Lecture 13: Population Heterogeneity in Genetic Association Studies

Jurg Ott, Ph.D.
Institute of Psychology
Chinese Academy of Sciences, Beijing
Rockefeller University, New York

Overview

Lecture 8 considered heterogeneity in linkage analysis, that is, differences of the recombination fraction from family to family. Here, we are interested in the effects of population heterogeneity in case-control association studies.

I. Fixed effects (potentially recognizable groups)
II. Random effects, admixture
   1. Measuring amount of admixture, $F_{st}$
   2. Methods for working with admixed populations
      a. STRUCTURE
      b. EIGENSTRAT
      c. AIMS markers
      d. MDS components
      e. Randomization tests
   3. Genetic heterogeneity in human disease
I. Fixed Effects

Allowing for Fixed Effects

- Identify potential confounders: Race, sex, age, height?, weight/BMI?
- Use as covariates in logistic regression with case/control as dependent variables, $f = \text{probability of being a case}$:

$$\log\left(\frac{f}{1-f}\right) = a_0 + a_1 x_1 + ...$$
II. Admixture
1. Assessing degree of admixture

- No unequivocally known subgroups
- Measure differences in allele frequencies, $p$, but via $2p(1 - p)$: Heterozygosity under HWE
- $H_T$ = heterozygosity in population
- $H_S$ = average heterozygosity over subpopulations
- $F_{ST} = (H_T - H_S)/H_T$

Sample $F_{ST}$ values

- Mechanisms leading so substructure: Assortative mating, population admixture, inbreeding
- Effect: Reduction in heterozygosity

<table>
<thead>
<tr>
<th>Population</th>
<th>$F_{ST}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human (Africa, Europe, East Asia)</td>
<td>0.069</td>
</tr>
<tr>
<td>Human (Yanomama Indian Villages)</td>
<td>0.077</td>
</tr>
<tr>
<td>House mouse</td>
<td>0.113</td>
</tr>
<tr>
<td>Drosophila equinoxialis</td>
<td>0.109</td>
</tr>
<tr>
<td>Lycopod plant</td>
<td>0.282</td>
</tr>
</tbody>
</table>
Assumed data: Single Locus

- Single locus; 2 alleles; 2 populations, each in HWE, mix in 1:1 ratio

<table>
<thead>
<tr>
<th>Pop</th>
<th>p</th>
<th>AA</th>
<th>Aa</th>
<th>aa</th>
<th>F_{ST}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.2</td>
<td>0.04</td>
<td>0.32</td>
<td>0.64</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.8</td>
<td>0.64</td>
<td>0.32</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>1 + 2 obs</td>
<td>0.5</td>
<td>0.34</td>
<td>0.32</td>
<td>0.34</td>
<td></td>
</tr>
<tr>
<td>1 + 2 exp</td>
<td>0.5</td>
<td>0.25</td>
<td>0.50</td>
<td>0.25</td>
<td>0.36</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pop</th>
<th>p</th>
<th>AA</th>
<th>Aa</th>
<th>aa</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1 + 2 obs</td>
<td>0.5</td>
<td>0.50</td>
<td>0.00</td>
<td>0.50</td>
</tr>
<tr>
<td>1 + 2 exp</td>
<td>0.5</td>
<td>0.25</td>
<td>0.50</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Assumed data: Two Loci

- Two loci, genotypes independent (unassociated) in each of two populations

<table>
<thead>
<tr>
<th>Pop A</th>
<th>BB</th>
<th>Bb</th>
<th>bb</th>
<th>F_{dU}</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>196</td>
<td>588</td>
<td>441</td>
<td>1225</td>
</tr>
<tr>
<td>Aa</td>
<td>168</td>
<td>504</td>
<td>378</td>
<td>1050</td>
</tr>
<tr>
<td>aa</td>
<td>36</td>
<td>108</td>
<td>81</td>
<td>225</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>1200</td>
<td>900</td>
<td>2500</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>P_{dU}</th>
<th>BB</th>
<th>Bb</th>
<th>bb</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>196</td>
<td>168</td>
<td>36</td>
</tr>
<tr>
<td>Aa</td>
<td>588</td>
<td>504</td>
<td>108</td>
</tr>
<tr>
<td>aa</td>
<td>441</td>
<td>378</td>
<td>81</td>
</tr>
<tr>
<td></td>
<td>900</td>
<td>1200</td>
<td>400</td>
</tr>
</tbody>
</table>
Assumed data: Two Loci

- Two loci, genotypes independent (unassociated) in each of two populations, mix in 1:4 ratio

\[\begin{array}{c|c|c|c|c}
 & BB & Bb & bb \\
\hline
AA & 0.078 & 0.101 & 0.047 & 0.226 \\
Aa & 0.202 & 0.202 & 0.065 & 0.468 \\
aa & 0.144 & 0.130 & 0.032 & 0.306 \\
\hline
& 0.424 & 0.432 & 0.144 & 1 \\
\end{array}\]

\[\begin{array}{c|c|c|c|c}
 & BB & Bb & bb \\
\hline
AA & 0.096 & 0.098 & 0.033 & 0.226 \\
Aa & 0.198 & 0.202 & 0.067 & 0.468 \\
aa & 0.130 & 0.132 & 0.044 & 0.306 \\
\hline
& 0.424 & 0.432 & 0.144 & 1 \\
\end{array}\]

Assumed data: Two Loci

- Two loci, chi-square test for genotype association

\[\begin{array}{c|c|c|c|c|c|c|c|c|c}
 & BB & Bb & bb \\
\hline
AA & 980 & 1260 & 585 & 2825 \\
Aa & 2520 & 2520 & 810 & 5850 \\
aa & 1800 & 1620 & 405 & 3825 \\
\hline
& 4000 & 6000 & 2500 & 12500 \\
\end{array}\]

\[\begin{array}{c|c|c|c|c}
 & BB & Bb & bb \\
\hline
AA & 0.82 & 1.03 & 1.44 & 1 \\
Aa & 1.02 & 1.00 & 0.96 & 1 \\
aa & 1.11 & 0.98 & 0.74 & 1 \\
\hline
& 1 & 1 & 1 & 1 \\
\end{array}\]

\[\begin{array}{c|c|c|c}
Pop & chi-sq & p \\
\hline
A & 0 & 1 \\
B & 0 & 1 \\
A + 4B & 176.16 & 5.00E-37 \\
\end{array}\]
Methods: Testing HWE


- Test of HWE *may* be indicative of admixture
- Size of deviation from HWE depends on allele frequency difference
- Cases tend to be distantly related → inbreeding → reduction of heterozygosity, increase in homozygotes
- Not generally used to allow for admixture
- Still useful to see differences between cases and controls
- Detect deviations from HWE:
  - Chi-square test: Observed versus expected numbers
  - Exact tests

Working with F and allele frequencies


- $p$ = frequency of A allele, $F$ = probability of autozygosity

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>$q_1 = Fp + (1 - F)p^2$</td>
</tr>
<tr>
<td>AB</td>
<td>$q_2 = 2p(1 - p)(1 - F)$</td>
</tr>
<tr>
<td>BB</td>
<td>$q_3 = F(1 - p) + (1 - F)(1 - p)^2$</td>
</tr>
<tr>
<td>$\sum q_i$</td>
<td>1</td>
</tr>
</tbody>
</table>

$p = q_1 + q_2/2$
$F = 1 - q_2/[2p(1 - p)]$

Case-control study:

<table>
<thead>
<tr>
<th></th>
<th>$p_A$ in cases</th>
<th>$p_U$ in controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F = 1$</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$F \neq 1$</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
**Methods: Genomic Control, GC**


- **Rationale:** Admixture increases association chi-square (allele test, 1 df) for all SNPs by a factor, \( \lambda \) (lambda).
- **Example:** 11-df test for SNP allele frequency differences among geographical regions (Wellcome Trust Case-Control Consortium [2007] *Nature* 447:661)
- **Q-Q plot (quantile-quantile plot) shows consistent allele frequency differences**

---

**Methods: GC at work**


- **Assumption:** At most 50% of smallest test statistics show inflation only due to population stratification, not due to genetic effects
- **Approach:** Define genomic control factor, \( \lambda = \frac{m_{\text{obs}}}{m_{\text{exp}}} \); \( m_{\text{obs}} \) = median of observed distribution of test statistic (chi-square for allelic association, 1 df); \( m_{\text{exp}} = 0.455 \) = median of 1-df chi-square distribution
- **Correction for admixture:** For each SNP, work with \( \chi^2/\lambda \).
- “Even when population substructure is present, our simulations agree with the theoretical calculations in Devlín and Roeder (1999) and suggest that, for a welldesigned study, the effect of population substructure is not sizable”
“Inflation” of HWE chi-square
Sing & Rothman (1975) Ann Hum Genet 39:141

- Inflation of HWE chi-square test due to correlation among study individuals. Quote:
  
  Deviation from the Hardy-Weinberg proportion of each genotype frequency may be attributable to a number of forces (selection, inbreeding, mutation, migration, subdivision).

- Define $\rho$ = correlation among any two individuals
- $E(\chi^2/\rho) = df/\beta$, $\beta = 1/[1 + \rho(n - 1)]$.
- Recommendation: Use $\beta \chi^2$ as test statistic
- GC became very well known, not so Sing & Rothman (1975). Why not?

Methods: STRUCTURE

- Method for clustering into subpopulations, each in HWE, and analyzing each subpopulation.
  
  1) "Applications of our method include demonstrating the presence of population structure, assigning individuals to populations, studying hybrid zones, and identifying migrants and admixed individuals."
  
  2) "Our method uses a set of unlinked genetic markers to infer details of population structure, and to estimate the ancestry of sampled individuals, before using this information to test for associations within subpopulations."

- Rather simple clustering algorithm. Many papers. STRUCTURE program available.
- Combine results from subpopulations, e.g. via p-values by Fisher method: $c = -2\ln(p) \sim \chi^2 \cdot 2df$
Methods: STRUCTURE

- Human population structure based on 377 microsatellites in 1056 individuals from 52 populations

Methods: EIGENSTRAT

- First, we apply principal components analysis to genotype data (codes 0, 1, 2) to infer continuous axes of genetic variation. These often have a geographic interpretation: For example, an axis describing a northwest-southeast cline in Europe would have values that gradually range from positive for samples from northwest Europe to negative in southeast Europe.
Methods: EIGENSTRAT

- *Second*, we continuously adjust genotypes and phenotypes by amounts attributable to ancestry along each axis, via computing residuals of linear regressions; intuitively, this creates a virtual set of matched cases and controls.
- *Third*, we compute association statistics using ancestry-adjusted genotypes and phenotypes. This procedure is equivalent to using the axes of variation as covariates in a multilinear regression, but is simpler because the axes of variation are orthogonal, and thus the adjustments can be performed independently for each axis of variation.
- **Practice:** Initially, remove outliers. Use all markers, even when in LD
- Cryptic relatedness: Detect related individuals and select only unrelated individuals for analysis.
- Additional covariates: If correlated with each other or with ancestry, use multivariate methods.

EIGENSTRAT applied to China

(A) 1708 Han Chinese and 767 non-Han Chinese individuals representing all samples studied.
(b) 1708 Han Chinese and 89 Japanese (JP) individuals (excluding non-East Asian samples).
Methods: AIMS markers

- Situation: Studies not using large numbers of SNPs (candidate genes, replications) cannot use EIGENSTRAT approach.
- Developed set of 300 highly informative SNPs for ancestry detection in European Americans.
- “Inferred genetic ancestry is more nuanced and informative than self-reported ancestry with regard to genetic similarity, particularly for individuals who may descend from multiple ancestral populations.”
- Grandparental origin...
- AIMS markers developed as early as 2001 (P-Y Kwok)

---

Methods: AIMS markers

- MS = multiple sclerosis
- BD = bipolar disorder
- PD = Parkinson’s disease
- IBD = inflammatory bowel disease
Methods: MDS components

- From SYSTAT manual:
  Multidimensional scaling (MDS) is a procedure for fitting a set of points in a space such that the distances between points correspond as closely as possible to a given set of dissimilarities between a set of objects. Dissimilarities may be measured directly, as in psychological judgments, or derived indirectly as in correlation matrices computed on rectangular data.

- Implemented in *plink* program
- Based on pairwise IBS, e.g.
  \((Aa, Aa) \rightarrow IBS = 2; (AA, aa) \rightarrow IBS = 0\)
- Using Euclidean distance → MDS same as PC analysis

---

Methods: MDS components

- The International Schizophrenia Consortium; 3,322 cases
  3,587 controls

*Figure S1*: Multidimensional scaling (MDS) plot for the individuals in the final post-QC dataset (both cases and controls). Known study samples are indicated by colour; the distinct clusters are labeled with the exception of the four British Isles samples (from Scotland, Ireland and England) that show near complete overlap on the first two dimensions.
Methods: Randomization test

“Our approach ... conditions on the genotype matrix and thus takes into account not only the population structure but also the complex linkage disequilibrium structure of the genome. As we show in simulation experiments, our method achieves higher power and significantly better control over false-positive rates than do existing methods. In addition, it can be easily applied to whole-genome association studies.”

- Complicated. Shown here for completeness.
- Cited in 24 papers
Genetic Heterogeneity in Human Disease

Abstract

“Strong evidence suggests that rare mutations of severe effect are responsible for a substantial portion of complex human disease. Evolutionary forces generate vast genetic heterogeneity in human illness by introducing many new variants in each generation. Current sequencing technologies offer the possibility of finding rare disease-causing mutations and the genes that harbor them.”

(A) The oldest human alleles originated in Africa well before the diasporas of modern humans 50,000–60,000 years ago. These oldest alleles are common in all populations worldwide. Approximately 90% of the variability in allele frequencies is of this sort. (Figure adapted from Cavalli-Sforza and Feldman, 2003.)
Genetic Heterogeneity in Human Disease

(B) Origins of common and rare alleles. KYA refers to “thousand years ago.” Horizontal arrows suggest continuing cross-migration between continental populations. Development of agriculture in the past 10,000 years and of urbanization and industrialization in the past 700 years has led to rapid populations growth and therefore to the appearance of vast numbers of new alleles, each individually rare and specific to one population or even to one family. (Figure adapted from Tishkoff and Verrelli, 2003.)

Meeting abstracts
http://www.ashg.org/2013meeting/pages/abstracts.shtml

Search for “heterogeneity”

115

Hypertriglyceridemia (HTG) is a risk factor for cardiovascular disease, a leading cause of death in the U.S. There are few pharmacological treatments, with non-compliance rates >5% due to side effects. Investigating the genetics of HTG may lead to new drug targets. There are ~35 known SNPs that explain only ~10% of variation in triglycerides (TG). Due to the genetic heterogeneity of HTG, a family study design is optimal for identification of novel loci with large effect size as the same mutation can be seen in many relatives and co-segregation with TG observed.
Meeting abstracts
http://2013.ispg.net/

**PSYCHIATRIC GENOMICS CONSORTIUM QUADRUPLES SCHIZOPHRENIA GWAS SAMPLE-SIZE TO 35,000 CASES AND 47,000 CONTROLS**

Stephan Ripke, M.D.¹, Schizophrenia Working Group²

¹The Broad Institute; Massachusetts General Hospital, ²Psychiatric Genomics Consortium

**Background:** The PGC (Psychiatric Genomics Consortium) is an international group of researchers whose major aim is to maximize the utility of extant psychiatric GWAS through mega-analysis. In a previous study, our first wave of genome-wide schizophrenia association analysis identified multiple loci involved in this genetically complex and clinically heterogeneous disorder (Nature Genetics, 2011). While around 20,000 individuals were necessary to achieve this result, detailed analysis of the data suggested that there are many more genes to discover, and that this should be possible by further increase of sample size.

**Methods:** Here we present an update of this international endeavor, which now comprises 35,476 schizophrenia cases and 46,839 controls coming from 52 sub-studies. The presented data is imputed into 1000 Genomes (Aug, 2012) and analyzed using standard logistic regression with ancestry components as covariates. All index SNPs with a p-value smaller than $1 \times 10^{-5}$ were used for replication lookup in an independent GWAS analysis with 1,500 cases and 66,000 controls.
LARGE-SCALE RNA-SEQUENCING OF SCHIZOPHRENIA BRAINS BY THE COMMONMIND CONSORTIUM

Menachem Fromer, Ph.D.¹, for the CommonMind Consortium (CMC)²

Background: Schizophrenia is a severe psychiatric disease with ~1% lifetime prevalence in the general population. And, although it is a highly heritable disease, a picture of the genetics is only now emerging from large-scale studies (of both common and rare variation) but is nowhere near complete. That is, the exact genes and pathways involved are still largely unknown, and no individual variant explains even a moderate fraction of the total genetic variance. Thus, two things have become increasingly clear: i) large studies are needed due to disease heterogeneity and polygenicity, and ii) study of any single level of biology (genetics, expression, imaging, etc.) will be insufficient to fully unravel the biology of disease, but rather information must be collected (and then combined) across multiple levels.