Genetic Disease Studies in Structured Populations

David Alexander

August 18, 2010

What is population structure?

People in the United States draw their ancestry from many distinct worldwide populations that evolved in partial isolation. This is the essence of a *structured* or *stratified* population. Furthermore, there are people who have partial ancestry stemming from several different ancestral populations (*admixture*).

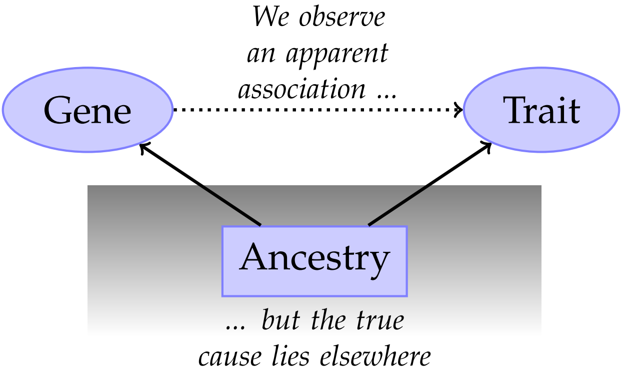
What are the implications for genetic epidemiology?

We need to account for ancestry in genetic disease studies?

An light-hearted example from [Lander & 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 …

Gene

Trait

Ancestry

… but the true cause lies elsewhere

A simple example of confounding

European sample, $n = 120$

	Disease	
	+	-
WT	60	20
Mutant	30	10

$$\widehat{OR} = 1$$

Amerindian sample, $n = 120$

	Disease	
	+	-
WT	10	30
Mutant	20	60

$$\widehat{OR} = 1$$

A simple example of confounding

European sample, $n = 120$

	Disease	
	+	-
WT	60	20
Mutant	30	10

$$\widehat{OR} = 1$$

Amerindian sample, $n = 120$

	Disease	
	+	-
WT	10	30
Mutant	20	60

$$\widehat{OR} = 1$$

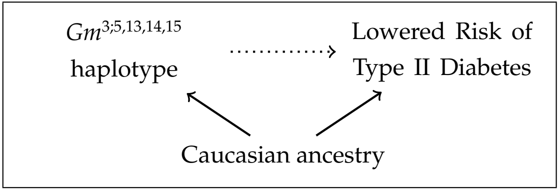
Pooled samples, $n = 240$

	Disease	
	+	-
WT	70	50
Mutant	50	70

$$\widehat{\mathbf{OR}} = \mathbf{1.96}$$

A real world example, and its lessons

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



$Gm^{3;5,13,14,15}$ haplotype

Lowered Risk of Type II Diabetes

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.
- Another lesson: Self-reported ancestry often inaccurate. “Cryptic ancestry.”

How can we deal with population structure?

Four 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 the effect of population structure
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!

A word on genomic control: don't use it!

Genomic control attempts to measure how much population structure increases the evidence for association, by calculating the genome-wide median χ^2_1 association statistic, termed λ_{GC}, the GC *inflation factor*.

All association χ^2_1 statistics are then divided by λ_{GC} before calculating p values.

GC is underpowered but *not conservative*—will *undercorrect* in regions where generic divergence is greater than average (perhaps due to selection or other factors).

Ancestry Estimation

- The problem:

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

- The solutions:

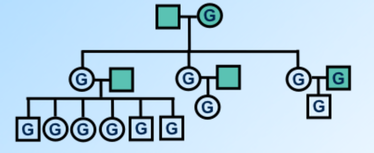
Pedigree	**Unrelated individuals**
• Mendel option 15	• PCA approaches (EIGENSTRAT)
	• Soft-clustering approaches (ADMIXTURE, STRUCTURE)

A little terminology

- A **SNP genotype** $g_{ij} \in \{0, 1, 2\}$ is the number of markers of allele type '1' occurring at a locus j in individual i.
- **Principal component analysis (PCA)** reduces each individual's SNP genotypes (100K-1M+) to a K-dimensional vector (K small, often 2 or 3 here) while retaining information about ancestry.
- **Model-based soft-clustering approaches** assume that there are K ancestral populations and attempt to learn what proportion of individual i's ancestry comes from population k. These proportions q_{ik} satisfy $\sum_k q_{ik} = 1$.
- **Ancestry-informative markers (AIMs)** are genetic markers that have different allele frequencies in different populations.

Pedigree-based ancestry estimation (Mendel option 15)

- MENDEL maximizes the pedigree likelihood as a function of the founders' ancestries. Ancestries of non-founders are then found by averaging their parents' ancestries.
- The allele frequencies in the K ancestral populations must be provided to MENDEL
- Can handle any kind of markers
- Works best when used with AIMs.



The pedigree likelihood

$$L(Q) = \sum_{\mathbf{g}_1} \cdots \sum_{\mathbf{g}_n} \prod_i \mathrm{Prior}(\mathbf{g}_i \mid \mathbf{q}_i) \prod_{\{j,k,l\}} \mathrm{Tran}(\mathbf{g}_l \mid \mathbf{g}_j, \mathbf{g}_k) \prod_m \mathrm{Pen}(\mathbf{x}_m \mid \mathbf{g}_m)$$

At locus j allele ‘1’ occurs with frequencies $\mathbf{f}_j = (f_{1j}, f_{2j}, \ldots, f_{Kj})$ in the K populations. A founder with ancestry $\mathbf{q}_i = (q_{i1}, q_{i2}, \ldots, q_{iK})$ has allele ‘1’ at locus j with probability

$$p_{ij} = \sum_{k=1}^{K} q_{ik} f_{kj}.$$

Then assuming gametes sampled independently,

$$\mathrm{Prior}(g_{ij} \mid q_i) = \Pr(g_{ij} \mid q_i) = \begin{cases} p_{ij}^2, & g_{ij} = 2; \\ 2p_{ij}(1 - p_{ij}), & g_{ij} = 1; \\ (1 - p_{ij})^2, & g_{ij} = 0. \end{cases}$$

Offspring ancestries

Once we estimate the founders' ancestries, each offspring's ancestry estimate is calculated as the average of her parents' ancestries:

$$\mathbf{q}_c = (\mathbf{q}_m + \mathbf{q}_f)/2.$$

Thus we have calculated ancestries for all individuals in the pedigree. MENDEL also will provide standard errors.

MENDEL Example

- 76 offspring and 33 founders in 6 extended Mestizo families (27 nuclear families).
- 89 of these individuals are genotyped at 9 unlinked markers. Ancestry Informative Markers (AIMs) chosen so that the allele frequencies differ by at least 0.30 between ethnic groups.
- There are a number of potential ancestral Amerindian nations. Use an average of Cheyenne, Mayan, Nahua, Pima, Pueblo allele frequencies. Markers chosen so the allele frequencies are within 0.10 between these nations.

MENDEL Example (2)

- Try it!

  ```
  % mendel -c Control15a.in
  ```

  in the “15 Ethnic Admixture” directory.
- Look at the summary file, and also look at the resulting new pedigree file Ped15a.out—what’s different?
- Note how # of pops and allele frequencies specified

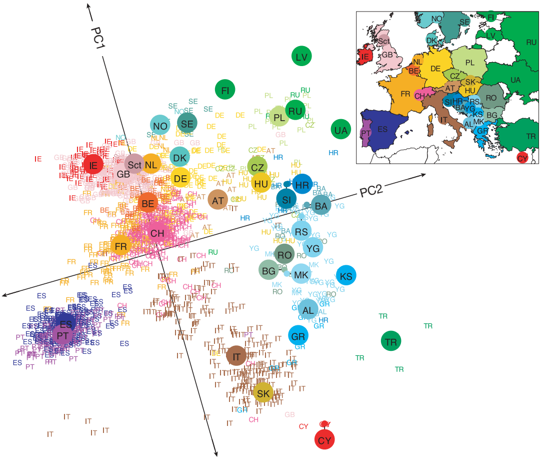
Ancestry estimation for unrelated individuals: Principal component analysis

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.

Principal component analysis (2)



... figure from [Novembre & Stephens, 2008] shows that PCA can recover ancestry very precisely.

Ancestry estimation for unrelated individuals: model-based approach

ADMIXTURE reads autosomal SNP datasets for unrelated individuals and estimates both the admixture fractions q_{ik} and the ancestral allele frequencies for the K populations, f_{kj}. Can handle large datasets and does not require the user to provide allele frequencies or specify which alleles are AIMs.

ADMIXTURE efficiently maximizes the (approximate) likelihood as a function of each individual's ancestry $\mathbf{q_i} = (q_{i1}, q_{i2}, \ldots, q_{iK})$ and the ancestral major allele frequencies $\mathbf{f_j} = (f_{1j}, f_{2j}, \ldots, f_{Kj})$:

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

Using ADMIXTURE

ADMIXTURE is not yet fully integrated into MENDEL, but let's see how to use it and use the results in a disease study we are doing with MENDEL.

Let's say you have a binary pedigree file myStudy.bed containing 100K SNP markers for 1000 Latino individuals. Historically, Latinos derive ancestry from Amerindian and European populations, so we will consider $K = 2$. You can process our dataset with ADMIXTURE with the following simple command:

```
% admixture myStudy.bed 2
```

Using ADMIXTURE (2)

ADMIXTURE’s output of estimated ancestry fractions and allele frequencies is put in files myStudy.2.Q and myStudy.3.F

myStudy.2.Q:

```
1        0.9999  0.0001
2        0.8000  0.2000
3        0.1000  0.9000
4        0.2000  0.8000
5        ...
```

The i^{th} line is the ancestry vector for the i^{th} individual. For example, individual 2 is estimated at having 80% of her genes from population 1, 20% from population 2.

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 regression model.

- Uncorrected model:

$$g(\mathrm{E}(y_i \mid x_i)) = \alpha + \beta x_i$$

  y trait, x genotype at test marker; $g(\cdot)$ link function.

- Corrected model:

$$g(\mathrm{E}(y_i \mid x_i, q_i)) = \alpha + \beta x_i + \gamma^t \mathbf{q}_i,$$

  $\mathbf{q}_i = (q_{i1}, q_{i2}, \dots q_{i(K-1)})$ the ancestry estimate.

Caveat!

When incorporating ancestry estimates into the regression model, be sure to only use $(K-1)$ entries of the ancestry vector from ADMIXTURE or MENDEL Option 15. Remember that the full vector is linearly constrained by $\sum_{k=1}^{K} q_{ik} = 1$!

There is no restriction on how many entries of the ancestry vector you can use from EIGENSTRAT.

Correcting for population structure with MENDEL

- MENDEL Option 15 generates a new pedigree file which includes the ancestry fractions. ADMIXTURE doesn't yet do this.
- Can use the new pedigree for a corrected analysis if we update the Def and Control files:
  - Def file: add

    ```
    POP1, Variable
    POP2, Variable
    ...
    ```

  - Control file: add

    ```
    PREDICTOR=POP1::TRAIT
    PREDICTOR=POP2::TRAIT
    ...
    ```

Mixed model approach

Does *not* require ancestry estimates.

Rather, the mixed model approach just tries to more correctly account for the variance-covariance matrix by using empirical kinship estimates.

$$\mathbf{Y} = X\beta + \varepsilon,$$

$$\text{Cov } \varepsilon = 2\sigma_a^2\Phi + \sigma_e^2 I$$

Better accounts for hidden relatedness amongst individuals than other methods. Implemented in programs EMMA, EMMAX, TASSEL.

Summary

- Ancestry should be controlled or corrected 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
- Mixed model approaches work best when there is substantial (hidden) interrelatedness amongst case-control samples

Recommended Reading

[1] A.L. Price, N.A. Zaitlen, D. Reich, and N. Patterson. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, 11(7):459–463, 2010.

[2] A.L. Price, N.J. Patterson, R.M. Plenge, M.E. Weinblatt, N.A. Shadick, and D. Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, 2006.

[3] D.H. Alexander, J. Novembre, and K. Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome research*, 19(9):1655, 2009.