Epistasis

02-715 Advanced Topics in Computational Genomics
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/T)</th>
<th>Extension genotype (MC1R)</th>
</tr>
</thead>
<tbody>
<tr>
<td>II</td>
<td>EE</td>
</tr>
<tr>
<td>ii</td>
<td>Ee</td>
</tr>
<tr>
<td>ii</td>
<td>ee</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
Epistasis Example

- 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
Epistasis Example

- 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.
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
BEAM Overview (Zhang and Liu, 2007)

• Bayesian epistasis association mapping for case/control studies

• Bayesian partitioning of the genetic markers to groups of markers with/without associations

• Use MCMC to learn the partitioning and obtain the posterior probability of partitions of markers

• It can handle up to ~100,000 markers
Assume the markers in the case group belong to Group 0, 1, or 2.
- Group 0: Markers with no effects $D_1$
- Group 1: Markers with marginal effects $D_2$
- Group 2: Markers with epistatic effects $D_3$

Markers in the control group $U$ belong to Group 0.

Goal: Learn the partition of markers $I$ into Groups 0, 1, 2, given the genotype data.
Bayesian Marker Partition Model

- Model for markers in case group with marginal effects
  \[ P(D_1|\Theta_1) = \prod_{j:I=1}^{3} \prod_{k=1}^{3} \theta_{jk}^{n_{jk}} \]
  - Assume a Dirichlet(\alpha) for \( \Theta_1 \)

- Marginal likelihood after integrating out parameters \( \Theta_1 \)
  \[ P(D_1|I) = \prod_{j:I=1}^{3} \left( \left( \prod_{k=1}^{3} \frac{\Gamma(n_{jk}+\alpha_k)}{\Gamma(\alpha_k)} \right) \frac{\Gamma(\alpha)}{\Gamma(N_d+\alpha)} \right) \]
Bayesian Marker Partition Model

• Model for markers in case group with epistatic effects

\[
P(D_2|I) = \left( \prod_{k=1}^{3^L} \frac{\Gamma(n_k+\beta_k)}{\Gamma(\beta_k)} \right) \frac{\Gamma(|\beta|)}{\Gamma(N_d+|\beta|)}
\]

• Markers with no effects in case group, and markers in the control group

\[
P(D_0, U|I) = \prod_{j=1}^{L} \left( \left( \prod_{k=1}^{3} \frac{\Gamma(n_{jk}+m_{jk}+\gamma_k)}{\Gamma(\gamma_k)} \right) \frac{\Gamma(|\gamma|)}{\Gamma\left(\sum_{k=1}^{3} (n_{jk}+m_{jk})+|\gamma|\right)} \right)
\]
Bayesian Marker Partition Model

• The posterior distribution of marker assignment to Groups 0, 1, 2 is given as

\[ P(I|D, U) \propto P(D_1|I)P(D_2|I)P(D_0, U|I)P(I) \]

  - The prior is given as

\[ P(I) \propto p_1^{l_1} p_2^{l_1} (1 - p_1 - p_2)^{L-l_1-l_2} \]

\[ p_1 = p_2 = 0.01 \]
BEAM: MCMC Sampling

- Initialize I according to P(I)
- Metropolis-Hastings (MH) algorithm
  - Propose to change the marker’s group membership
  - Propose to randomly exchange two markers between Groups 0, 1, 2.
Evaluating Markers for Associations

• Use the posterior probability $P(I|D,U)$

• Use B-statistic

$$B_M = \ln \frac{P_A(D_M, U_M)}{P_0(D_M, U_M)} = \ln \frac{P_{\text{join}}(D_M)(P_{\text{ind}}(U_M) + P_{\text{join}}(U_M))}{P_{\text{ind}}(D_M, U_M) + P_{\text{join}}(D_M, U_M)}$$

- $P_0(D_M, U_M)$ Bayes factor under null model
- $P_A(D_M, U_M)$ Bayes factor under the alternative model
Simulation Study

- Model 1: two disease loci with independent effects
- Model 2: disease risk only when both loci have at least one disease allele
- Model 3: additional disease alleles at each locus do not further increase the disease risk
Simulation Study

• Other scenarios to be considered
  – Model 4: three disease loci.
  – Model 5: multiple causal epistasis by a mixture of two two-way interactions. Disease risk if at least one epistatic interaction is present.
  – Model 6: six-way interaction
Results for Models 1-3

Model 1 ($\lambda = 0.3, r^2 = 0.7$)

Model 1 ($\lambda = 0.3, r^2 = 1.0$)

Model 2 ($\lambda = 0.2, r^2 = 0.7$)

Model 2 ($\lambda = 0.2, r^2 = 1.0$)

Model 3 ($\lambda = 0.2, r^2 = 0.7$)

Model 3 ($\lambda = 0.2, r^2 = 1.0$)

- 2000 cases, 2000 controls
- 1000 cases, 1000 controls

power

<table>
<thead>
<tr>
<th>MAF</th>
<th>MAF</th>
<th>MAF</th>
<th>MAF</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>0.1</td>
<td>0.2</td>
<td>0.5</td>
</tr>
</tbody>
</table>
Results on AMD Dataset

- AMD dataset, 116,204 SNPs genotyped for 96 cases and 50 controls. Posterior probabilities for having an association
Sensitivity to Prior Distributions

$p_1 = p_2 = 0.0033$

$p_1 = p_2 = 0.001$

$p_1 = p_2 = 0.0001$

$p_1 = p_2 = 0.00001$

P1: prior probabilities for marginal effects
P2: prior probabilities for epistatic effects
MCMC Trace and Autocorrelation Plot

Simulated data

AMD data
Screen and Clean (Wu et al., 2010)

• Multivariate analysis of all SNPs and SNP interactions using lasso

• Assume that the SNPs with epistatic effects are likely to have at least small marginal effect
  – Select the SNPs with marginal effects.
  – Consider only those SNPs with marginal effects for epistatic interaction.
  – SNPs will be found to have epistatic effects but no marginal effects
Screen: Two-stage Lasso for Detecting Epistasis

- **Two-stage lasso**
  - **Step 1**: Apply lasso with no consideration of epistasis to detect SNPs with significant individual effects
    \[
g(E[Y|X]) = \beta_0 + \sum_{j=1}^{L} \beta_j X_j
\]
  - **Step 2**: Apply lasso with pairs of only those SNPs selected in Step 1
    \[
g(E[Y|X]) = \beta_0 + \sum_{j=1}^{L} \beta_j X_j + \sum_{i<j; i,j=1,...,L} \beta_{ij} X_i X_j
\]
Clean: Assessing Significance of Results

• Split the data into Stage 1 and 2 datasets
  – Using Stage 1 data, apply lasso to select SNPs with marginal and epistatic effects
  – Using Stage 2 data, assess the significance of SNPs
    • Apply the least squared error method and obtain the traditional t-statistic

• Multi-split method
  – Randomly split the data into Stage 1 and Stage 2 datasets multiple times
  – Perform the two-way split method for each split
  – Combine the p-values from each split.
Advantages and Disadvantages

• Reduces the computational burden.

• SNPs with epistatic effects often do not have detectable individual (marginal) effects, and many of these association signals will be missed.

• Need to split the data into two parts for Stage 1 and Stage 2 analysis
Screen and Clean

\[ X_1 \ldots X_{15} \]

\[ X_3 \quad X_5 \quad X_6 \quad X_{10} \quad X_{11} \quad X_{14} \quad X_{15} \]

Main + Interactions

\[ X_3 X_5 \quad X_6 X_9 \quad X_9 X_10 \quad X_6 X_{10} \quad X_9 X_{11} \quad X_{10} X_{11} \quad X_3 X_{14} \quad X_9 X_{14} \quad X_6 X_{14} \quad X_{10} X_{14} \quad X_{11} X_{14} \quad X_9 X_{15} \quad X_6 X_{15} \quad X_{10} X_{15} \quad X_{11} X_{15} \quad X_{14} X_{15} \]

Screen

\[ X_9 X_6 \quad X_3 X_{11} \quad X_9 X_{11} \quad X_{10} X_{11} \quad X_{10} X_{14} \]

Clean

\[ X_9 X_6 \quad X_{10} X_{11} \]
Simulation Results

- Solid line: multi-split, dotted line: single-split

M1. $Y = \beta X_5 X_5 + \varepsilon$
M2. $Y = \beta (X_5 X_6 + X_{10} X_{11}) + \varepsilon$
M3. $Y = \beta (X_5 X_6 + 0.8 X_{10} X_{11} + 0.6 X_{15} X_{16} + 0.4 X_{20} X_{21} + 0.2 X_{25} X_{26}) + \varepsilon$. 
Results

- WTCCC data
  - 1,963 cases with T1D and 2,938 controls
  - Multi-split screen and clean with 56 random splits

<table>
<thead>
<tr>
<th>Chr</th>
<th>Position (bp)</th>
<th>SNP</th>
<th>Chr</th>
<th>Position (bp)</th>
<th>SNP</th>
<th>SC</th>
<th>Univariate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>p-value</td>
<td>p-value</td>
</tr>
<tr>
<td>6</td>
<td>3291181</td>
<td>rs241429</td>
<td>6</td>
<td>32910181</td>
<td>rs241432</td>
<td>3.5 x 10^{-66}</td>
<td>8.2 x 10^{-5}</td>
</tr>
<tr>
<td>6</td>
<td>3291181</td>
<td>rs241429</td>
<td>6</td>
<td>32712350</td>
<td>rs9272346</td>
<td>1.9 x 10^{-25}</td>
<td>8.2 x 10^{-5}</td>
</tr>
<tr>
<td>12</td>
<td>110918887</td>
<td>rs11066119</td>
<td>12</td>
<td>110971201</td>
<td>rs17696736</td>
<td>2.3 x 10^{-15}</td>
<td>9.6 x 10^{-5}</td>
</tr>
</tbody>
</table>