Introduction to Genome-wide Association Studies

Shi-Heng Wang
Postdoctoral Fellow, Institute of Epidemiology and Preventive Medicine, National Taiwan University
September 4, 2014
Genome-wide Association Studies (GWAS)

• An examination of many common genetic variants in different individuals, to see if any variant is associated with disease
  – Traditionally comparing two groups: cases and controls
  – Millions of genetic variants are read using SNP arrays for each subject
    • The entire genome, non-candidate-driven studies
Manhattan plot for the GWAS of uterine fibroids (子宮纖維瘤)
Cha et al., 2011 *Nature Genetics*
False positives/negatives in GWAS

- Genotyping errors, differences in sample collection, population differences...
False positives/negatives in GWAS

- Genotyping errors, differences in sample collection, population differences...
Demographic statistics

- Genotype frequencies
- Allele frequencies
  - 利用 Genotype frequencies 估計 Allele frequencies
    - $f(A) = \frac{120}{200} = 0.6$
    - $f(a) = \frac{80}{200} = 0.4$

- Hardy-Weinberg Equilibrium Proportion Test
- SNP call rate
- Individual sample call rate
Data Quality control

• Batch effect
  – In genotyping experiment
    • Randomized plating & plate positions (Case-control phenotypes)
  – In genotyping calling

• Sample quality
  – Call rate (>98% or 99%)
  – Sample relatedness
  – Population stratification

• Marker quality
Sample quality: Sample relatedness

- To compute pairwise kinship estimates between every individual (kinship ≤0.05, unrelated)
- The proportion of loci where two individuals share zero, one, or two alleles identical by descent (IBD)

Sharing two alleles IBD at every locus are monozygotic twins
Sharing one allele IBD at every locus are parent-child pairs
Siblings share zero, one, and two alleles IBD at 25%, 50%, and 25%
Sample quality: Population stratification

• Data comprise multiple groups of individuals who differ systematically in both genetic ancestry and the phenotype
  – Spurious associations would be due to differences in ancestry rather than true association of alleles to disease
• Data from a relatively homogenous population
• Principal components analysis
  – Adjusting in association test
Marker quality

• SNP call rate
  – >98% or 99%

• Minor allele frequencies (MAF)
  – low statistical power for rare SNPs

• Hardy-Weinberg Equilibrium (HWE)
  – Departure from HWE can be indicative of potential genotyping errors, population stratification
Association test

- Test one SNP at a time
  - Inheritance mode
    - Additive inheritance mode
    - Dominant inheritance mode
    - Recessive inheritance mode
  - Q-Q plot
  - Manhattan plot
  - Multiple testing

- Test more than one SNP at a time
  - Haplotype analysis
  - Two-loci interaction
  - Polygenic risk score
Quantile-quantile (Q-Q) plot

- The expected distribution of association test statistics (X-axis) and the observed values (Y-axis)
  - Association statistics: $\chi^2$, -log (p-value)
- A clean QQ plot show a solid line matching X=Y until it sharply curves at the end
  - The small number of true associations

Pearson and Manolio, 2008 JAMA
Polygenic model (1)

- Training sample and replication sample
- Training sample
  - Association test for each variant
  - The markers were ranked by p-values
- Replication sample
  - Computing polygenic score: $\hat{G} = \sum_{i=1}^{m} \hat{b}_i \hat{x}_i$
    - $\hat{x}_i = 0, 1$ or $2$ reference alleles of a marker
    - $\hat{b}_i$ : the estimated effect size from the training sample
Polygenic model (2)

• Testing the association between the score and the disease status in the replication sample
• To quantify how much of the variation in disease status is explained by the score
Fig. Replication of the ISC-derived polygenic component in independent schizophrenia and bipolar disorder samples

Purcell et al. *Nature* 2009
Reference

