Human Genetics and Gene Mapping of Complex Traits

Advanced Genetics, Spring 2018
Human Genetics Series
Thursday 4/5/18
Nancy L. Saccone, Ph.D.
Dept of Genetics
nlims@genetics.wustl.edu / 314-747-3263
What is different about Human Genetics
(recall from Cristina Strong's lectures)

• Imprinting – uniquely mammalian

• Trinucleotide repeat diseases – "anticipation"

• Can study complex behaviors and cognition (neurogenetics)

• Extensive sequence variation leads to common/complex disease
  1. Common disease, common variant hypothesis
  2. Large # of small-effect variants
  3. Large # of large-effect rare variants
  4. Combo of genotypic, environmental, epigenetic interactions

Greg Gibson, Nature Rev Genet 2012
Mapping disease genes

- **Linkage**
  - quantify co-segregation of trait and genotype in families

  ![Linkage Diagram]

  LOD score traditionally used to measure statistical evidence for linkage

- **Association**
  - Common design: case-control sample, analyzed for allele frequency differences

  ![Association Diagram]

  cases

  controls
## Comparing Linkage and Association

<table>
<thead>
<tr>
<th>Linkage mapping:</th>
<th>Association mapping:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Requires family data</td>
<td>Unrelated cases/controls OR Case/parents OR family design</td>
</tr>
<tr>
<td>Disease travels with marker allele within families</td>
<td>Disease is associated with marker allele that may be</td>
</tr>
<tr>
<td>(close genetic distance between disease locus and</td>
<td>either causative or in linkage disequilibrium with causal</td>
</tr>
<tr>
<td>marker)</td>
<td>variant</td>
</tr>
<tr>
<td>Relationship between same allele and trait need not</td>
<td>Works only if association exists at the population level</td>
</tr>
<tr>
<td>exist across the full sample (e.g. across different</td>
<td></td>
</tr>
<tr>
<td>families)</td>
<td></td>
</tr>
<tr>
<td>robust to allelic heterogeneity: if different</td>
<td>not robust to allelic heterogeneity</td>
</tr>
<tr>
<td>mutations occur within the same gene/locus, the</td>
<td></td>
</tr>
<tr>
<td>method works</td>
<td></td>
</tr>
<tr>
<td>signals for complex traits tend to be broad (~20 Mb)</td>
<td>association signals generally not as broad</td>
</tr>
</tbody>
</table>
Human DNA sequence variation

- Single nucleotide polymorphisms (SNPs)

Strand 1: A A C C A T A T C ... C G A T T T ...
Strand 2: A A C C A T A T C ... C A A T T T ...
Strand 3: A A C C C T A T C ... C G A T T T ...

- Provide biallelic markers
- Coding SNPs may directly affect protein products of genes
- Non-coding SNPs still may affect gene regulation or expression
- Low-error, high-throughput technology
- Common in genome
Number of SNPs in dbSNP over time
solid: cumulative # of non-redundant SNPs.
dotted: validated. dashed: double-hit

from: The Intl HapMap Consortium
Number of SNPs in dbSNP over time

dotted: validated

From: Fernald et al., Bioinformatics challenges for personalized medicine, Bioinformatics 27:1741-1748 2011
Number of SNPs over time

Questions that we can answer with SNPs:

• Which genetic loci influence risk for common human diseases/traits? (Disease gene mapping studies, including GWAS – genome-wide association studies)

• Which genetic loci influence efficacy/safety of drug therapies? (Pharmacogenetics)

• Population genetics questions
  • evidence of selection
  • identification of recombination hotspots
Part I: Human linkage studies

Need to track co-segregation of trait and markers (number of recombination events among observed meioses)

General “linkage screen" approach:

Recruit families

Genotype individuals at marker loci along the genome

If a marker locus is "near" the trait-influencing locus, the parental alleles from the same grandparent at these two loci "tend to be inherited together" (recombination between the two loci is rare)

\[ \theta = \text{the probability of recombination between 2 given loci} \]

Defn: max LOD score = \( \log_{10}[L(\theta = \hat{\theta}) / L(\theta = 1/2)] \)  
( \( \hat{\theta} \) is the maximum likelihood estimate of theta)
Example of autosomal dominant (fully penetrant, no phenocopies)
General hallmarks:
All affected have at least one affected parent, so the disease occurs in all generations above the latest observed case.
The disease does not appear in descendants of two unaffecteds.

Possible molecular explanation: disease allele codes for a functioning protein that causes harm/dysfunction.
Example of autosomal recessive (fully penetrant, no phenocopies).

General hallmarks:
Many/most affecteds have two unaffected parents, so the disease appears to skip generations.
On average, 1/4 of (carrier x carrier) offspring are affected.
(Affected x unaffected) offspring are usually unaffected (but carriers)
(Affected x affected) offspring are all affected.
Example of autosomal recessive (fully penetrant, no phenocopies).

Possible molecular explanation: disease allele codes for a nonfunctional protein or lack of a protein, and one copy of the wild-type allele produces enough protein for normal function.
**Classic models of disease**

Classical autosomal dominant inheritance (no phenocopies, fully penetrant).

Penetrance table:

<table>
<thead>
<tr>
<th>f++</th>
<th>f+_d</th>
<th>f_d d</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Often the dominant allele is rare, so that probability of homozygous dd individuals occurring is negligible.

Classical autosomal recessive inheritance (no phenocopies, fully penetrant).

Penetrance table:

<table>
<thead>
<tr>
<th>f++</th>
<th>f+_d</th>
<th>f_d d</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Genetic models of disease

Other examples of penetrance tables (locus-specific):

<table>
<thead>
<tr>
<th>$f_{++}$</th>
<th>$f_{+d}$</th>
<th>$f_{dd}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0.9</td>
</tr>
<tr>
<td>0.1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.1</td>
<td>0.8</td>
<td>0.8</td>
</tr>
</tbody>
</table>

Incomplete/reduced penetrance: when the risk genotype's effect on phenotype is not always expressed/observed. (e.g. due to environmental interaction, modifier genes)

Phenocopy: individual who develops the disease/phenotype in the absence of "the" risk genotype (e.g. through environmental effects, heterogeneity of genetic effects)
Part II: Genetic Association Testing

Typical statistical analysis models:

Quantitative continuous trait:
  linear regression

Dichotomous trait – e.g. case/control:
  logistic regression
  - more flexible than chi-square / Fisher’s exact test
  - can include covariates
  - provides estimate of odds ratio
Linear regression

Let $y = \text{quantitative trait value}$

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \text{error}$$

OR

$$\hat{y} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

$\hat{y}$ = predicted quantitative trait value

$x_1$ = SNP genotype (e.g. # copies of designated allele: 0,1,2)

$x_2, \ldots, x_n$ are covariate values (e.g. age, sex)

Null hypothesis $H_0: \beta_1 = 0$.

The SNP “effect size” is represented by $\beta_1$, the coefficient of $x_1$.

Is there significant evidence that $\beta_1$ is non-zero?
Least squares linear regression: general example

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

Fitted line,
Slope = \beta

Residual deviations

The least squares solution finds \( \alpha \) and \( \beta \) that minimize the sum of the squared residuals.
Least squares linear regression: general example

The least squares solution finds $\alpha$ and $\beta$ that minimize the sum of the squared residuals.

Fitted line,
Slope = $\beta$

Residual deviations

Would NOT minimize the sum of squared residuals

$y = \alpha + \beta x$
SNP Marker Additive Coding:

<table>
<thead>
<tr>
<th>Genotype</th>
<th>$x_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/1</td>
<td>0</td>
</tr>
<tr>
<td>1/2</td>
<td>1</td>
</tr>
<tr>
<td>2/2</td>
<td>2</td>
</tr>
</tbody>
</table>

Codes number of “2” alleles
Least squares linear regression

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

\( \beta \) = slope of Fitted line

x-axis: number of alleles
"Phenotypic variance explained"

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

- \( \alpha \): intercept of the fitted line
- \( \beta \): slope of the fitted line
- \( r^2 \): squared correlation coefficient

\( r^2 \) indicates the proportion of phenotypic variance in \( y \) that is explained by \( x \).

**x-axis:** number of alleles

\[ \beta = \text{slope of Fitted line} \]
Another use of linear regression: Traditional sib pair linkage analysis “Model-free / non-parametric”

- Idea: if two sibs are alike in phenotype, they should be alike in genotype near a trait-influencing locus.
- to measure "alike in genotype" : Identity by descent (IBD). Not the same as identity by state.

```
1 | 2
1 | 3
```

```
1 | 2
1 | 3
1 | 2
1 | 3
```

IBS=1
IBD=0

```
2 | 3
2 | 3
1 | 3
1 | 3
```

IBS=1
IBD=1
Sib pair linkage analysis of quantitative traits

- Haseman-Elston regression: Compare IBD sharing to the squared trait difference of each sib pair.

\[(\text{trait difference})^2\]
Example sib-pair based LOD score plot, from Saccone et al., 2000
Logistic regression for dichotomous traits

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

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

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

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

Why?
Logit function

• Usual regression expects a dependent variable that can take on any value, \((-\infty, \infty)\)

• A probability 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 $\infty$
Logistic regression

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

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

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...) 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
To get odds ratio per copy of the allele ("effect size")

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

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

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

- 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 $\logit(P)$ ranges from $-\infty$ to $\infty$

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

Odds ratio

$$e^{\beta_1}$$

Similar to case of linear regression, can compute an analog to “variance explained,” usually also called $r^2$
Published Genome-Wide Associations (p≤5×10⁻⁸)

http://www.ebi.ac.uk/gwas/diagram
This diagram shows all SNP-trait associations with p-value $\leq 5.0 \times 10^{-8}$, published in the GWAS Catalog.

SNP-trait associations for "Nervous system disease" on the diagram: 729
Displaying GWAS results

“Manhattan plot”

x-axis: chromosomal position
y-axis: $-\log_{10}(p$-value)

$p = 1 \times 10^{-8}$ is plotted at $y=8$,
$p = 5 \times 10^{-8}$ is plotted at $y = 7.3$

“Q-Q plot”

Quantile-Quantile plot

Idea: Rank tested SNPs by association evidence;
compare number of observed vs expected associations under the null at a given significance level

Bloom et al., Ann Am Thorac Soc 2014
Displaying GWAS results

– “Q-Q plot”: Quantile-quantile plot
  Idea: Rank tested SNPs by association evidence; compare number of observed vs expected associations under the null at a given significance level

Helps detect systematic bias in data:
  Most datapoints should be close to the y=x line
  Exception: signals (lowest, most significant p-values)
Displaying GWAS results

— “Q-Q plot”: Quantile-quantile plot

Before-and-after adjustment for population stratification

$\chi^2$ Statistics
- Unadjusted
- Adjusted

GWAS

Successful by several metrics:
• Identifying genetic variants underlying complex diseases
• Highlighting novel genes, pathways, biology
• Motivating functional followup, collaborative meta-analyses

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

But even by that metric, there's good news:
• As a whole, the variation assayed by GWAS may be able to explain even more of the phenotypic variance (work of Peter Visscher et al.)

GWASes rely on linkage disequilibrium (LD) to "tag" variation, and thus must be interpreted in the context of LD: the signal SNP may be different from the causal SNP.
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: \textit{CHRNA5-CHRNA3-CHRNB4} on chromosome 15q25

\textbf{Nicotinic receptor gene cluster}

rs16969968

D398N in \textit{CHRNA5}

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

**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
- Amos et al., 2008
- Pillai et al., 2009

Also others
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

- Potential sources of LD:
  1. Genetic linkage between loci
  2. Random drift
  3. Founder effect
  4. Mutation
  5. Selection
  6. Population admixture / stratification
Linkage Disequilibrium (LD) involves haplotype frequencies.

Focus on pair-wise LD, SNP markers

Genotypes do not necessarily determine haplotypes:
Consider 2-locus genotype $A_1 A_2 B_1 B_2$.
Two possible phases:
Linkage Disequilibrium (LD) involves haplotype frequencies

Focus on pair-wise LD, SNP markers

Genotypes do not necessarily determine haplotypes:
Consider 2-locus genotype $A_1 A_2 B_1 B_2$.

Two possible phases:

$A_1 \parallel A_2$
$B_1 \parallel B_2$

$A_1 \parallel A_2$
$B_2 \parallel B_1$
Linkage Disequilibrium

Linkage Disequilibrium (LD), aka allelic association:

For two loci A and B:
LD is said to exist when alleles at A and B tend to co-occur on haplotypes in proportions different than would be expected under statistical independence.
Linkage Disequilibrium

Example: Consider 2 SNPs:

SNP 1:      A 50%  C 50%
SNP 2:      A 50%  G 50%

<table>
<thead>
<tr>
<th></th>
<th>snp1</th>
<th>snp2</th>
<th>expected freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 possible haplotypes:</td>
<td>A</td>
<td>A</td>
<td>0.5 * 0.5</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>G</td>
<td>0.5 * 0.5</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>A</td>
<td>0.5 * 0.5</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>G</td>
<td>0.5 * 0.5</td>
</tr>
</tbody>
</table>

But perhaps in your sample you observe only the following:

A A C C A T A T C ... C G A T T ...

and

A A C C C T A T C ... C A A T T ...
Linkage Disequilibrium

• 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:

\[ p_{A₁}p_{B₁} = \left( \frac{n_{11} + n_{12}}{2N} \right) \left( \frac{n_{11} + n_{21}}{2N} \right) \]

Define the disequilibrium coefficient:

\[ D = h_{11} - p_{A₁}p_{B₁} \]
Common LD measures

Disequilibrium coefficient:
\[ D = h_{11} - p_A p_B \]

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

\[ |D|_{\text{max}} = \begin{cases} 
\min(p_A p_B, p_A p_B) & \text{if } D > 0 \\
\min(p_A p_B, p_A p_B) & \text{if } D < 0 
\end{cases} \]

Range of D' is [-1,1]

Squared correlation coefficient:
\[ r^2 = \frac{D^2}{(p_A p_A p_B p_B)} \]
Measuring LD

Example:

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

<table>
<thead>
<tr>
<th></th>
<th>$B_1$</th>
<th>$B_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1$</td>
<td>50</td>
<td>0</td>
</tr>
<tr>
<td>$A_2$</td>
<td>0</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>50</td>
</tr>
</tbody>
</table>

$D = h_{11} - p_{A_1}p_{B_1} = (0.5) - (0.5)(0.5) = 0.5 - 0.25 = 0.25$

$D_{\text{max}} = \min(p_{A_1}p_{B_2}, p_{A_2}p_{B_1}) = \min(0.25, 0.25) = 0.25$

$|D'| = 1$, $r^2 = 1$
Measuring LD

Example:

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A₁</td>
<td>50</td>
<td>0</td>
<td>50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A₂</td>
<td>0</td>
<td>50</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>

Only observe 2 haplotypes: A₁B₁ and A₂B₂

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-sqr = 100, r² = 1, |D'| = 1