Human Genetics and Gene Mapping of Complex Traits

Advanced Genetics, Spring 2017
Human Genetics Series
Tuesday 4/4/17
Nancy L. Saccone, nlims@genetics.wustl.edu
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

  LOD score traditionally used to measure statistical evidence for linkage

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

  cases
  controls
<table>
<thead>
<tr>
<th><strong>Linkage mapping:</strong></th>
<th><strong>Association mapping:</strong></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 (close genetic distance between disease locus and marker)</td>
<td>Disease is associated with marker allele that may be either causative or in linkage disequilibrium with causal variant</td>
</tr>
<tr>
<td>Relationship between same allele and trait need not exist across the full sample (e.g. across different families)</td>
<td>works only if association exists at the population level</td>
</tr>
<tr>
<td>robust to allelic heterogeneity: if different mutations occur within the same gene/locus, the method works</td>
<td>not robust to allelic heterogeneity</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
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.

Figure from Strachan and Read, Human Molecular Genetics
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></th>
<th>f_{++}</th>
<th>f_{+d}</th>
<th>f_{dd}</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<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></th>
<th>f_{++}</th>
<th>f_{+d}</th>
<th>f_{dd}</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Genetic models of disease

Other examples of penetrance tables (locus-specific):

\[
\begin{array}{ccc}
  f_{++} & f_{+d} & f_{dd} \\
  0 & 1 & 1 \\
  0 & 0 & 1 \\
  0 & 0 & 0.9 \\
  0.1 & 1 & 1 \\
  0.1 & 0.8 & 0.8 \\
\end{array}
\]

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 \( y = \) 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)
\( x_2, \ldots, x_n = \) 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

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

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

**Fitted line, Slope = \( \beta \)**

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

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

\( \alpha \) = intercept

\( \beta \) = slope of fitted line

x-axis: number of alleles
“Phenotypic variance explained”

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

\( \beta = \text{slope of Fitted line} \)

\( \alpha \)

x-axis: number of alleles

\( r^2 = \text{squared correlation coefficient} \)

Indicates proportion of phenotypic variance in y that’s explained by x
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.

\[\begin{array}{c}
\text{IBS}=1 \\
\text{IBD}=0 \\
\end{array} \quad \begin{array}{c}
\text{IBS}=1 \\
\text{IBD}=1 \\
\end{array}\]
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 $x_1 = \text{genotype (additive coding)}$

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

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

Note that can exponentiate both sides to get odds $= P / (1-P)$:

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

= $\Omega$

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

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

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

- Odds Ratio:
\[
\left( \frac{P_1 / (1 - P_1)}{P_0 / (1 - P_0)} \right) = \frac{e^{[\alpha + \beta_2 x_2 + ... + \beta_n x_n ]} + \beta_1}{e^{[\alpha + \beta_2 x_2 + ... + \beta_n x_n ]}} = e^{\beta_1}
\]
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 \(r^2\), though not squared correlation
Published Genome-Wide Associations (p≤5×10⁻⁸)

http://www.ebi.ac.uk/gwas/diagram
| Abdominal aortic aneurysm | Acute lymphoblastic leukemia | Adhesion molecules | Adiponectin levels | Age-related macular degeneration | AIDS progression | Alcohol dependence | Alopecia areata | Alzheimer disease | Amyloid A levels | Amyotrophic lateral sclerosis | Angiotensin-converting enzyme activity | Ankylosing spondylitis | Arterial stiffness | Asparagus anosmia | Asthma | Atherosclerosis in HIV | Atrial fibrillation | Attention deficit hyperactivity disorder | Autism | Basal cell cancer | Behcet’s disease | Bipolar disorder | Biliary atresia | Bilirubin | Bitter taste response | Birth weight | Bladder cancer | Bleomycin sensitivity | Blond or brown hair | Blood pressure | Blue or green eyes | BMI, waist circumference | Bone density | Breast cancer | C-reactive protein | Calcium levels | Cardiac structure/function | Cardiovascular risk factors | Carcinoembryonic antigen | Carotenoid/tocopherol levels | Celiac disease | Celiac disease and rheumatoid arthritis | Cerebral atrophy measures | Chronic lymphocytic leukemia | Chronic myeloid leukemia | Cleft lip/palate | Coffee consumption | Cognitive function | Conduct disorder | Colorectal cancer | Corneal thickness | Coronary disease | Creutzfeldt-Jakob disease | Crohn’s disease | Crohn’s disease and celiac disease | Cystic fibrosis severity | Dermatitis | DHEA-s levels | Diabetic retinopathy | Dilated cardiomyopathy | Drug-induced liver injury | Drug-induced liver injury (hepatocyte damage) | Endometrial cancer | Endometriosis | Eosinophil count | Eosinophilic esophagitis | Erectile dysfunction and prostate cancer treatment | Erythrocyte parameters | Esophageal cancer | Essential tremor | Exfoliation glaucoma | Eye color traits | F cell distribution | Fibrinogen levels | Folate pathway vitamins | Follicular lymphoma | Fuch’s corneal dystrophy | Freckles and burning | Gallstones | Gastric cancer | Glialoma | Glycemic traits | Hair color | Hair morphology | Handedness in dyslexia | HDL cholesterol | Heart failure | Heart rate | Height | Hemostasis parameters | Hepatic steatosis | Hepatitis | Hepatocellular carcinoma | Hirschsprung’s disease | HIV-1 control | Hodgkin’s lymphoma | Homocysteine levels | Hypoplasias | Idiopathic pulmonary fibrosis | IFN-related cytophenia | IgA levels | IgE levels | Inflammatory bowel disease | Insulin-like growth factors | Intracranial aneurysm | Iris color | Iron status markers | Ischemic stroke | Juvenile idiopathic arthritis | Keloid | Kidney stones | LDL cholesterol | Leprosy | Leptin receptor levels | Liver enzymes | Longevity | Lp(a) levels | LpPLA2 (activity and mass) | Lung cancer | Magnesium levels | Major mood disorders | Malaria | Male pattern baldness | Mammographic density | Matrix metalloproteinase levels | MCP-1 | Melanoma | Menarche & menopause | Meningococcal disease | Metabolic syndrome | Migraine | Moyamoya disease | Multiple sclerosis | Myeloproliferative neoplasms | Myopia (pathological) | N-glycan levels | Narcolepsy | Nasopharyngeal cancer | Natriuretic peptide levels | Neuroblastoma | Nicotine dependence | Obesity | Open angle glaucoma | Open personality | Optic disc parameters | Osteoarthritis | Osteoporosis | Osteosclerosis | Other metabolic traits | Ovarian cancer | Pancreatic cancer | Pain | Paget’s disease | Panic disorder | Parkinson’s disease | Periodontitis | Peripheral arterial disease | Personality dimensions | Phosphatidylcholine levels | Phosphorus levels | Photonic sneeze | Phytosterol levels | Platelet count | Polycystic ovary syndrome | Primary biliary cirrhosis | Primary sclerosing cholangitis | PR interval | Progranulin levels | Progressive supranuclear palsy | Prostate cancer | Protein levels | PSA levels | Psoriasis | Psoriatic arthritis | Pulmonary function COPD | QRS interval | QT interval | Quantitative traits | Recombination rate | Red vs. non-red hair | Refractive error | Renal cell carcinoma | Renal function | Response to antidepressants | Response to antipsychotic therapy | Response to carbamazepine | Response to clopidogrel therapy | Response to hepatitis C treatment | Response to interferon beta therapy | Response to metformin | Response to statin therapy | Restless legs syndrome | Retinal vascular caliber | Rheumatoid arthritis | Ribavirin-induced anemia | Schizophrenia | Serum metabolites | Skin pigmentation | Smoking behavior | Speech perception | Sphingolipid levels | Statin-induced myopathy | Stroke | Sudden cardiac arrest | Suicide attempts | Systemic lupus erythematosus | Systemic sclerosis | T-tau levels | Tau AB1-42 levels | Telomere length | Testicular germ cell tumor | Thyroid cancer | Thyroid volume | Tooth development | Total cholesterol | Triglycerides | Tuberculosis | Type 1 diabetes | Type 2 diabetes | Ulcerative colitis | Urate | Urinary albumin excretion | Urinary metabolites | Uterine fibroids | Venous thromboembolism | Ventricular conduction | Vertical cup-disc ratio | Vitamin B12 levels | Vitamin D insufficiency | Vitiligo | Warfarin dose | Weight | White cell count | White matter hyperintensity | YKL-40 levels |
Displaying GWAS results

“Manhattan plot”

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

$p = 1 \times 10^{-8}$ 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
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)
GWAS results

– “Q-Q plot”: Quantile-quantile plot

Before-and-after adjustment for population stratification

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

\textit{Nicotinic receptor gene cluster}

rs16969968
D398N in \textit{CHRNA5}
Saccone SF et al., 2007
Example: \textit{CHRNA5-CHRNA3-CHRNB4} and rs16969968
Associated with nicotine dependence, smoking, lung cancer, COPD.

\textbf{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

\textbf{rs1317286}
Berrettini et al., 2008

\textbf{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

\textbf{rs8034191}
Hung 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 $A_1A_2B_1B_2$.

Two possible phases:

\[
\begin{align*}
A_1 & \mid \mid A_2 \\
B_1 & \mid \mid B_2
\end{align*}
\]

\[
\begin{align*}
A_1 & \mid \mid A_2 \\
B_2 & \mid \mid B_1
\end{align*}
\]
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>snp1</th>
<th>snp2</th>
<th>expected freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A</td>
<td>0.5 * 0.5</td>
</tr>
<tr>
<td>A</td>
<td>G</td>
<td>0.5 * 0.5</td>
</tr>
<tr>
<td>C</td>
<td>A</td>
<td>0.5 * 0.5</td>
</tr>
<tr>
<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 ...
• Extra slides follow
Number of SNPs in dbSNP over time

solid: cumulative # of non-redundant SNPs.
dotted: validated. dashed: double-hit

from:
Fernald et al., Bioinformatics 2011