Genome-Wide Association Study

02-710 Computational Genomics
Seyoung Kim
Overview

• How can we identify the genetic loci responsible for determining phenotypes?
  • Linkage analysis
    – Data are collected for family members
    – Difficult to collect data on a large number of families
    – Effective for rare diseases
    – Low resolution on the genomes due to only few recombinations
      » a large region of linkage

• Genome-wide association studies
  – Data are collected for unrelated individuals
  – Easier to find a large number of affected individuals
  – Effective for common diseases, compared to family-based method
  – Relatively high resolution for pinpointing the locus linked to the phenotype
Overview

- **Statistical methods for testing genotype/phenotype associations**
  - Discrete-valued phenotype: case/control study
  - Continuous-valued phenotype: quantitative traits
  - Sparse regression method for considering all of the SNP markers
  - Multimarker association test

- **Issues arising in GWAS**
  - Genotype imputation
  - From common to rare variants
  - Epistasis for multiple interacting loci
  - Correcting for population structure
Population Genotype/Phenotype Data

Phenotype data

\[ y = \begin{pmatrix} y^1 \\ \vdots \\ y^N \end{pmatrix}_{N \text{ individuals}} \]

Genotype data

\[ X = \begin{pmatrix} x^1_1 & \cdots & x^1_J \\ \vdots & \ddots & \vdots \\ x^N_1 & \cdots & x^N_J \end{pmatrix}_{N \text{ individuals}} \]

- 0 or 1 for case/control studies
  - e.g., healthy/diabetic
- Real-valued phenotypes
  - e.g., cholesterol level
Single SNP Association Test: Case/Control Study

- For each marker locus, find the 3x2 contingency table containing the counts of three genotypes

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Case</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>$N_{case,AA}$</td>
<td>$N_{control,AA}$</td>
</tr>
<tr>
<td>Aa</td>
<td>$N_{case,Aa}$</td>
<td>$N_{control,Aa}$</td>
</tr>
<tr>
<td>aa</td>
<td>$N_{case,aa}$</td>
<td>$N_{control,aa}$</td>
</tr>
<tr>
<td>Total</td>
<td>$N_{case}$</td>
<td>$N_{control}$</td>
</tr>
</tbody>
</table>

- $\chi^2$ test with 2 df under the null hypothesis of no association

Genotype score = the number of minor alleles
Single SNP Association Analysis: Case/Control Study

- Alternatively, assume the heterozygote risk is approximately between the two homozygotes.
- Form a 2x2 contingency table. Each individual contributes twice from each of the two chromosomes.

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Case</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>G_{case,A}</td>
<td>G_{control,A}</td>
</tr>
<tr>
<td>a</td>
<td>G_{case,a}</td>
<td>G_{control,a}</td>
</tr>
<tr>
<td>Total</td>
<td>2xN_{case}</td>
<td>2xN_{control}</td>
</tr>
</tbody>
</table>

- $\chi^2$ test with 1df
Manhattan Plot of p-values from Breast Cancer GWAS

- Analysis of 582,886 SNPs for 3,659 cases with family history and 4,897 controls

Often $-\log_{10} (p\text{-value})$ is plotted instead of p-value.
Single SNP Association Test: Continuous-valued Traits

- Continuous-valued traits
  - Also called quantitative traits
  - Cholesterol level, blood pressure etc.

- For each locus, fit a linear regression at each locus
  \[ y_i = x_i \beta + \varepsilon \]

- \( t \)-test with null hypothesis “No associations, i.e., \( \beta = 0 \)”
Genetic Model for Association

• Additive effect of minor allele, assuming effect size $a$ for each minor allele
  – Major allele homozygote: 0
  – Heterozygote: $a$
  – Minor allele homozygote: $2a$

• Generalizing additive genetic models for heterozygotes: $a + a \times k$
  – $k=1$: dominant effect of the minor allele
  – $k=0$: no dominance
  – $k=-1$: dominant effect of the major allele

• Penetrance
  – Proportions of individuals carrying a particular allele that possess an associated trait
  – Alleles with high penetrance are easier to detect
Correcting for Multiple Testing

• What happens when we scan the genome of 1 million genetic markers for association with $\alpha = 0.05$?
  – 50,000 (=1 million x 0.05) SNPs are expected to be found significant just by chance
  – We need to be more conservative when we decide a given marker is significantly associated with the trait.

• Correction methods
  – Bonferroni correction
  – Permutation test
Bonferroni Correction

- If N markers are tested, we correct the significance level as $\alpha' = \frac{\alpha}{N}$
  - Assumes the N tests are independent, although this is not true because of the linkage disequilibrium.
  - Overly conservative for tightly linked markers
Permutation Procedure

- In order to generate the null distribution
  - Step 1: Set $N_{\text{sig}} = 0$
  - Step 2: Repeat 1:$N_{\text{perm}}$
    - Step 3a: Randomly permute the individuals in the phenotype data to generate datasets with no association (retain the original genotype)
    - Step 3b: Find the test statistics $T_{\text{perm}}$ of SNPs using the permuted dataset
      - $T_1, ..., T_{N_{\text{perm}}}$ form a null distribution

- Compute the test statistic $T$ using the original dataset and test with the above null distribution

This approach is computationally demanding because often a large $N_{\text{perm}}$ is required.
Sparse regression method to evaluate the effect of each SNP in the context of all other SNPs

\[ y = X \beta + \varepsilon \]

• Sparsity constraint: Only few SNPs are influencing the given phenotype and the rest of the SNPs have effect size 0, no multiple-hypothesis-testing problem
L1 Regularization (LASSO)

- A convex relaxation.

**Constrained Form**
\[
\hat{\beta} = \arg\min_{\beta} \| Y - X\beta \|^2 \quad \text{subject to:} \quad \sum_{j=1}^{p} |\beta_j| \leq C
\]

**Lagrangian Form**
\[
\hat{\beta} = \arg\min_{\beta} \| Y - X\beta \|^2 + \lambda \|\beta\|_1
\]

- Still enforces sparsity!
# Lasso for Reducing False Positives

<table>
<thead>
<tr>
<th>Trait</th>
<th>Genotype</th>
<th>Association Strength</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>=</td>
<td>G ≡ A A C G A A A T G TA</td>
</tr>
</tbody>
</table>

\[
\text{argmin}_{\beta} (y - X\beta)^T(y - X\beta) + \lambda \sum_j |\beta_j|
\]

Many zero associations *(sparse results)*

Wu et al., Bioinformatics 2009
Multi-marker (Haplotype) Association Test

• Idea: a haplotype of multiple SNPs is a better proxy for a true causal SNP than a single SNP

• Form a new allele by combining multiple SNPs for a haplotype

<table>
<thead>
<tr>
<th>SNP A</th>
<th>SNP B</th>
<th>Auxiliary Markers for Haplotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1 0 0 0 0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0 1 0 0 0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0 0 1 0 0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0 0 0 1 0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 0 0 0 1</td>
</tr>
</tbody>
</table>

• Test the haplotype allele for association
Multi-marker Association Test

- Multi-marker approach can capture dependencies across multiple markers
  - SNPs in LD form a haplotype that can be tested as a single allele
  - Can achieve the higher power
    - Haplotypes are more powerful discriminators between cases and controls in disease association studies

- Challenge as the size of haplotype increases
  - Haplotype of $K$ SNPs results in $2^K$ different haplotypes, but the number of samples corresponding to each haplotype decreases quickly as we increase $K$
  - Large $K$ requires a large sample size
Overview

• Statistical methods for testing genotype/phenotype associations
  • Discrete-valued phenotype: case/control study
  • Continuous-valued phenotype: quantitative traits
  • Sparse regression method for considering all of the SNP markers
  • Multimarker association test

• Issues arising in GWAS
  • Genotype imputation
  • From common to rare variants
  • Epistasis for multiple interacting loci
  • Correcting for population structure
Causal Mutations and Genetic Markers

With SNP array data:

Unknown
Causal
Mutation

Known (genotyped)
SNP Marker

Linkage
Disequilibrium

• What happens when SNP density increases?
• Fine mapping required to locate the causal mutation
• What happens with whole genome sequencing data?
Increasing SNP Density via Genotype Imputation

- Reference data: dense SNP data from HapMap III, or 1000 genome project

- New data: SNP data for individuals in a given study

- Data after imputation with the reference data (leverage LD!)
Genotype Imputation

Reference set of haplotypes, for example, HapMap

Genotype data with missing data at untyped SNPs (grey question marks)

Each sample is phased and the haplotypes are modelled as a mosaic of those in the haplotype reference panel

The reference haplotypes are used to impute alleles into the samples to create imputed genotypes (orange)

PHASE can be used for imputation!
Imputation-Based Methods
(Servin & Stephens, 2007)
Common Variants vs. Rare Variants

• First-generation genome-wide association study (GWAS): common variant common disease hypothesis

• Common variants with minor allele frequency (MAF)>5%
  – dbGap: ~11 million SNPs
  – HapMap: 3.5 million SNPs
  – A successful GWAS requires a more complete catalogue of genetic variations

• Rare variants (MAF<0.5%), low-frequency variants (MAF:0.5%~5%)
  – Captured by sequencing with next-generation sequencing technology
  – Possibly significant contributors to the genetic architecture of disease
    • Causal variants are subject to negative selection
**Associations to Rare Variants**

- Often GWA studies are underpowered for functional rare variants

<table>
<thead>
<tr>
<th>Common Variant Association</th>
<th>Case</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allele a</td>
<td>60</td>
<td>20</td>
</tr>
<tr>
<td>Allele A</td>
<td>40</td>
<td>80</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rare Variant Association</th>
<th>Case</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allele a</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>Allele A</td>
<td>93</td>
<td>98</td>
</tr>
</tbody>
</table>

- Common variant GWA approaches are appropriate only for common variants
Feasibility of Identifying Disease Loci

- Rare alleles causing Mendelian disease
- Low-frequency variants with intermediate effect
- Rare variants of small effect very hard to identify by genetic means
- Common variants implicated in common disease by GWA

- Few examples of high-effect common variants influencing common disease
Epistasis

• Definition: The effect of one locus depends on the genotype of another locus
  – Epistatic effects vs. marginal effects
Epistasis for Mendelian Traits

Dominant epistasis (Mendelian)

<table>
<thead>
<tr>
<th>Dominant white genotype (K/I)</th>
<th>Extension genotype (MC1R)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EE</td>
<td>Ee</td>
</tr>
<tr>
<td>ii</td>
<td></td>
</tr>
<tr>
<td>(\bar{i})</td>
<td></td>
</tr>
<tr>
<td>(i)</td>
<td></td>
</tr>
<tr>
<td>(\bar{i})</td>
<td></td>
</tr>
</tbody>
</table>

Carlborg & Haley, Nature Reviews Genetics 2004
When There is No Epistasis

- Two additive (non-epistatic) loci
- The three lines run in parallel
Epistasis Example

- **Dominant epistasis**
- One locus in a dominant way suppresses the allelic effects of a second locus

- **Co-adaptive epistasis**
- Genotypes that are homozygous for alleles of the two loci that originate from the same line (JJ with JJ, or LL with LL) show enhanced performance.
- Almost no marginal effects: average effect of JJ, JL, LL do not differ

- **Dominance-by-dominance epistasis**
- Double heterozygote (LS, LS) deviates from the phenotype that is expected from the phenotypes of the other heterozygotes.
- Double heterozygotes have a lower phenotype than expected.
Epistatic and Individual QTLs
Detecting Epistasis

• Epistatic effects of SNPs can often be detected only if the interacting SNPs are considered jointly
  – The number of candidate SNP interactions is very large
    • For $J$ SNPs, $J \times J$ SNP pairs need to be considered for epistasis
    • In general for $J$ SNPs and $K$-way interactions, there are $O(J^K)$ candidate interactions
    • Computationally expensive to consider all possible groups of interacting SNPs
    • For a reliable detection of $K$-way interactions, a large sample size is required
  – Multiple testing problem
Population Structure and Association Analysis

- Population structure in data causes false positives
  - Samples in the case population are usually more related
  - Any SNPs more prevalent in the case population will be found significantly associated with the trait.
Accounting for Population Structure in Association Analysis

• Needs to account for population structure in association mapping.

• Careful study design with each population represented in case/control groups in a balanced way.
  – Can be hard to control for population structure during data collection
  – The effect of cryptic population structure
Family-based Design vs. Population-based Design

• Family-based studies
  – The effect of population structure can be controlled by the use of parents’ genotypes (e.g., Transmission disequilibrium test (TDT))
  – In practice, collecting genotypes from multiple individuals in a family can be hard. (e.g., late-onset diseases)

• Population-based design
  – Data collection is easier for a large number of unrelated individuals than families.
  – The control samples can be reused in different studies.
Accounting for Population Structure in Association Analysis

- Population-based method
  - Genomic control (Devlin & Roeder, Biometrics 1999)
    - Use the SNPs that are not associated with the trait to remove the effect of population stratification
    - Ignores admixture
  - Structured association (Pritchard et al., AJHG 2000)
    - First run STRUCTURE on genotype data. Within each subpopulation, an association between a genetic marker and the trait is a true association.
  - EigenStrat: principal component analysis (Price et al., Nature Genetics 2006)
    - First run PCA on genotype data to infer the population structure. Perform association analysis after correcting for the population effects in genotype/phenotype data
- Linear mixed model (Lippert et al., Nature Methods 2011)
  - Model the population effects with random effects