Why do we need to account for ancestry in genetic disease studies?

An illustrative example from Lander and Schork [2006]:

... suppose that a would-be geneticist set out to study the “trait” of ability to eat with chopsticks in the San Francisco population by performing an association study with the HLA complex. The allele HLA-A1 would turn out to be positively associated with ability to use chopsticks—not because immunological determinants play any role in manual dexterity, but simply because the allele HLA-A1 is more common among Asians than Caucasians.
Ancestry acts as a hidden confounder

We observe an apparent association ...

... but the true cause lies here.
A real world example, and its lessons

- The classic Pima Indian Diabetes study [Knowler et al., 1988]

![Diagram showing Gm^3;5,13,14,15 haplotype leading to lowered risk of Type II Diabetes due to Caucasian ancestry]

- Lessons from this study:
  - Ancestry needs to be accounted for or else we risk false positive associations.
  - Ancestry is fractional, not discrete. Classical stratified designs inadequate.
  - Self-reported ancestry often inaccurate. “Cryptic ancestry.”
How can we deal with population structure?

Five options:

1. Only do studies within homogeneous populations
2. Family-based designs (TDT, FBAT, ...)
3. Genomic Control: correct the critical value for the test statistic based on an estimate of heterogeneity
4. Estimate the ancestry of each individual, and then use the estimates to statistically correct for the population structure
5. Use a linear mixed model approach

Recommendation: (4) and (5) have the best power. (1) is unrealistic!
Ancestry Estimation

- The problem:
  
  **Input:** a matrix of multilocus genotypes,
  
  **Output:** an ancestry vector for each individual.

- The solutions:

<table>
<thead>
<tr>
<th>Pedigree</th>
<th>Unrelated individuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Mendel option 15</td>
<td>• PCA approaches (EIGENSTRAT)</td>
</tr>
<tr>
<td></td>
<td>• Soft-clustering approaches (ADMIXTURE, STRUCTURE)</td>
</tr>
</tbody>
</table>
Principal component analysis

Condense a high dimensional genotype matrix (say 500K markers × 1000 individuals) to a lower-dimensional matrix (say 2 coordinates × 1000 individuals) that retains as much variation as possible among the 1000 individuals.

New coordinates retain a lot of information about the individuals’ population-level ancestry.

Implemented in popular program EIGENSTRAT.
The direction of the PC1 axis and its relative strength may reflect a special role for this geographic axis in the demographic history of Europeans (as first suggested in ref. 10). PC1 aligns north-northwest/south-southeast (NNW/SSE, $2^{\circ}16$ degrees) and accounts for approximately twice the amount of variation as PC2 (0.30% versus 0.15%, first eigenvalue $54.09$, second eigenvalue $52.04$). However, caution is required because the direction and relative strength of the PC axes are affected by factors such as the spatial distribution of samples (results not shown, also see ref. 9). More robust evidence for the importance of a roughly NNW/SSE axis in Europe is that, in these same data, haplotype diversity decreases from south to north (A.A. et al., submitted). As the fine-scale spatial structure evident in Fig. 1 suggests, European DNA samples can be very informative about the geographical origins of their donors. Using a multiple-regression-based assignment approach, one can place 50% of individuals within 310 km of their reported origin and 90% within 700 km of their origin (Fig. 2 and Supplementary Table 4, results based on populations with $n$.6). Across all populations, 50% of individuals are placed within 540 km of their reported origin, and 90% of individuals within 840 km (Supplementary Fig. 3 and Supplementary Table 4). These numbers exclude individuals who reported mixed grandparental ancestry, who are typically assigned to locations between those expected from their grandparental origins (results not shown). Note that distances of assignments from reported origin may be reduced if finer-scale information on origin were available for each individual.

Population structure poses a well-recognized challenge for disease-association studies (for example, refs 11–13). The results obtained here reinforce that the geographic distribution of a sample is important to consider when evaluating genome-wide association studies.

... figure from Novembre et al. [2008] shows that PCA can recover ancestry very precisely.
Model-based ancestry estimation

Given SNP genotype matrix, computes maximum likelihood estimates for each individual’s ancestry fractions $q_i = (q_{i1}, q_{i2}, \ldots, q_{iK})$ and the ancestral major allele frequencies $f_j = (f_{1j}, f_{2j}, \ldots, f_{Kj})$.

$$L(Q, F) = \sum_i \sum_j \left\{ g_{ij} \ln \left[ \sum_k q_{ik} f_{kj} \right] + (2 - g_{ij}) \ln \left[ \sum_k q_{ik} (1 - f_{kj}) \right] \right\}$$

Program ADMIXTURE
How to use ancestry estimates to correct for population structure

For an association study, the simplest approach is to add ancestry as an additional covariate (vector valued) in the generalized linear regression model.

- Uncorrected model:
  \[ g(\mathbb{E}(y_i | x_i)) = \alpha + \beta x_i \]
  \( y \) trait, \( x \) genotype at test marker; \( g(\cdot) \) link function.

- Corrected model:
  \[ g(\mathbb{E}(y_i | x_i, q_i)) = \alpha + \beta x_i + \gamma^t q_i, \]
  \( q_i = (q_{i1}, q_{i2}, \ldots q_{i(K-1)}) \) the ancestry estimate.
Another approach: don’t use ancestry estimates, just try to more correctly account for the variance-covariance matrix by using empirical kinship estimates.

\[
Y = X\beta + \epsilon,
\]

\[
\text{Cov} \ \epsilon = 2\sigma_a^2 \Phi + \sigma_e^2 I
\]

Better accounts for hidden relatedness amongst individuals. Implemented in programs EMMA, EMMAX.
Summary

- Ancestry should be accounted for when doing disease studies.
- There are methods for estimating ancestry in related or unrelated individuals, and these estimates can be used to correct the analysis.
- Or use a mixed model approach for association.

