Lecture 10: Gene Environment Interactions, Meta Analysis, Emerging Issues

Timothy Thornton and Michael Wu

Summer Institute in Statistical Genetics 2015
Lecture Outline

Yet more on rare variants...

- Gene-Environment Interaction Testing
- Meta-analysis
- Additional Concerns
Gene-Environment Interactions ($G \times E$)

Complex diseases are caused by interplay of genes & environment.
Identification of $G \times E$ aids in disease prevention.
GxE Association Testing

Objective: Identify statistical interactions (synergism/antagonism) between environmental variable and rare variants in sequencing studies

Standard Approach:

- Test SNV individually
- Regress outcome on single variant, environment, and interaction
- Under-powered for rare variant analysis! (possibly worse than main effects)

How do we conduct region based analysis of GxE interactions?
Motivation

- Circulating levels of adiponectin are highly heritable and associated with many conditions.
- SNVs at the adiponectin-encoding gene, ADIPOQ, are associated with adiponectin levels.
- Dataset consists of adiponectin levels and rare SNVs (MAF < 5%) within the ADIPOQ gene.
Notation

Consider the following generalized linear model:

\[ g(\mu_i) = X_i^T \alpha_1 + \alpha_2 E_i + G_i^T \alpha_3 + E_i G_i^T \beta = \tilde{X}_i^T \alpha + S_i^T \beta. \]

- Outcome: \( Y_i \), has distribution from exponential family and \( \mu_i = E(Y_i|X_i) \).
- \( q \) non-genetic covariates: \( X_i \).
- Environmental factor: \( E_i \).
- Group of \( p \) variants: \( G_i = (G_{i1}, \ldots, G_{ip})^T \).
- \( p \) \( G \times E \) interaction terms: \( S_i = (E_i G_{i1}, \ldots, E_i G_{ip})^T \).

We are interested in testing if there is any \( G \times E \):

\[ H_0 : \beta = 0. \]
Collapsing tests

Intuition behind collapsing tests

- General problem: $p$ is large and $G_1, \cdots, G_p$ are rare.
- Solution: for each individual $i$, summarize rare SNV-set $(G_{i1}, \cdots, G_{ip})$ using a single summary variable and conduct inference using this single summary variable.
- For example, define “collapsing” variable as $G_i^* = \sum_{k=1}^{p} G_{ik} =$ Total No. of rare alleles.
Collapsing Tests for Interactions

To test for main effects:

\[ H_{1m} : g(\mu_i) = \alpha_1^* + \alpha_2^* E_i + \alpha_3^* G_i^* \]
\[ H_{0m} : \alpha_3^* = 0 \]

Can we use it to test for interactions?

\[ H_{1x} : g(\mu_i) = \alpha_1^* + \alpha_2^* E_i + \alpha_3^* G_i^* + \beta^* E_i G_i^* \]
\[ H_{0x} : \beta^* = 0 \]
Bias analysis for Collapsing $G \times E$ tests

**Intuition**

Null model has to be correctly specified for valid inference. Collapsing $G \times E$ tests may not give valid inference as main effects of the SNVs may not be sufficiently accounted for.

**Continuous Outcome: No, even if $G \perp E$.**

- **$G$ and $E$ are independent:**
  
  Model for mean of $Y$ is valid;
  Model for variance of $Y$ is not valid.

- **$G$ and $E$ not independent:**
  
  Model for mean of $Y$ is not valid;
  Model for variance of $Y$ is not valid.
Binary Outcome: Yes if disease is rare and $G \perp E$.

- $G$ and $E$ are independent:
  Model for mean of $Y$ is valid;
  Model for variance of $Y$ is valid approximately.

- $G$ and $E$ not independent:
  Model for mean of $Y$ is not valid;
  Model for variance of $Y$ is valid approximately.
iSKAT: Model

To test if there is any $G \times E$ ($H_0 : \beta = 0$):

$$H_0 : \text{logit}[P(Y_i = 1|E_i, X_i, G_i)] = X_i^T \alpha_1 + \alpha_2 E_i + G_i^T \alpha_3$$

$$H_A : \text{logit}[P(Y_i = 1|E_i, X_i, G_i)] = X_i^T \alpha_1 + G_i^T (\alpha_3 + E_i \beta) + \alpha_2 E_i$$

In principle, we can do the same thing as with SKAT, but ...

Difficulties

Need to fit null model:

- Need to estimate main effect of variants
- Lots of variants
- LD and rarity make fitting difficult

Modifications are necessary.

iSKAT: Extension of SKAT for GxE
iSKAT: Test Statistic

- Assume \((\beta_1, \cdots, \beta_p)^T\) are random and independent with mean zero and common variance \(\tau\).
- Testing \(H_0\) reduces to testing \(H_0 : \tau = 0\).
- Following Lin (1997), the score test statistic is
  \[ T = (Y - \hat{\mu})^T SS^T (Y - \hat{\mu}) = [Y - \mu (\hat{\alpha})]^T SS^T [Y - \mu (\hat{\alpha})]. \]
- \(\hat{\mu} = \mu (\hat{\alpha})\) is estimated under the null model,
  \[ g (\mu_i | X_i, E_i, G_i) = X_i^T \alpha_1 + \alpha_2 E_i + G_i^T \alpha_3 = \tilde{X}_i^T \alpha. \]
- Use ridge regression to estimate \(\alpha\), impose a penalty only on \(\alpha_3\).
- Under \(H_0\), \(T \sim \sum_{v=1}^{p} d_v \chi^2_1\) approximately.
- Invert characteristic function to get p-value (Davies, 1980).
iSKAT

Consider a GLMM framework:

\[ H_0 : g(\mu_i) = \alpha_1 + \alpha_2 E_i + G_i^T \alpha_3 \]
\[ H_1 : g(\mu_i) = \alpha_1 + \alpha_2 E_i + G_i^T \alpha_3 + [E_i G_i^T \beta] \]

- Let \( \beta_j \sim F(0, w_j^2 \tau) \) and let \( \beta \) have exchangeable correlation structure with pairwise correlation \( \rho \).
- \( \rho = 0 \) and \( \rho = 1 \) correspond to \( H_{1b} \) and \( H_{1a} \) respectively.
- For a fixed \( \rho \), a score test statistic for testing \( H_0 : \tau = 0 \) is:

\[ Q_\rho = (Y - \hat{\mu})^T SW [\rho 11^T + (1 - \rho) I] WS^T (Y - \hat{\mu}) = \rho Q_{1a} + (1 - \rho) Q_{1b} \]

- Find optimal \( \rho \) to maximize \( Q_\rho \):

\[ Q_{iSKAT} = \min_{0 \leq \rho \leq 1} p_\rho, \]

where \( p_\rho \) is the p-value computed based on \( Q_\rho \).
Size Simulations: iSKAT vs. Collapsing $G \times E$ test

- Bootstrap dataset samples to obtain genotypes/covariates.
- Main effects of SNVs chosen to mimic dataset.
- Simulate outcome under null hypothesis to investigate the size.

Collapsing $G \times E$ test

iSKAT

![Graphs comparing iSKAT and Collapsing $G \times E$ test](image-url)
Data Application: Adiponectin Levels

- Adiponectin levels are associated with many diseases.
- SNVs at the adiponectin-encoding gene, ADIPOQ, are associated with adiponectin levels.
- Adiponectin levels from 1945 individuals.
- 11 rare SNVs within the exon region of ADIPOQ.
- Test for $G \times E$.

<table>
<thead>
<tr>
<th>Test</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>iSKAT</td>
<td>0.037</td>
</tr>
<tr>
<td>$\rho = 0$</td>
<td>0.23</td>
</tr>
<tr>
<td>$\rho = 1$</td>
<td>0.022</td>
</tr>
</tbody>
</table>
Rare variant Meta-analysis

- Meta-analysis is an effective approach to combine data from multiple studies.
- Rare variant meta-analysis: **desirable properties**
  - Use summary statistics
  - Same power as mega-analysis (joint analysis)
  - Account for varying levels of heterogeneity of genetic effects across studies.
Rare Variant Meta-Analysis References


Multi-Study Model

- For the $k^{th}$ study ($k = 1, \ldots, K$),
  - Genotype $G_{ki} = (g_{k1}, \ldots, g_{kp})'$
  - Covariates $X_{ki} = (x_{k1}, \ldots, x_{kiq_k})'$
  - Model:
    $$g(\mu_{ik}) = X_{ki} \alpha_k + G_{ki} \beta_k$$

- Test $H_0: \beta_k = 0$ ($k = 1, \ldots, K$)
Meta-Analysis for Rare Variants in Sequencing Association Studies

▶ Challenges
  ▶ Jointly analyze multiple SNPs in a region.
  ▶ Hard to estimate $\beta_k$ for rare variants
  ▶ Wald-based meta-analysis for vector $\beta_k$ is challenging
Meta-Analysis for Rare Variants in Sequencing Association Studies

- Idea of Meta-SKAT family tests:
  - Work with the scores of $\beta_k$ by fitting only null models.
  - Assume a distribution for $\beta_{kj}$ ($j = 1, \cdots, p$)
  - Perform variance component score test by allowing homogeneous and heterogeneous genetic effects across studies.
Single study $k$

- Score statistic of variant $j$
  \[
  S_{kj} = \sum_{i=1}^{n} g_{ijk}(y_{ij} - \hat{\mu}_{ij})/\hat{\phi}_k
  \]

- SKAT and Burden test statistics:
  \[
  Q_{SKAT} = \sum_{j=1}^{p} (w_{jk}S_{kj})^2, \quad Q_{Burden} = \left(\sum_{j=1}^{p} w_{jk}S_{kj}\right)^2
  \]
Single study $k$

- SKAT-O (combined approach):

$$T = \min_{0 \leq \rho \leq 1} P_\rho$$

where $P_\rho$ is the p-value of

$$Q_\rho = (1 - \rho)Q_{SKAT} + \rho Q_{Burden}$$
Input Summary Statistics for meta-analysis

- Input summary statistics from each study
  - MAF
  - $S_{kj}$: score statistic of each marker
  - Between-variant relationship matrix ($p \times p$)

$$\Phi_k = G_k' P_k G_k,$$

where $P_k = V_k^{-1} - V_k^{-1} X_k (X_k' V_k^{-1} X_k)^{-1} X_k' V_k^{-1}$
Homogeneous genetic effects

- Assume the same SNP effects across studies.
  - \( \beta_1 = \beta_2 = \cdots = \beta_K = \beta \)
  - \( E(\beta_j) = 0, \text{var}(\beta_j) = w_j \tau \) and \( \text{cor}(\beta_j, \beta_{j'}) = \rho \).

- Derive the VC score test for \( H_0 : \tau = 0 \).
- Multivariate score-based analog of univariate fixed effect meta-analysis
Meta-SKAT: Homogeneous genetic effects

- Meta-SKAT assuming homogeneous genetic effects:

\[
Q_{\text{hom\_meta\_SKAT}} = \sum_{j=1}^{p} \left( \sum_{k=1}^{K} w_{kj} S_{kj} \right)^2
\]

- Meta-Burden:

\[
Q_{\text{meta\_Burden}} = \left( \sum_{j=1}^{p} \sum_{k=1}^{K} w_{kj} S_{kj} \right)^2
\]

- Meta-SKAT-O:

\[
Q_{\text{hom\_meta}}(\rho) = (1 - \rho)Q_{\text{hom\_meta\_SKAT}} + \rho Q_{\text{meta\_Burden}}
\]
Meta-SKAT: Homogeneous genetic effects

- Test statistics are essentially *identical* to those of the mega-analysis SKAT and burden test
  \[\Rightarrow\text{ As powerful as mega-analysis}\]
- \(P\)-values can be computed using the Davies method.
  \[\Rightarrow\text{ Fast computation}\]
- SKAT-O can be conducted with adaptively selecting \(\rho\).
Meta-SKAT: Heterogeneous genetic effects

- Assume genetic effects vary between studies
  - $\beta_1, \ldots, \beta_K$ are iid
  - $E(\beta_{kj}) = 0$, $\text{var}(\beta_j) = w_{kj}\tau$ and $\text{cor}(\beta_{kj}, \beta_{kj'}) = \rho$.
- Multivariate score-based analog of the univariate random effect model meta-analysis.
- P-values can be calculated analytically
- Useful for meta analysis of studies of the same ethnicity or different ethnicities.
Meta-SKAT: Heterogeneous genetic effects

- Meta-SKAT assuming heterogeneous genetic effects:

\[ Q_{\text{het-meta-SKAT}} = \sum_{j=1}^{p} \sum_{k=1}^{K} (w_{kj} S_{kj})^2 \]

- Meta-SKAT-O:

\[ Q_{\text{hom-meta}}(\rho) = (1 - \rho) Q_{\text{het-meta-SKAT}} + \rho Q_{\text{meta-Burden}} \]
Meta-SKAT for multi-ethnicities:

- Multi-ethnic studies:
  - within-group homogeneity and between-group heterogeneity
  - $\beta_k = \beta_l$ for the same group and $\beta_k \perp \beta_l$ for the different groups

- Meta-SKAT with $B$ ancestry groups

$$Q_{het\_meta\_SKAT} = \sum_{j=1}^{p} \sum_{b=1}^{B} \left( \sum_{k=k_{b-1}+1}^{k_b} w_{kj} S_{kj} \right)^2$$

- Meta-SKAT-O:

$$Q_{hom\_meta}(\rho) = (1 - \rho) Q_{het\_meta\_SKAT} + \rho Q_{meta\_Burden}$$
Simulation Studies

- 3kb randomly selected regions
- Three scenarios:
  - 1: homogeneous
  - 2: moderately heterogeneous (studies share 50% of causal variants)
  - 3: two different ancestry groups (study 1,2: EUR and study 3: AA)
Powers comparison: Meta vs Mega (Joint)

![Graph A](image1)

Power of Joint analysis SKAT vs Power of Hom-Meta-SKAT

![Graph B](image2)

Power of Joint analysis SKAT-O vs Power of Hom-Meta-SKAT-O
Simulation Results, $\beta + / - = 100/0$
### Analysis of the lipid data (LPL gene)

- 11,000 samples from 7 studies.
- Adjusted for the index SNP
- For HDL and TG, **Het-Meta-SKAT-O** achieved the smallest p-values

<table>
<thead>
<tr>
<th>Traits</th>
<th>Hom-Meta SKAT-O</th>
<th>Het-Meta SKAT-O</th>
<th>Meta-Fisher</th>
<th>Meta-Burden</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDL</td>
<td>$2.5 \times 10^{-2}$</td>
<td>$1.2 \times 10^{-4}$</td>
<td>$1.7 \times 10^{-2}$</td>
<td>$3.5 \times 10^{-1}$</td>
</tr>
<tr>
<td>LDL</td>
<td>1.00</td>
<td>$4.0 \times 10^{-1}$</td>
<td>$3.9 \times 10^{-1}$</td>
<td>$2.1 \times 10^{-2}$</td>
</tr>
<tr>
<td>TG</td>
<td>$5.3 \times 10^{-3}$</td>
<td>$2.8 \times 10^{-5}$</td>
<td>$6.0 \times 10^{-4}$</td>
<td>$7.7 \times 10^{-2}$</td>
</tr>
</tbody>
</table>
Summary of Meta-Analyses for Rare Variants

- Based on study-specific summary score statistics
- As powerful as the joint analysis
- Flexible to accommodate a wide range of heterogeneity of genetic effects
Additional Concerns

▶ Quality control:
  ▶ Are the observed variants really variants?
  ▶ Batch effects
  ▶ Some standard pipelines now in place

▶ Population stratification:
  ▶ Common strategy: just use same PCs from common variant analysis to correct for PS
  ▶ Some evidence that rare variants require special accommodation (much larger number of PCs)

▶ Accommodating common variants:
  ▶ What do you do with common variants?
  ▶ (a) Assess joint effect with rare variants
  ▶ (b) Adjust for effect of common variants
Additional Concerns

- **Prediction**
  - In a new population (sample), we’re unlikely to see the same variants and we’re likely to see a lot of variants not previously observed

- **Prioritization of individual variants**
  - How to choose individual causal variants?
  - Some work on variable selection methods, but no ability to control type I error.
  - Bioinformatics and functionality tools may be useful

- Incorporation of functional information and other genomic data
Additional Concerns

▶ Design Choices
  ▶ Want to enrich for variants (extreme phenotypes)
  ▶ Some of these designs require specialized methods
  ▶ Stuck with the design chosen
▶ Dealing with admixed populations
▶ Related individuals
▶ Tim: what is a “rare variant”? 
▶ (Statistically) complex phenotypes