Human Genetics and Gene Mapping of Complex Traits

Advanced Genetics, Spring 2021
Human Genetics Series
Tuesday 4/20/21
Nancy L. Saccone
Dept of Genetics
nlims@genetics.wustl.edu
Outline and learning objectives remaining from last lecture

• Quick follow-up on LOD scores

• Logistic regression (for dichotomous outcomes – case/control)

• More on multiple test correction
Outline continued

• Linkage disequilibrium (LD)
  • Measures of LD, how to compute
  • Practical implications

• GWAS to "Post-GWAS"
  • Meta-analyses, consortia (imputation)
  • Resources (e.g. available GWAS data, results)
  • Diverse populations
  • Getting from statistical signal to biology
  • Polygenic risk scores (PRS)
  • Translating to clinical care – “precision medicine”
Note on linkage analysis

The "traditional" LOD score threshold for significance is 3. This corresponds to

\[
3 = \log_{10} \left( \frac{L(\theta = \hat{\theta})}{L(\theta = 1/2)} \right)
\]

i.e.

\[
\frac{L(\theta = \hat{\theta})}{L(\theta = 1/2)} = 1000
\]

(This is different from a p-value. The p-value corresponding to LOD=3 is 0.0001.)
Recall: linear regression model

\[ y = \alpha + \beta x \]

\( \beta \) = slope of Fitted line

\( \alpha \) = y-intercept

y-axis: quantitative trait value

x-axis: number of alleles

\( \beta \) is a measure of the EFFECT SIZE for the SNP

Alternative: \( r^2 \) = squared correlation coefficient, indicates proportion of phenotypic variance in \( y \) that’s explained by \( x \)
Logistic regression summary

Let $y = 1$ if case, 0 if control (2 values)

Let $P = \text{probability that } y = 1 \text{ (case)}, \text{ranges from 0 to 1}$

Then $\text{logit}(P)$ ranges from $-\infty$ to $\infty$

\[
\text{logit}(P) = \ln\left(\frac{P}{1-P}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n
\]

$x_1 =$ SNP genotype (# of copies of the “coded” allele)

Odds ratio measures “effect size”: $e^{\beta_1}$, involves $\beta_1$ again!

The $p$-value is generated from a likelihood ratio-based test (comparing models with and without the SNP term)

Remember: $p$-value is probability of seeing data “as or more extreme as you observed” when the null hypothesis holds.
Logistic regression for dichotomous traits

Let $y = 1$ if case, 0 if control (2 values)

Let $P = \text{probability that } y = 1$ (case)

Let $x_1 = \text{genotype (additive coding)}$

\[
\text{logit}(P) = \ln\left(\frac{P}{1-P}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n
\]

Why?
Logit function

• Usual regression expects a dependent variable that can take on any value, $(-\infty, \infty)$
• A probability $P$ is in $[0,1]$, so not a good dependent variable
• Odds = $p/(1-p)$ is in $[0,\infty)$
• Logit = $\ln(\text{odds})$ is in $(-\infty, \infty)$
Think of the shapes of the graphs

- $y = \frac{x}{1-x}$  (x in place of P)

As x varies from 0 to 1, y varies from 0 to $\infty$

- $y=\ln(x)$
  
  varies from $-\infty$ to $\infty$
  
  (as input varies from 0 to $\infty$)
Logistic regression

Let \( y = 1 \) if case, 0 if control (2 possible values)

Let \( P = \) probability that \( y = 1 \) (case)

\[
\logit(P) = \ln\left( \frac{P}{1-P} \right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n = \Omega
\]

Note that can exponentiate both sides to get odds = \( \frac{P}{1-P} \):

\[
\text{Odds} = \left( \frac{P}{1-P} \right) = e^{\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n} = e^\Omega
\]

What about the “effect size”?
It’s the “odds ratio”, and it is still related to \( \beta_1 \)!
Odds ratio

- The number $e (=2.718\ldots)$ is the base of natural logarithms
- $e^0 = 1$
- $e^{\beta_1}$ is the odds ratio; if $\beta_1=0$ then odds ratio is 1 (no effect of the SNP)
To get odds ratio per copy of the allele ("effect size")

- Full model: \[
\frac{P}{1 - P} = e^{[\alpha + \beta_1 x_1 + \ldots + \beta_n x_n]}
\]

- Odds when \( x_1 = 1 \) (1 copy of the allele)
  \[
P_1 / (1 - P_1) = e^{[\alpha + \beta_1 x_1 + \ldots + \beta_n x_n]} \bigg|_{x_1=1} = e^{[\alpha_0 + \beta_2 x_2 + \ldots + \beta_n x_n]} + \beta_1
\]

- Odds when \( x_1 = 0 \) (0 copies of the allele)
  \[
P_0 / (1 - P_0) = e^{[\alpha + \beta_1 x_1 + \ldots + \beta_n x_n]} \bigg|_{x_1=0} = e^{[\alpha_0 + \beta_2 x_2 + \ldots + \beta_n x_n]}
\]

- Odds Ratio:
  \[
\left( \frac{P_1 / (1 - P_1)}{P_0 / (1 - P_0)} \right) = \frac{e^{[\alpha + \beta_2 x_2 + \ldots + \beta_n x_n]} + \beta_1}{e^{[\alpha + \beta_2 x_2 + \ldots + \beta_n x_n]}} = e^{\beta_1}
\]
Logistic regression summary

Let \( y = 1 \) if case, 0 if control (2 values)

Let \( P = \) probability that \( y = 1 \) (case), ranges from 0 to 1

Then \( \text{logit}(P) \) ranges from \(-\infty\) to \(\infty\)

\[
\text{logit}(P) = \ln\left( \frac{P}{1 - P} \right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n
\]

Odds ratio

\[
e^{\beta_1}
\]

Similar to case of linear regression, can compute an analog to “variance explained,” usually also called \( r^2 \)
Multiple test correction

P-value: probability of observing the data or more "extreme" data when the null hypothesis holds = $\Pr(\text{as or more extreme} \mid \text{null})$

For a single test, traditional required threshold is $p$-value $\leq 0.05$

**Bonferroni correction**

- If $n =$ number of independent tests, then under Bonferroni correction, $p$-value threshold for "significance" is $0.05/n$, or more generally, require $P_{\text{singletest}} \leq P_{\text{experimentwide}} / n$

In the GWAS setting – how to determine the number of "independent tests"?

- Account for linkage disequilibrium (LD), or correlation between alleles (more on this later)
- Typical threshold corrects for 1 million independent tests, requiring $p \leq 5 \times 10^{-8}$
Genome-wide Association Studies (GWAS)

Successful by several metrics: have led to

- **Replicable** genetic variants underlying complex diseases
- Novel genes, pathways, biology
- Meaningful functional followup and clinical benefit
  - Recall Jim Skeath’s lecture: fetal hemoglobin, BCL11A and CRISPR-Cas9 for sickle cell & thalessemia

Less successful by other metrics:
- "Top" associated SNPs explain limited phenotypic variance (typical odds ratios ~ 1.3, variance explained ~ 1%)

Additional good news:
- Polygenic risk scores explain more phenotypic variance
Genome-wide Association Studies (GWAS)

GWASes rely on linkage disequilibrium (LD) to "tag" variation, and thus must be interpreted in the context of LD: The signal SNP may not be the biologically causal SNP.
Outline and learning objectives

• Linkage disequilibrium (LD)
  • Measures of LD, how to compute
  • Sources of LD
  • Practical implications

• GWAS to "Post-GWAS"
  • Resources (e.g. available GWAS data, results)
  • GWAS meta-analyses, consortia
  • Diverse populations
  • Polygenic risk scores (PRS)
  • Translating to clinical care
  • Getting from statistical signal to biology
Interpretation of GWAS results must account for LD

- Suppose a SNP is significantly associated with a disease
- Other SNPs correlated (high $r^2$) with that SNP are additional, potentially “causative” variants

Example: GWAS of fetal hemoglobin detected $BCL11A$

Some (not all!) of these SNPs are highly correlated with the lead SNP rs11886868 (intronic C $>$ T)

Uda et al. (2008), Genome-wide association study shows BCL11A… , PNAS 105:1620-1625
Interpretation of GWAS results must account for LD

Strongest signals are in intron 2 of BCL11A, point to group of SNPs

Uda et al. (2008), Genome-wide association study shows BCL11A..., PNAS 105:1620-1625
Example: \textit{CHRNA5-CHRNA3-CHRNB4} and rs16969968
Associated with nicotine dependence, smoking, lung cancer, COPD.

\textbf{CHRNA5-CHRNA3-CHRNB4}

rs16969968
Saccone SF et al., 2007
Example: *CHRNA5-CHRNA3-CHRNB4* and rs16969968
Associated with nicotine dependence, smoking, lung cancer, COPD.

**CHRNA5-CHRNA3-CHRNB4**

**rs16969968**
Saccone SF et al., 2007  
Bierut et al., 2008  
Stevens et al., 2008  
Sherva et al., 2008  
Chen et al., 2009  
Weiss et al., 2009  
Liu et al., 2008  
Young et al., 2008

**rs1317286**
Berrettini et al., 2008

**rs1051730**
Saccone SF et al., 2007  
Thorgeirsson et al., 2008  
Caporaso et al., 2009  
Hung et al., 2008  
Amos et al., 2008  
Pillai et al., 2009

Association signal extends across multiple genes!
Example: \textit{CHRNA5-CHRNA3-CHRNB4} and rs16969968
Associated with nicotine dependence, smoking, lung cancer, COPD.

A second independent "bin" of correlated SNPs represents a distinct statistical signal
A typical GWAS signal consists of multiple SNPs due to LD

- Often multiple distinct statistical signals exist
- "Conditional analysis" is used to identify distinct signals:
  - In the regression, include signal SNP as a covariate, observe if additional SNPs remain significant in the model
LD and Human Sequence Variation

ancestral chromosome

present day chromosomes:

alleles on the preserved "ancestral background" tend to be in linkage disequilibrium (LD)
Linkage Disequilibrium

• "Non-random" associations between alleles at different loci

• How to formally measure LD between alleles at 2 loci?
To measure LD between alleles at 2 biallelic loci

Locus A | Locus B
---------|---------
A₁, A₂   | B₁, B₂

Given 2N haplotypes:
Haplotype freq for AᵢBⱼ is

\[ h_{ij} = \frac{n_{ij}}{2N} \]

Compare \( h_{ij} \) to the frequency expected under no association:

\[ p_{A_1} p_{B_1} = \left( \frac{n_{11} + n_{12}}{2N} \right) \left( \frac{n_{11} + n_{21}}{2N} \right) \]

Define the disequilibrium coefficient:

\[ D = h_{11} - p_{A_1} p_{B_1} \]
Common LD measures

Disequilibrium coefficient:
\[ D = h_{11} - p_{A1} p_{B1} \]

Normalized disequilibrium coefficient:
\[ D' = \frac{D}{|D|_{\text{max}}} \], where

\[ |D|_{\text{max}} = \left\{ \begin{array}{ll} 
\min(p_{A1}p_{B2}, p_{A2}p_{B1}) & \text{if } D > 0 \\
\min(p_{A1}p_{B1}, p_{A2}p_{B2}) & \text{if } D < 0
\end{array} \right. \]

Range of \( D' \) is \([-1,1]\)

Squared correlation coefficient:
\[ r^2 = \frac{D^2}{(p_{A1}p_{A2}p_{B1}p_{B2})} \]
Notes:

1. \( D = h_{11} - p_{A1}p_{B1} = h_{22} - p_{A2}p_{B2} \)

2. Choice of allele labeling may affect sign but not absolute value of \( D \).
Measuring LD

Example:

<table>
<thead>
<tr>
<th>A₁</th>
<th>B₁</th>
<th>B₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0</td>
<td>50</td>
</tr>
<tr>
<td>0</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>100</td>
</tr>
</tbody>
</table>

Only observe 2 haplotypes: A₁B₁ and A₂B₂

\[ D = h_{11} - p_{A1}p_{B1} = (0.5) - (0.5)(0.5) = 0.5 - 0.25 = 0.25 \]

\[ D_{\text{max}} = \min(p_{A1}p_{B2}, p_{A2}p_{B1}) = \min(0.25, 0.25) = 0.25 \]

\[ |D'| = 1, r^2 = 1 \]
Measuring LD

Example:

Only observe 2 haplotypes: $A_1B_1$ and $A_2B_2$

To measure significance: $\chi^2$ (1 df):

$$\chi^2 = \frac{(50 - 25)^2}{25} + \frac{(0 - 25)^2}{25} + \frac{(0 - 25)^2}{25} + \frac{(50 - 25)^2}{25}$$

$$\chi^2 = \frac{(50 - 25)^2}{25} + \frac{(0 - 25)^2}{25} + \frac{(0 - 25)^2}{25} + \frac{(50 - 25)^2}{25}$$

Chi-sqr = 100, p-value very small
LD measures

Another useful example:

<table>
<thead>
<tr>
<th></th>
<th>B₁</th>
<th>B₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₁</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>90</td>
</tr>
<tr>
<td>A₂</td>
<td>10</td>
<td>90</td>
</tr>
</tbody>
</table>

\[
D' = \frac{(.1 - (.1)(.1))}{(.09)} = 1
\]

\[
\chi^2 = 100, \ p\text{-value} \sim 0.0
\]

\[
r^2 = 1
\]

<table>
<thead>
<tr>
<th></th>
<th>B₁</th>
<th>B₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₁</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>80</td>
</tr>
<tr>
<td>A₂</td>
<td>10</td>
<td>90</td>
</tr>
</tbody>
</table>

\[
D' = \frac{(0 - (.1)(.1))}{(.01)} = -1
\]

\[
\chi^2 = 1.23, \ p\text{-value} = 0.27
\]

\[
r^2 = 0.012
\]
LD measures

|D'| is 1 when the alleles of the two markers are as correlated as they can be, given the allele frequencies of the co-occurring alleles.

The range of $r^2$ depends on the marker allele frequencies.

$r^2$ equals 1 if and only if 1) the MAFs at the two loci match and 2) the minor alleles always co-occur.

D': useful for identifying regions of reduced recombination.

$r^2$: useful for identifying markers that are good predictors of allelic status at other markers.
LD across TCF7L2 in CEU HapMap.

Grant et al., Nat Genet 2006, Figure 1
Where does LD come from?

- Potential sources of LD:
  1. Genetic linkage between loci
  2. Random drift
  3. Founder effect
  4. Mutation
  5. Selection
  6. Population admixture / stratification
LD is not a simple monotonic function of distance
Dawson et al., Nature 2002

Panel (a): D' by distance between markers
Panel (b): r² by distance between markers
LD patterns can vary by population: differing population history, allele frequencies

Uda et al. (2008), Genome-wide association study shows BCL11A... , PNAS 105:1620-1625
The 1000 Genomes Project catalogs variants and LD structure
www.1000genomes.org
http://www.internationalgenome.org/

(Previously: The Haplotype Mapping (HapMap) Project)

Browsable via www.ensembl.org
Be aware of LD in **design** and in **interpretation**

GWAS SNP arrays rely on LD to “tag” common variation

A popular LD-based "tagging" SNP approach:
- "r² bin tags" (Carlson et al., 2004): greedy algorithm that identifies bins of SNPs such that at least one member of each bin has $r^2 > T$ (threshold, e.g. 0.8) with all bin members.
  - Bin members are not necessarily contiguous

Similar method: “haplotype tagging” – next slide

Imputation of untyped SNPs is possible due to LD and haplotype structure
Using LD in study design

The International HapMap Project, Nature 2003
Where does LD come from?

Potential sources of LD:

1. Genetic linkage between loci
2. Random drift
3. Founder effect
4. Mutation
5. Selection
6. Population admixture / stratification
An important concern: population stratification

Spurious association between a locus and disease can occur if there are two (unknown) subpopulations.

Exaggerated example: if an allele occurs only in stratum 1, then any trait with very different prevalence in stratum 1 versus stratum 2 could appear to be associated with this SNP.

Subpop1 (i.e. stratum 1): allele A1
Subpop2 (i.e. stratum 2): alleles A1, A2

unshaded = affected cases. Unaffected controls are more prevalent in stratum 1. Here allele A₁ appears to have much higher freq in controls compared to cases
An example of spurious association due to admixture/stratification:

<table>
<thead>
<tr>
<th></th>
<th>B₁</th>
<th>B₂</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>population 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A₁</td>
<td>9</td>
<td>1</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A₂</td>
<td>81</td>
<td>9</td>
<td>90</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>90</td>
<td>10</td>
<td>100</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

chi-square = 0

<table>
<thead>
<tr>
<th></th>
<th>B₁</th>
<th>B₂</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>population 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B₁</td>
<td>25</td>
<td>25</td>
<td>50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B₂</td>
<td>25</td>
<td>25</td>
<td>50</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>50</td>
<td>100</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

chi-square = 0

<table>
<thead>
<tr>
<th></th>
<th>B₁</th>
<th>B₂</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Combined</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A₁</td>
<td>34</td>
<td>26</td>
<td>60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A₂</td>
<td>106</td>
<td>34</td>
<td>140</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>140</td>
<td>60</td>
<td>200</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

chi-square = 7.26
p-value = 0.007
How to protect against population stratification?

1. Match geographic/ancestry background of the case and control samples. (at the very least!)

2. Check to make sure there is no underlying stratification
   - Can use GWAS data
   - If study is not GWAS, would need a set of “ancestry-informative markers (AIMs)”

3. If there is, account for this in the analysis

4. Software:
   - EIGENSTRAT software (principal components) (Price et al.)
   - STRUCTURE software (Pritchard et al.)
   - Devlin, Roeder, Bacanu "genomic control"
Next steps (post-GWAS)?

1. Narrow down to "true" (biologically causal) variants in the associated region
2. Determine how/why these variants alter disease risk
3. Translate to *clinical care, outcomes* (next lec)

How to do 1 and 2?

- Look across multiple diverse populations, leverage LD differences
- Bioinformatic prioritization
- Imputation, genotyping, sequencing to query the regions of association more thoroughly (1000 Genomes Project)
- Meta-analysis for larger, more powerful samples
- Functional follow-up
- Effects in model organisms
**CHRNA5-CHRNA3-CHRNB4** region: the $r^2 \geq 0.8$ bin.
Associated with nicotine dependence, smoking, lung cancer, cocaine dependence.

**Conclusive evidence in European-ancestry populations**

- **rs16969968**
  - Saccone SF et al., 2007
  - Bierut et al., 2008
  - Liu et al., 2008
  - Saccone NL et al., 2010

- **rs8034191**
  - Hung et al., 2008
  - Amos et al., 2008

- **rs1317286**
  - Berrettini et al., 2008

- **rs1051730**
  - Thorgeirsson et al., 2008
  - Hung et al., 2008
  - Amos et al., 2008

Blue: nicotine dependence / smoking
Red: lung cancer
**CHRNA5-CHRNA3-CHRNB4 region:** the $r^2 > 0.8$ bin. Associated with nicotine dependence, smoking, lung cancer

In HapMap YRI, there are only 2 non-trivial $r^2$ bins

The other SNPs are singleton bins!

Opportunity to narrow down the signal
Combining data through sharing/collaboration

For example, meta-analysis and/or combined ("mega") analysis

Benefits:
- Improved power
- Extends value of existing data (often costly to collect)

Challenges:
1. Harmonizing phenotypes
2. Harmonizing genotypes: genetic imputation
- Meta-analysis: statistically combines summary statistics across multiple datasets

- Thus meta-analysis can be applied to published data/results

- Collaborative meta-analyses goes further: not just retrospective literature review, but carrying out new, coordinated analyses across multiple datasets
  - Not limited to the "published analyses" for a given dataset
  - Can include unpublished datasets
Example: Targeted smoking meta-analysis (rs16969968)

Summary

<table>
<thead>
<tr>
<th>Study</th>
<th>OR</th>
<th>95% C.I.</th>
<th>Case/Control</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.33</td>
<td>(1.26-1.39)</td>
<td></td>
</tr>
</tbody>
</table>

P-value: $5.96 \times 10^{-31}$

Saccone et al., PLoS Genetics 2010
Smoking GWAS meta-analysis

N = 38,181 subjects

Tobacco and Genetics Consortium, Nat Genet 2010
Human genetic resources, cohort databases, opportunities

- dbGaP
- UK Biobank (UKB)
- Million Veteran Program (MVP)
- All of Us
- Consumer genomics companies, e.g. 23andMe
- Electronic Health Records (EHR) based research