Genome Wide Association Studies (GWAS)

Samantha Breaux
GWAS, What Are They?

A study design to identify phenotype-genotype association across the genome

- Identifies regions of the genome for further analysis
Why is determining genotype from phenotype hard?

Most traits do not follow mendelian genetics

Genome is very large

Interaction with several different genes

Observed phenotype is not wholly genetics based
Common Disease Common Variant Hypothesis

Individuals with a common disease will share common genetic variations

... So we should be able to identify these common variants by comparing genomes of affected individuals and controls
Typical Study Design

- Large number of participants 1k-100ks
- Grouped based on their specific phenotype criteria
- Assaying \( \approx 10^6 \) alleles per microarray
- Quality control
- Imputation
- Analysis

http://knowgenetics.org/genome-wide-association-studies-gwas/
DNA Microarrays

Investigational DNA sequences (probe/ oligo) that are densely ($10^{-12}$ pmol) bound at microscopic spots along a solid surface.

https://www.mun.ca/biology/scarr/DNA_Chips.html
SNPs
Single Nucleotide Polymorphisms
- Ex. A -> G or C -> G
- Most common type of mutation

Haplotype
SNPs on a chromosome that are inherited together
- Ie. alleles that are at linkage disequilibrium
Knowing a SNP in a common haplotype allows you to infer the other SNPs at that location
Imputation

The process of using statistical inference of known haplotypes to predict genotype

Allows investigation of association between a trait of interest and unsequenced genotypes

Increases SNP density vs genome length

https://www.nature.com/articles/nature01400/figures/3
Preprocessing Genotyping Data

- Missing SNPs/ individuals
- Sex Discrepancy
- Minor Allele Frequency
- Hardy-Weinberg equilibrium

- Heterozygosity
- Relatedness
- Population stratification
Association Analysis

Typically a regression analysis of each SNP on a given trait

Adjusted for individual clinical, environmental, and ethnic factors

Allele affect is considered

- An additive allele model is generally chosen
- All models can be chosen but must correct for multiple testing

Odds ratio for each SNP calculated and p-value assigned
Odds Ratios

Measure of association between an event and an outcome

So for GWA the odds ratio for each SNP would be..

\[
\text{Odds Ratio} = \frac{\text{# of SNP Cases} \times \text{# of no SNP Controls}}{\text{# of SNP control} \times \text{# of no SNP Cases}}
\]

Number of cases with C at SNP 1 = 2200
Number of cases without C at SNP 1 = 1800
Number of controls with C at SNP 1 = 3400
Number of controls without C at SNP 1 = 4600

\[
\text{OR} = \frac{2200 \times 4600}{3400 \times 1800} = 1.6536
\]
Data Visualization: Manhattan Plot

- Bonferroni level of significance threshold $5 \times 10^{-8}$
- Study specific threshold $5 \times 10^{-6}$

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5019244/figure/sim6605-fig-0003/
Limitations

Populations available very very biased

Standardized phenotyping

Large samples sets

Poorly suited to detect rare variants

Results require correlating

Multiple testing problem

Data dependent on technology used
Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression

Completed a genome-wide association meta-analysis of 9.6 million imputed SNPs in 135,458 MDD and major depression cases and 344,901 controls

- SNPs involved with brain development

https://www.nature.com/articles/s41588-018-0090-3
Questions?
Sources

https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002822
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2925172/
https://www.nature.com/scitable/definition/haplotypes
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2938757/
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5019244/
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6001694/
http://www.biostathandbook.com/multiplecomparisons.html
http://knowgenetics.org/genome-wide-association-studies-gwas/
https://www.nature.com/articles/s41588-018-0090-3