Genome-wide association studies (GWAS)
Further reading (not required)

- Nature reviews genetics:
  http://www.nature.com/nrg/series/gwas/index.html
- McCarthy et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges
- Marchini and Howie. Genotype imputation for genome-wide association studies
- Ott. Family-based designs for genome-wide association studies.
- Gibson. Rare and common variants: twenty arguments.
Outline for GWAS

• Review / Overview

• Design

• Analysis
  – QC
  – Prostate cancer example
  – Imputation
  – Replication & Meta-analysis

• Advanced analysis intro (more next lecture)
  – Limitations & “missing heritability”
  – Gene/pathway tests
  – Polygenic models
Outline for GWAS

• Review / Overview
• Design
• Analysis
  – QC
  – Prostate cancer example
  – Imputation
  – Replication & Meta-analysis
• Advanced analysis intro (more next lecture)
  – Limitations & “missing heritability”
  – Gene/pathway tests
  – Polygenic models
Published Genome-Wide Associations through 07/2012
Published GWA at \( p \leq 5 \times 10^{-8} \) for 18 trait categories

NHGRI GWA Catalog
www.genome.gov/GWASTudies
www.ebi.ac.uk/fgpt/gwas/
Outline for GWAS

• Review / Overview

• Design

• Analysis
  – QC
  – Prostate cancer example
  – Imputation
  – Replication & Meta-analysis

• Advanced analysis intro (more next lecture)
  – Limitations & “missing heritability”
  – Gene/pathway tests
  – Polygenic models
Design 1: Microarray – Genotype a subset of markers available

Direct association
Disease locus directly typed

Indirect association (guilt by association)
Marker correlated with typed disease locus

- Previously we talked about using this to “tag” candidate genes (only Genotype a subset of the SNPs).
- Same principal applies to tagging the whole genome.

Hirschhorn & Daly, Nat Rev Genet 2005
Technology behind a GWAS Microarray

Target prep
- Amplify
- Fragment

Hybridization
- Capture
- Labeled solution probe
- Sample binds to array

Ligation
- Differentiate
- Labeled probes bind to sample, differentiating between the two alleles

Signal amplification
- Make it bright enough then measure intensity of array

Assay ~ 0.7 - 5M SNPs (keeps increasing)
Raw microarray data: Genotype calls

Good calls!

Bad calls!
Design 2: Next generation sequencing

- Genotype essentially “all” SNPs in genome
  - A bunch of random reads all over the genome are assembled to produce the genome
  - Some regions of the genome have more reads than others; can choose how much you want to sequence (5x coverage 1000 Genomes, 60-80x Complete Genomics, …)

- More expensive
- Extreme phenotypes, since can't do everyone?
- More in “Next Generation Sequencing” lecture
Design choices

- **GWAS Microarray**
  - Only assay SNPs designed into array (0.7-5 million)
  - Cheaper (so more subjects tradeoff)

- **GWAS Sequencing**
  - “De novo” discovery (particularly good for rare variants)
  - More expensive (but costs are falling) (many less subjects)
  - Need much more expansive IT support
  - Lots of interesting interpretation problems (field rapidly evolving)

Can also do a hybrid design, more later in the lecture.
Design choices

- **Exome Microarray**
  - Only assay SNPs designed into array (~300K+custom); in **exons only** and that could affect protein coding function
  - Cheapest (so many more subjects)

- **Exome Sequencing**
  - “De novo” discovery (particularly good for rare variants); % of exons only (first step is a pull down that does not capture all exons)
  - More expensive than microarrays, less expensive than gwas sequencing
  - Need more expansive IT support, but not as much as whole genome
Size matters

- Large # SNPs tested – multiple comparisons
- Very small effect sizes
- Tag SNP, rather than actual SNP

Visscher, AJHG 2012
Biggest studies...

• GWAS Microarray: 100,000 People in the Kaiser RPGEH, still to be analyzed (Hoffmann et al., Genomics, 2011a&b)
• Sequencing: 1000 Genomes Project, UK10K
• Exome Sequencing: GO ESP (12,031 subjects, for exome microarray design)
Outline for GWAS

• Review / Overview
• Design

• Analysis
  – QC
  – Prostate cancer example
  – Imputation
  – Replication & Meta-analysis

• Advanced analysis intro (more next lecture)
  – Limitations & “missing heritability”
  – Gene/pathway tests
  – Polygenic models
QC Steps

• “garbage rises to the top”
• Filter SNPs and Individuals
  – MAF (e.g., 1%, but very sample size dependent!)
  – Low call rates (means genotype probably didn't work correctly)
• Test for HWE among controls & within ethnic groups. Use conservative alpha-level ($10^{-6}$, e.g., but opinions vary) (again, to remove artifacts)
• Check for relatedness.
  – MZ twins, or accidentally genotyped same sample twice?
  – Remove first degree, or account for.
• Check genotype gender (mislabeled samples)
• Filter Mendelian inheritance (family-based, or potentially cryptics, if large enough sample)
• plink software: pngu.mgh.harvard.edu/~purcell/plink/reference.shtml
Filter for Mendelian inheritance

A|A------A|A
   |
   A|T ← Offspring Genotype not Possible!
   (De novo mutation, but more likely genotyping error.)
Check for relatedness, e.g., HapMap

- Take overall average of all SNPs of how many alleles are shared.
- E.g., parent-offspring never shares zero alleles.
- HapMap was supposed to be unrelateds (this was a bit of a surprise!)

Pemberton et al., AJHG 2010
GWAS analysis

• Most common approach: Model each SNP separately
• Additive coding of SNP most common, just a covariate in a regression framework
• Dichotomous phenotype: logistic regression
• Continuous phenotype: linear regression
• Correct for multiple comparisons
  • e.g., Bonferroni, 1 million gives $\alpha=5\times10^{-8}$
  • more next time
• Adjust for potential population stratification
  • principal components (PC’s), on best performing SNPs
  • software usually does LD filter (e.g., Eigensoft)
Recap of population substructure

• Two populations have different disease frequency, and different allele frequency.
• Association picks up they are different populations!
Control for race/ethnicity/ancestry: Principal Components (PCs) as covariates in regression model

- Li et al., Science 2008
PCs pick up fine population structure

- Razib, Current Biology 2008
Adjusting for PC's

Do not want separate cluster for cases as controls! – Why want individuals from similar population for controls.
After run analysis

- Assess the fit with QQ-plot
- Look at results in Manhattan plots
- **Replication** of hits in a separate cohort
Quantile-quantile (QQ) plot

Most SNPs are on the line, but want a few hits off the line (true significant associations!)
QQ-plots and PC adjustment recap

- Exactly on line if no signal at all.
- If don't adjust for PC's, then p-values are all inflated artificially.
- Adjusting for PC's fixes the problem.

Wang, BMC Proc 2009
Manhattan plot example

HLA region

-10

-5

5

10

15

\(-\log_{10}(P)\)

Chromosome

Systemic Schlerosis (auto-immune disease) Radstake et al., Nature Genetics 2010
But they are not edible!
They are Manhattan Plots

http://www.biocomicals.com
Replication, Replication, Replication

• To replicate:
  – Association test for replication sample significant at $0.05/\{\text{Number of SNPs replicating}\}$ alpha level
  – Same genetic model (e.g. additive, dominant)
  – Same direction
  – Sufficient sample size for replication

• Non-replications not necessarily a false positive
  – LD structures, different populations (e.g., flip-flop)
  – covariates, phenotype definition, underpowered
Meta-analysis

• Combine multiple studies to increase power
• Either combine p-values (Fisher’s test),
• or coefficient estimates + standard error (better)
Replication & Meta-analysis

Biological, clinical and population relevance of 95 loci for blood lipids

A list of authors and their affiliations appears at the end of the paper.

Table 1 | Meta-analysis of plasma lipid concentrations in $\geq 100,000$ individuals of European descent.

<table>
<thead>
<tr>
<th>Locus</th>
<th>Chr</th>
<th>Lead SNP</th>
<th>Lead trait</th>
<th>Other traits</th>
<th>Alleles/MAF</th>
<th>Effect size</th>
<th>P</th>
<th>eQTL</th>
<th>CAD</th>
<th>Ethnic</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDLRA1</td>
<td>1</td>
<td>rs12027135</td>
<td>TC</td>
<td>LDL</td>
<td>T/A/0.45</td>
<td>-1.22</td>
<td>$4 \times 10^{-11}$</td>
<td>Y</td>
<td>Y</td>
<td>+++?</td>
</tr>
<tr>
<td>PABPC4</td>
<td>1</td>
<td>rs4660293</td>
<td>HDL</td>
<td></td>
<td>A/G/0.23</td>
<td>-0.48</td>
<td>$4 \times 10^{-10}$</td>
<td>Y</td>
<td>Y</td>
<td>++++</td>
</tr>
<tr>
<td>PCSK9</td>
<td>1</td>
<td>rs2479409</td>
<td>LDL</td>
<td>TC</td>
<td>A/G/0.30</td>
<td>+2.01</td>
<td>$2 \times 10^{-28}$</td>
<td>++++</td>
<td>++</td>
<td></td>
</tr>
<tr>
<td>ANGPTL3</td>
<td>1</td>
<td>rs2131925</td>
<td>TG</td>
<td>TC, LDL</td>
<td>T/G/0.32</td>
<td>-4.94</td>
<td>$9 \times 10^{-43}$</td>
<td>Y</td>
<td>Y</td>
<td>++++</td>
</tr>
<tr>
<td>EVIS</td>
<td>1</td>
<td>rs7515577</td>
<td>TC</td>
<td></td>
<td>A/C/0.21</td>
<td>-1.18</td>
<td>$3 \times 10^{-8}$</td>
<td>++++</td>
<td>++</td>
<td></td>
</tr>
<tr>
<td>SORT1</td>
<td>1</td>
<td>rs629301</td>
<td>LDL</td>
<td>TC</td>
<td>T/G/0.22</td>
<td>-5.65</td>
<td>$1 \times 10^{-170}$</td>
<td>Y</td>
<td>Y</td>
<td>++++</td>
</tr>
<tr>
<td>ZNF648</td>
<td>1</td>
<td>rs1689800</td>
<td>HDL</td>
<td></td>
<td>A/G/0.35</td>
<td>-0.47</td>
<td>$3 \times 10^{-10}$</td>
<td>++++</td>
<td>++</td>
<td></td>
</tr>
</tbody>
</table>

The gene name listed in ‘Locus’ column is either a plausible biological candidate gene in the locus or the nearest annotated gene to the lead SNP. Listed in ‘Lead trait’ column is the lipid trait with best $P$-value among all four traits. Listed in ‘Other traits’ are additional lipid traits with $P \leq 5 \times 10^{-8}$. Listed in ‘Alleles/MAF’ column are: major allele, minor allele and minor allele frequency (MAF) within the combined cohorts included in this meta-analysis (alleles designated with respect to the ‘+’ strand; Supplementary Table 2). Numbers in ‘Effect size’ column are in mg dl$^{-1}$ for the lead trait, modelled as an additive effect of the minor allele. $P$-values are listed for the lead traits. In the ‘eQTL’ column, ‘Y’ indicates that lead SNP has an eQTL with at least one gene within 500 kb with $P \leq 5 \times 10^{-8}$ in at least one of the three tissues tested (liver, omental fat, subcutaneous fat). In the ‘CAD’ column, ‘Y’ indicates that the lead SNP meets the pre-specified statistical significance threshold of $P < 0.001$ for association with CAD and being concordant between the direction of lipid effect and the change in CAD risk. In the ‘Ethnic’ column, ‘+’ indicates concordant effect on lead trait of the variant between the primary meta-analysis cohort and the European or non-European group, ‘-’ indicates discordant effect on lead trait, and ‘?’ indicates data not available for the group; in order, the ethnic groups are European, East Asian, South Asian and African American (Supplementary Table 11).

Chr, chromosome.
Beyond Genotyped SNPs: Imputation of SNP Genotypes

- Combine data from different platforms (e.g., Affy & Illumina) (for replication / meta-analysis).
- Estimate unmeasured or missing genotypes.
- Based on measured SNPs and external info (e.g., haplotype structure of HapMap).
- Increase GWAS power (impute and analyze all), e.g. Sick sinus syndrome, most significant was 1000 Genomes imputed SNP (Holm et al., Nature Genetics, 2011)
- HapMap as reference, now 1000 Genomes Project?
- Hybrid sequencing/genotyping approaches
Generalization of Tag SNPs: Imputed SNPs using a Reference Panel

- Generalization of pairwise $r^2$ of tagSNPs to multi-marker correlation

**Study sample**

```
.....A.....A.....A
.....G.....C.....A...
```

**Reference haplotypes**

```text
CGAGATCTCCTTCTTCTGTGC
CGAGATCTCCCGACCTCATGG
CCAAGCTCTTTTCTTCTGTGC
CGAGATCTTTTTCTTCTGTGC
CGAGATCTCCACCTCATGG
TGGGATCTCCCGACCTCATGG
CGAGATCTCCACCTCATGG
CGAGATCTTTTTCTTCTGTGC
CGAGATCTCCACCTCATGG
CGAGATCTTTTTCTTCTGTGC
CGAGATCTTTTTCTTCTGTGC
CGAAAGCTTTTTTTTTTTTTTTC
CGAGATCTCCCGACCTCATGG
CCAAGCTCTTTTCTTCTGTGC
CGAGATCTTTTTCTTCTGTGC
CGAGATCTCCACCTCATGG
TGGGATCTCCCGACCTCATGG
CGAGATCTCCACCTCATGG
CGAGATCTTTTTCTTCTGTGC
CGAGATCTCCACCTCATGG
CGAGATCTTTTTCTTCTGTGC
CGAGATCTTTTTCTTCTGTGC
CGAAAGCTTTTTTTTTTTTTTTC
```

**Study sample**

```
.....A.....A.....A
.....G.....C.....A...
```

**Reference haplotypes**

```text
cgagAtctccccgACctcAtgg
cgaAGctttttttttttttttcAtgg
```

**Reference haplotypes**

```text
CGGCCCCCGGCAATTTTTTTTTTT
CGAGATCTCCCGACCTCATGG
CCAAGCTCTTTTCTTCTGTGC
CGAGATCTTTTTCTTCTGTGC
CGAGATCTCCACCTCATGG
TGGGATCTCCCGACCTCATGG
CGAGATCTCCACCTCATGG
CGAGATCTTTTTCTTCTGTGC
CGAGATCTCCACCTCATGG
CGAGATCTTTTTCTTCTGTGC
CGAGATCTTTTTCTTCTGTGC
CGAAAGCTTTTTTTTTTTTTTTC
CGAGATCTCCCGACCTCATGG
CCAAGCTCTTTTCTTCTGTGC
CGAGATCTTTTTCTTCTGTGC
CGAGATCTCCACCTCATGG
TGGGATCTCCCGACCTCATGG
CGAGATCTCCACCTCATGG
CGAGATCTTTTTCTTCTGTGC
CGAGATCTCCACCTCATGG
CGAGATCTTTTTCTTCTGTGC
CGAGATCTTTTTCTTCTGTGC
CGAAAGCTTTTTTTTTTTTTTTC
```

- [http://mathgen.stats.ox.ac.uk/impute/impute_v2.html](http://mathgen.stats.ox.ac.uk/impute/impute_v2.html)
- [http://faculty.washington.edu/browning/beagle/beagle.html](http://faculty.washington.edu/browning/beagle/beagle.html)
- [http://www.sph.umich.edu/csg/abecasis/MACH/download/](http://www.sph.umich.edu/csg/abecasis/MACH/download/)

Li et al., Ann Rev Human Genet, 2009
Imputation Application

*TCF7L2* gene region & T2D from the WTCCC data

Observed genotypes black
Imputed genotypes red.

Imputed SNP more significant than typed SNP!
Outline for GWAS

• Review / Overview
• Design
• Analysis
  – QC
  – Prostate cancer example
  – Imputation
  – Replication & Meta-analysis
• Advanced analysis intro (more next lecture)
  – Limitations & “missing heritability”
  – Gene/pathway tests
  – Polygenic models
Limitations of GWAS

- Not very predictive

Example:
AUC for Breast Cancer Risk

58%: Gail model (# first degree relatives w bc, age menarche, age first live birth, number of previous biopsies) + age, study, entry year
58.9%: SNPs
61.8%: Combined

Wacholder et al., NEJM 2010

Witte, Nat Rev Genet 2009
Limitations of GWAS

- Not very predictive
- Explain little heritability
- Focus on common variation
- Many associated variants are not causal
Where's the heritability?

Table 1. Population Variation Explained by GWAS for a Selected Number of Complex Traits

<table>
<thead>
<tr>
<th>Trait or Disease</th>
<th>$h^2$ Pedigree Studies</th>
<th>$h^2$ GWAS Hits$^a$</th>
<th>$h^2$ All GWAS SNPs$^b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type 1 diabetes</td>
<td>0.9$^{98}$</td>
<td>0.6$^{99,c}$</td>
<td>0.3$^{12}$</td>
</tr>
<tr>
<td>Type 2 diabetes</td>
<td>0.3–0.6$^{100}$</td>
<td>0.05–0.10$^{34}$</td>
<td></td>
</tr>
<tr>
<td>Obesity (BMI)</td>
<td>0.4–0.6$^{101,102}$</td>
<td>0.01–0.02$^{36}$</td>
<td>0.2$^{14}$</td>
</tr>
<tr>
<td>Crohn’s disease</td>
<td>0.6–0.8$^{103}$</td>
<td>0.1$^{11}$</td>
<td>0.4$^{12}$</td>
</tr>
<tr>
<td>Ulcerative colitis</td>
<td>0.5$^{103}$</td>
<td>0.05$^{12}$</td>
<td></td>
</tr>
<tr>
<td>Multiple sclerosis</td>
<td>0.3–0.8$^{104}$</td>
<td>0.1$^{45}$</td>
<td></td>
</tr>
<tr>
<td>Ankylosing spondylitis</td>
<td>&gt;0.90$^{105}$</td>
<td>0.2$^{106}$</td>
<td></td>
</tr>
<tr>
<td>Rheumatoid arthritis</td>
<td>0.6$^{107}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Schizophrenia</td>
<td>0.7–0.8$^{108}$</td>
<td>0.01$^{79}$</td>
<td>0.3$^{109}$</td>
</tr>
<tr>
<td>Bipolar disorder</td>
<td>0.6–0.7$^{108}$</td>
<td>0.02$^{79}$</td>
<td>0.4$^{12}$</td>
</tr>
<tr>
<td>Breast cancer</td>
<td>0.3$^{110}$</td>
<td>0.08$^{111}$</td>
<td></td>
</tr>
<tr>
<td>Von Willebrand factor</td>
<td>0.66–0.75$^{112,113}$</td>
<td>0.13$^{114}$</td>
<td>0.25$^{14}$</td>
</tr>
<tr>
<td>Height</td>
<td>0.8$^{115,116}$</td>
<td>0.1$^{13}$</td>
<td>0.5$^{13,14}$</td>
</tr>
<tr>
<td>Bone mineral density</td>
<td>0.6–0.8$^{117}$</td>
<td>0.05$^{118}$</td>
<td></td>
</tr>
<tr>
<td>QT interval</td>
<td>0.37–0.60$^{119,120}$</td>
<td>0.07$^{121}$</td>
<td>0.2$^{14}$</td>
</tr>
<tr>
<td>HDL cholesterol</td>
<td>0.5$^{122}$</td>
<td>0.1$^{57}$</td>
<td></td>
</tr>
<tr>
<td>Platelet count</td>
<td>0.8$^{123}$</td>
<td>0.05–0.1$^{58}$</td>
<td></td>
</tr>
</tbody>
</table>

$^a$ Proportion of phenotypic variance or variance in liability explained by genome-wide-significant and validated SNPs. For a number of diseases, other parameters were reported, and these were converted and approximated to the scale of total variation explained. Blank cells indicate that these parameters have not been reported in the literature.

$^b$ Proportion of phenotypic variance or variance in liability explained when all GWAS SNPs are considered simultaneously. Blank cell indicate that these parameters have not been reported in the literature.

$^c$ Includes pre-GWAS loci with large effects.

Visccher, AJHG 2011
Where’s the heritability?

Common disease rare variant (CDRV) hypothesis: diseases due to multiple rare variants with intermediate penetrances (allelic heterogeneity)

See: NEJM, April 30, 2009

McCarthy et al., 2008
Will GWAS results explain more heritability?

Possibly, if...

1. Causal SNPs not yet detected due to power issues – weak effects
   - Previous GWAS designed/powered for MAF>0.05/0.10; lower MAFs for newer arrays and larger reference panels

1. More complicated modeling necessary, e.g. gene-gene interaction, gene-environment interaction, etc.
Gene/pathway-based tests

• Various ways of collapsing the genotype information in multiple genes
  – Less multiple comparison adjustment

• \( \text{logit (Prob}(y=1|\ x, \ c)) = \alpha + \beta x + \gamma c \)
e.g. = \( \alpha + \beta_1 x_1 + \ldots + \beta_{12} x_{12} + \{\text{PC’s, other covariates}\} \)
  – \( y \): disease status
  – \( x \) is a vector of genotypes (e.g., a gene, or a pathway)
  – \( c \) is a vector of covariates

• \( H_0: \beta = 0 \)
Pathways - how to define?

• Many websites / companies provide ‘dynamic’ graphic models of molecular and biochemical pathways.

• Examples: GO (Gene Ontology), KEGG, BioCarta, Reactome

• May be interested in potential joint and/or interaction effects of multiple genes in one pathway.
Other Extreme: Polygenic Models

- Many weak associations combine to risk?
- Score model (use all GWAS SNPs):

\[
\sum_{i=1}^{m} \ln (OR_i) \times SNP_{ij}
\]

\[
x_j = \frac{\sum}{m}
\]

where
- \( \ln(OR_i) \) = ‘score’ for SNP\(_i\) from ‘discovery’ sample
- \( SNP_{ij} \) = # of alleles (0,1,2) for SNP\(_i\), person j in ‘validation’ sample.
- Large number of SNPs (\( m \))

- \( x_j \) associated with disease?

ISC / Purcell et al. Nature 2009
Application of Model

\[ P = 2 \times 10^{-28} \]

![Graph showing variance explained for different psychiatric conditions](image)

- **Schizophrenia**:
  - MGS-EA
  - MGS-AA
  - O'Donovan

- **Bipolar disorder**:
  - STEP-BD
  - WTCCC

- **Non-psychiatric (WTCCC)**:
  - CAD
  - CD
  - HT
  - RA
  - T1D
  - T2D

\[ P_T < 0.1 \]
\[ P_T < 0.2 \]
\[ P_T < 0.3 \]
\[ P_T < 0.4 \]
\[ P_T < 0.5 \]
Other things than SNPs

• Copy number variants (CNVs)
• Epigenetics, e.g., methylation, histone modification
• Gene expression (RNA levels)
• Proteomics (measures of proteins in cells)
• ...