frequency (<0.01) were flagged. The number of SNPs remaining after quality control procedures was 556,445.

Analysis of population structure

Population structure analysis of the North American CF population was performed using the smartpca program in eigenstrat. SNPs with low minor allele frequency (<0.05) were excluded from principal component analyses. To reduce the LD between SNPs, PLINK was used to conduct SNP-pruning: we considered a sliding window of 1500 SNPs and calculated \( r^2 \) between each pair of SNPs in that window; one of a pair of SNPs was removed if the pairwise \( r^2 \) was greater than 0.2, then the window shifted 100 SNPs forward and the procedure was repeated. We used three-dimensional (3D) scatter plots of the top three significant principal components inferred by smartpca to visualize population structure. Self-reported ethnicity was compared to population structure inferred by principal component analyses. To ensure that the ethnic diversity and the degree of admixture were comparable, Tracy-Widom statistics and p values were calculated using eigenstrat to assess the number of statistically significant principal components. Scree plots were also constructed using the top 10 principal components, and they are commonly

<table>
<thead>
<tr>
<th>HapMap3 population</th>
<th>African American</th>
<th>CEPH Caucasian</th>
<th>Han Chinese</th>
<th>Gujarati Indian</th>
<th>Japanese</th>
<th>Mexican</th>
<th>Yoruba</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of samples</td>
<td>42</td>
<td>109</td>
<td>82</td>
<td>83</td>
<td>82</td>
<td>45</td>
<td>108</td>
</tr>
</tbody>
</table>
used to graphically determine the number of principal components to be used as covariates in an association analysis to adjust for population structure. Here we provide an example of the effect of using these principal components as covariates to adjust for population structure in a GWAS of CF patients. Although the consortium conducted a combined genome-wide association analysis for lung disease severity in a well-powered sample consisting of all the samples collected by the three study groups, here, our example consists of an analysis in the Canadian sample alone, which is the only population-based sample within the consortium. We present two quantile-quantile (Q-Q) plots from a genome-wide association analysis in this Canadian sample with and without incorporating principal components in the analysis. Scree plot was used to determine the number of principal components needed for population substructure adjustment in the Canadian sample. The Q-Q plot provides the observed by expected p values under the null hypothesis of no association across 555,168 separate multiple regression analyses; each multiple regression consists of a quantitative phenotype regressed on a genetic marker, adjusted for 7 principal components. Deviation from the identity line of a Q-Q plot can provide evidence of possible population structure. The R statistical package (13) was used to produce the Q-Q plots.

Results
Population structure

In the Canadian population sample, the top 15 principal components were statistically significant (p < 0.05) based on the Tracy-Widom test. The scree plot (Fig. S1, supporting information online) suggested that the top seven principal components accounted for a substantial fraction (1.37%) of the total variability, recognizing that even minor substructure can affect the most extreme p values. Two snapshots from a 3D plot of the top three principal components (accounting for 0.86% of the total variability) are shown in Fig. 1. Four continuous arms of variation were apparent instead of tight clusters, indicating population admixture in the Canadian sample. The limited ethnicity information available on each individual did not provide additional insight into the population substructure observed. Neither geographic differences (analysis by province) nor CFTR mutation type (Fig. S2, supporting information online) seemed to explain the population substructure. Moreover, mutation type was not significantly associated with any of the top 50 principal components in the logistic regression analysis.

When the Canadian sample was seeded with seven HapMap3 populations (Fig. 2), the CEPH Caucasian (red) and Yoruba (blue) samples formed two separate tight clusters, while the two Asian samples, Han Chinese (orange) and Japanese (aquamarine), clumped together and formed another tight cluster. The Gujarati Indian (gray), Mexican (green) and African American (purple) samples formed three short continuous arms pointing to the CEPH Caucasian cluster. The Canadian sample exhibited a tripod structure with most of the subjects clustering around the apex where the CEPH Caucasian samples were located. The rest of the Canadian subjects formed three long arms, distributing along the CEPH Caucasian-Indian,
Understanding the population structure of North American patients with cystic fibrosis

Fig. 2. Four snapshots of the 3D plot of the top three principal components extracted from the combined Canadian population study and HapMap3 samples. Color legend: Black = Canadian population study sample, Red = CEPH Caucasian, Purple = African American, Green = Mexican, Gray = Gujarati Indian, Orange = Han Chinese, Aquamarine = Japanese, Blue = Yoruba. CEPH Caucasian and Yoruba samples formed two separate tight clusters; Han Chinese and Japanese samples clumped together and formed another tight cluster; African American, Mexican and Gujarati Indian samples formed three short continuous arms pointing to the CEPH Caucasian cluster. The Canadian population study samples exhibited a tripod structure, with most of the samples forming a cluster at the apex overlapping the CEPH Caucasian cluster, and the rest of the samples formed three long arms, distributing along the CEPH Caucasian-African American, CEPH Caucasian-Mexican and CEPH Caucasian-Gujarati Indian arms.

CEPH Caucasian-Mexican and CEPH Caucasian-African American arms. The HapMap3 Asian populations did not encompass any of the Canadian subjects. Canadian patients that scattered along the CEPH Caucasian-African American arm were either self-reported ‘African’ or ‘Mixed.’ Both of the CEPH Caucasian-Mexican and CEPH Caucasian-Gujarati Indian arms consisted of individuals that self-reported as either ‘Caucasian’ or as ‘Other.’ The three self-reported ‘Asian’ individuals were not mapped onto the HapMap3 Asian cluster; one scattered around the CEPH Caucasian cluster, and the other two fell along the CEPH Caucasian-Gujarati Indian arm.

Tracy-Widom tests indicated 9 and 13 significant principal components in the extremes of phenotype and sibling studies, respectively. Separate scree plots (Fig. S1) indicated that the top four principal components accounted for a substantial proportion...
Fig. 3. (a) A snapshot of the 3D plot of the top three principal components extracted from the US extremes of phenotype study sample. Points were colored by self-reported ethnicity. Color legend: Black = ’Caucasian’, Green = ’Hispanic’, Purple = ’African American’, Orange = ’Other’. (b) A snapshot of the 3D plot of the top three principal components extracted from the US sibling study sample. Points are colored by self-reported ethnicity. Color legend: Black = ’Caucasian’, Green = ’Hispanic’ or ’Hispanic, Caucasian’, Purple = ’African American’ or ’African American, Caucasian’, Gray = ’Middle Eastern’ or ’Middle Eastern, Caucasian’, Brown = ’Aleut, Caucasian’, Blue = ’Native American, Caucasian’. Three distinct arms are observed in each of the two US study samples, with one arm encompassing the majority of the subjects. The major arm consists of mostly self-reported ’Caucasian’ samples; and samples with ’Hispanic’ or ’African American’ ancestral background dominated the other two arms, respectively.

(1% and 2.15%, respectively) of the total variability in each of the two samples. The extremes of phenotype and sibling study samples were very similar to the Canadian sample in that they both exhibited arm-like structures in the 3D plot of the top three principal components, with most of the samples scattered around one of the arms. However, both US samples displayed three arms in contrast to the Canadian sample, which revealed four distinct arms (Figs. S3 and S4, supporting information online, respectively). In both US samples, the self-reported ’Caucasian’, ’African American’ and ’Hispanic’ samples dominated each of the three arms (Fig. 3). When each of the US samples was seeded with HapMap3 data, the three arms were mapped onto the CEPH Caucasian cluster, the CEPH Caucasian-African American arm and the CEPH Caucasian-Mexican arm, respectively. In general, the population substructure detected by principal component analysis agreed with self-reported ethnicity in both US samples; however, there were exceptions. For example, in the extremes of phenotype sample, one self-reported ’Hispanic’ subject (Fig. S5, supporting information online) fell along the CEPH Caucasian-African American arm. In the siblings sample, one self-reported ’Caucasian’ subject (circled in green) distributed away from the CEPH Caucasian cluster; and one self-reported ’Hispanic’ individual (circled in red) scattered along the CEPH Caucasian-Gujarati Indian arm; and three individuals (circled in blue) with self-reported ’Asian’ or ‘Asian, Caucasian’ ethnicity did not scatter around either of the CEPH Caucasian or the Asian clusters, one fell along the CEPH Caucasian-African American arm and the other two fell along the CEPH Caucasian-Mexican arm (Fig. S6, supporting information online).

When we analyzed the US and Canadian samples combined, four arms were present in the 3D plot of the top three principal components (Fig. S7, supporting information online). Individuals from each of the three study samples distributed evenly along the arms, with the exception of a single arm, which consisted almost exclusively of individuals from Canada. When we seeded the three study samples with HapMap3 data (Fig. 4), the structure was very similar to that shown in Fig. 2, and the four arms we observed (Fig. S7, supporting information online) clearly projected onto the CEPH Caucasian cluster, the CEPH Caucasian-Mexican, CEPH Caucasian-African American and CEPH Caucasian-Gujarati Indian arms, with the majority of the samples in a cluster overlapping the CEPH Caucasian cluster. The arm that consisted exclusively of Canadian samples overlapped the CEPH Caucasian-Gujarati Indian arm; all these patients had their self-reported ethnicity as ’Other’.

This analysis indicates that there appears to be population substructure in all three study samples, with the ethnic make-up of the patients in each study comparable. How might this substructure
Fig. 4. Four snapshots of the 3D plot of the top three principal components extracted from the combined North American CF consortium and HapMap3 samples. Color legend: Black = North American CF samples, Purple = African American, Red = CEPH Caucasian, Orange = Han Chinese, Gray = Gujarati Indian, Aquamarine = Japanese, Green = Mexican, Blue = Yoruba. The HapMap3 samples formed three continuous arms and two separate tight clusters similar to that shown in Fig. 2. The North American CF samples scattered along the three arms and around the CEPH Caucasian cluster, and the CEPH Caucasian-Gujarati Indian arm consists of almost exclusively samples from the Canadian population study.

affect an association analysis in the North American CF population?

Effects on GWAS

We performed a genome-wide association analysis in the Canadian sample only to illustrate the effect of population substructure. Q-Q plots with and without adjusting for population substructure using the top seven principal components, as suggested by the scree plot, are provided in Fig. 5a,b, respectively. Although the Tracy-Widom test nominated the top 15 principal components as significant, we chose to use the top 7 principal components for the sake of parsimony; and the association results did not differ with the inclusion of the extra principal components. Deviation from the identity line is observed in the absence of covariate adjustment with the top 7 principal components, but removed with adjustment, suggesting bias in the observed p values because of population substructure.

Discussion

This is the first study to evaluate the ethnic make-up of the North American CF population using genetic data. Our study revealed clearly defined population substructures in this CF population, and
a similar structure was present across the Canadian and US CF samples. The continuous arms (instead of tight clusters) observed in the North American CF population are indicative of population admixture. In both of the US samples, substructures identified by principal component analysis showed relatively good agreement with self-reported ethnicity; although, there were individual cases where self-reported ethnicity disagreed with genotype-inferred ethnicity.

Self-reported ethnicity is widely used in association studies to account for ancestral differences. However, it has limited reliability (14) and is inconsistent in reporting over time (15). Self-reported ethnicity has also been shown to be insufficient in removing population stratification in genetic association analyses (16), presumably due to many of the known limitations, including those observed in this study. In populations with a large degree of complexity, such as the North American CF population, it is extremely difficult to define discrete ethnic categories that can fully capture the population diversity and the degree of admixture. Moreover, the lack of agreement between genotype-inferred ancestry and self-reported ethnicity may be because of the complexity of ethnicity itself. Ethnicity is a combination of not only genetic heritage but also one’s self-identity, making it a complex construct involving many factors.

Despite the limitations for population structure adjustment, we feel capturing detailed self-reported ethnicity is an important complement to genetic ancestry information in order to understand the ethnic make-up of a sample. This study highlights that composite ethnic categories that are highly detailed would be necessary to capture multiple ancestral backgrounds of admixed individuals.

The CEPH Caucasian, Han Chinese together with Japanese and Yoruba samples from HapMap3, representing the three major ethnic groups, ‘Caucasian’, ‘Asian’ and ‘African’, are the most commonly used reference populations in population structure analysis. In our principal component analysis, the majority of the self-reported ‘Caucasian’ samples clustered well with the CEPH Caucasian samples, as expected. There was, however, some degree of distribution away from the CEPH Caucasian cluster, highlighting both admixture and limitations of self-reported ethnicity. The two relatively homogeneous HapMap3 Asian populations, Han Chinese and Japanese, did not encompass any of the North American CF samples. In addition, self-reported ‘Asian’ subjects showed no clustering with the HapMap3 Asian populations, which suggests the limitations of using pure Asian populations as reference when studying population structure in the North American CF population.

Our principal component analyses suggest that the North American CF samples lying along the CEPH Caucasian-Mexican arm are admixed with Mexican and European ancestries, and those along the CEPH Caucasian-African American arm are admixed with African and European ancestries. Individuals that scattered along the
Understanding the population structure of North American patients with cystic fibrosis

CEPH Caucasian-Gujarati Indian arm are admixed with northwestern Indian and European ancestries, and this group of CF patients was only observed in Canada. One possible explanation for the lack of admixed Indian patients in the US samples is that they do indeed exist, but they were simply not captured by the two US studies. Although the US samples included patients distributed across 48 US states, the majority of the subjects were collected from the eastern United States. An alternative explanation is a diagnosis bias (17). Since only the Canadian sample was a population-based sample, we cannot purport to conclude that we have a clear ethnic representation of the US CF patients.

In admixed populations, it is important to guard against biased and spurious association results by carefully estimating structure, understanding the study sample and adjusting for population stratification when possible. In our CF admixed population we have seen that we can avoid spurious findings by carefully estimating structure and adjusting for population stratification. We used genome-wide data to infer continuous ancestry to fully capture the diversity and admixture of a given population. In the absence of genome-wide data, there are other methods available to adjust for population structure in genetic association studies (4, 18).

Although predominantly Caucasian in origin, the North American CF population clearly has a complex ethnic make-up that is likely to change over time. For this reason we advocate for careful analysis of population structure in all CF genetic association studies. Moreover, these findings have implications for genetic association studies in other simple Mendelian diseases presumed to predominantly affect single ethnic groups, such as alpha-1 antitrypsin deficiency and sickle cell disease. Genetic association studies in these populations may also benefit from careful analysis of population structure.

Supporting Information

The following Supporting information is available for this article:

Fig. S1. Scree plots for the Canadian population study, the US extremes of phenotype study and the US sibling study.

Fig. S2. Snapshot of the 3D plot of the top three principal components extracted from the Canadian population study sample alone. CF patients who are homozygous for F508del are colored in red; and those with other mutation types are colored in black. The red and black dots are evenly distributed along each of the four arms, suggesting the population structure observed in the Canadian population study samples is not due to difference in the CFTR mutation type.

Fig. S3. Two snapshots of the 3D plot of the top three principal components extracted from the US extremes of phenotype study samples. Three distinct arms are present, with most of the samples distributing along one arm.

Fig. S4. Two snapshots of the 3D plot of the top three principal components extracted from the US sibling study samples. Three distinct arms are present, with most of the samples distributing along one arm.

Fig. S5. A snapshot of the 3D plot of the top three principal components extracted from the combined US extremes of phenotype study and the HapMap3 samples. Color Legend: Black = US extremes of phenotype study sample. Purple = African American, Red = CEPH Caucasian, Orange = Han Chinese, Grey = Gujarati Indian, Aquamarine = Japanese, Green = Mexican, Blue = Yoruba. The point highlighted in the red circle indicates discrepancy between self-reported ethnicity and ethnicity estimated using one’s genotype, where a self-reported ‘Hispanic’ sample distributed along the CEPH Caucasian-African American arm, suggesting his/her admixed Caucasian and African ancestries.

Fig. S6. A snapshot of the 3D plot of the top three principal components extracted from the combined US sibling study samples and HapMap3 samples. Color Legend: Black = US sibling study samples, Purple = African American, Red = CEPH Caucasian, Orange = Han Chinese, Grey = Gujarati Indian, Aquamarine = Japanese, Green = Mexican, Blue = Yoruba. Points highlighted in circles are indicative of discrepancy between self-reported ethnicity and ethnicity estimated using one’s genotype. One self-reported ‘Caucasian’ subject (circled in green) distributed away from the CEPH Caucasian cluster; one self-reported ‘Hispanic’ individual (circled in red) scattered along the CEPH Caucasian-Gujarati Indian arm; and three individuals (circled in blue) with self-reported ‘Asian’ or ‘Asian, Caucasian’ ethnicity did not scatter around either of the CEPH Caucasian or the Asian clusters, one fell along the CEPH Caucasian-African American arm and the other two fell along the CEPH Caucasian-Mexican arm.

Fig. S7. A snapshot of the 3D plot of the top three principal components extracted from the North American CF samples. Color legend: Red = Canadian population study samples, Green = US extremes of phenotype study samples, Blue = US sibling study samples. Individuals from each of three studies distributed evenly along 3 of the 4 arms; one of the arms consisted almost exclusively of individuals from the Canadian population study.

Additional Supporting information may be found in the online version of this article.

Please note: Wiley-Blackwell Publishing is not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Acknowledgements

The authors would like to thank Nicole Anderson, Katherine Keenan, Kathleen Naughton, and Rhonda Pace for coordinating patient recruitment and race/ethnicity verification; Johanna Rommens, Mike Collaco, Jaclyn Stonebraker, Wanda O’Neal and Jack Kent for their helpful discussions and comments on study direction; David Cutler for initial cleaning of SNP data; and the International HapMap Consortium for the genotype data.

The authors sincerely appreciate the significant contribution from the CF Directors and Clinic Coordinators at each CF center that enrolled subjects and ascertained phenotype data and DNA, as part of the Canadian Cystic Fibrosis Genetic Modifier Study, the University of North Carolina/Case Western Reserve University Case-Control Study and the US CF Twin and Sibling Study. The authors express deep gratitude to all CF patients and their families. This work was supported by Genome Canada through the Ontario Genomics Institute [2004-OGI-3-05 to P.D.]; the Canadian Cystic Fibrosis Foundation [L.J.S., P.D., A.J.S.]; the Ontario Research Foundation [L.J.S., P.D.]; the Natural Sciences and Engineering Sciences Institute; the National Institutes of Health, National Human Genome Research Institute [HL-052031 to B.C.C.]; the National Institutes of Health, National Institute on Aging [AG-029431 to B.C.C.]; and the Ontario CF Foundation [L.J.S., P.D., A.J.S.].
References


