Lecture Overview

1. Variance Component Tests
2. Omnibus Tests
3. Weights
Recall: Region Based Analysis of Rare Variants

- Single variant test is not powerful to identify rare variant associations
- Strategy: Region based analysis
  - Test the joint effect of rare/common variants in a gene/region while adjusting for covariates.

Major Classes of Tests

- Burden/Collapsing tests
- Supervised/Adaptive Burden/Collapsing tests
- Variance component (similarity) based tests
- Omnibus tests: hedge against difference scenarios
Variance component test

- Burden tests are not powerful, if there exist variants with different association directions or many non-causal variants
- Variance component tests have been proposed to address it.
- “Similarity” based test
C-alpha test


- Case-control studies without covariates.
- Assume the $j$th variant is observed $n_{j1}$ times, with $r_{j1}$ times in cases.

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>A</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case</td>
<td>$r_{j1}$</td>
<td>$r_{j2}$</td>
<td>$r$</td>
</tr>
<tr>
<td>Control</td>
<td>$s_{j1}$</td>
<td>$s_{j2}$</td>
<td>$s$</td>
</tr>
<tr>
<td>Total</td>
<td>$n_{j1}$</td>
<td>$n_{j2}$</td>
<td>$n$</td>
</tr>
</tbody>
</table>

- Under $H_0$

$$r_{j1} \sim \text{Binomial}(n_{j1}, q) \quad (q = r/n)$$
C-alpha test

- Risk increasing variant:

\[ r_{j1} - qn_{j1} > 0 \]

- Risk decreasing variant:

\[ r_{j1} - qn_{j1} < 0 \]

- Test statistic:

\[ T_\alpha = \sum_{j=1}^{p} (r_{j1} - qn_{j1})^2 - \sum_{j=1}^{p} n_{j1}q(1 - q) \]

- This test is robust in the presence of the opposite association directions.
C-alpha test

- Weighting scheme

\[ T_\alpha = \sum_{j=1}^{p} w_j (r_{j1} - qn_{j1})^2 - \sum_{j=1}^{p} w_j n_{j1} q(1 - q) \]

- Test for the over-dispersion due to genetic effects
  - Neyman’s \( C(\alpha) \) test.
C-alpha test, P-value calculation

- Using normal approximation, since the test statistic is the sum of random variables.

\[
T_\alpha = \sum_{j=1}^{p} (r_{j1} - qn_{j1})^2 - \sum_{j=1}^{p} n_{j1}q(1 - q)
\]

- Doesn’t work well when \( p \) is small (or moderate).
  - P-value is computed using permutation.
C-alpha test

- C-alpha test is robust in the presence of the different association directions
- Disadvantages:
  - Permutation is computationally expensive.
  - Cannot adjust for covariates.
Sequence Kernel Association Test (SKAT)

Wu et al. (2010, 2011). AJHG

Recall the original regression models:

$$\mu_i / \logit(\mu_i) = \alpha_0 + X_i^T \alpha + G_i^T \beta$$

Variance component test:

- Assume $\beta_j \sim \text{dist.}(0, w_j^2 \tau)$.
- $H_0 : \beta_1 = \cdots = \beta_p = 0 \iff H_0 : \tau = 0$. 
Sequence Kernel Association Test (SKAT)

- $\beta_j \sim \text{dist.}(0, w_j^2 \tau)$: $\tau = 0$ is on the boundary of the hypothesis.
- Score test statistic for $\tau = 0$:
  \[ Q_{SKAT} = (y - \hat{\mu}_0)'K(y - \hat{\mu}_0), \]
- $K = GWWG'$: weighted linear kernel ($W = \text{diag}[w_1, \ldots, w_p]$).
Sequence Kernel Association Test (SKAT)

- The C-alpha test is a special case of SKAT
  - With no covariates and flat weights:
    \[
    Q_{SKAT} = \sum_{j=1}^{p} (r_{j1} - qn_{j1})^2
    \]
SKAT

- $Q_{SKAT}$ is a weighted sum of single variant score statistics

$$Q_{SKAT} = (y - \hat{\mu}_0)'GWWG'(y - \hat{\mu}_0)$$

$$= \sum_{j=1}^{p} w_j^2 [g_j'(y - \hat{\mu}_0)] = \sum_{j=1}^{p} w_j^2 U_j^2$$

where $U_j = \sum_{i=1}^{n} g_{ij} (y_i - \hat{\mu}_0i)$.

- $U_j$ is a score of individual SNP $j$ only model:

$$\mu_i / \text{logit}(\mu_i) = \alpha_0 + X_i^T \alpha + g_{ij} \beta_j$$
SKAT

- $Q_{SKAT}$ (asymptotically) follows a mixture of $\chi^2$ distribution under the NULL.

$$Q = (y - \hat{\mu}_0)'K(y - \hat{\mu}_0)$$
$$= (y - \hat{\mu}_0)'\hat{V}^{-1/2}\hat{V}^{1/2}K\hat{V}^{1/2}\hat{V}^{-1/2}(y - \hat{\mu}_0)$$
$$= \sum_{j=1}^{p}\lambda_j [u_j'\hat{V}^{-1/2}(y - \hat{\mu}_0)]^2$$
$$\approx \sum_{j=1}^{p}\lambda_j \chi^2_{1,j}$$
SKAT

- $\lambda_j$ and $u_j$ are eigenvalues and eigenvectors of $P^{1/2}KP^{1/2}$.

where $P = \hat{V}^{-1} - \hat{V}^{-1}\tilde{X}(\tilde{X}'\hat{V}^{-1}\tilde{X})^{-1}\tilde{X}'\hat{V}^{-1}$ is the project matrix to account that $\alpha$ is estimated.
SKAT: P-value calculation

- P-values can be computed by **inverting the characteristic function** using Davies’ method (1973, 1980)
  - Characteristic function
    \[ \varphi_x(t) = E(e^{itx}). \]
  - Characteristic function of \( \sum_{j=1}^{p} \lambda_j \chi^2_{1,j} \)
    \[ \varphi_x(t) = \prod_{i=j}^{p} (1 - 2\lambda_j it)^{-1/2}. \]
  - Inversion Formula
    \[ P(X < u) = \frac{1}{2} - \frac{1}{\pi} \int_{0}^{\infty} \frac{Im[e^{-itu} \varphi_x(t)]}{t} \, dt. \]
Small sample adjustment

Lee et al. (2012). AJHG

- When the sample size is small and the trait is binary, asymptotics does not work well.
- SKAT test statistic:

\[ Q_{SKAT} = (y - \hat{\mu}_0)'K(y - \hat{\mu}_0) \]

\[ = \sum_{v=1}^{p} \lambda_v \eta_v^2, \]

- \( \eta_v \)'s are asymptotically independent and follow N(0,1).
Small sample adjustment

- When the trait is binary and the sample size is small:
  - $\text{Var}(\eta_v) < 1$.
  - $\eta_v$s are negatively correlated.
Small sample adjustment

- Mean and variance of the $Q_{SKAT}$

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large Sample</td>
<td>$\sum \lambda_j$</td>
<td>$\sum \lambda_j^2$</td>
</tr>
<tr>
<td>Small Sample</td>
<td>$\sum \lambda_j$</td>
<td>$\sum \lambda_j \lambda_k c_{jk}$</td>
</tr>
</tbody>
</table>

- Adjust null distribution of $Q_{SKAT}$ using the estimated small sample variance.
Small sample adjustment

- Variance adjustment is not enough to accurately approximate far tail areas.
- **Kurtosis** adjustment:
  - Estimate the kurtosis of $Q_{SKAT}$ using parametric bootstrapping:
  - $\hat{\gamma}$ (estimated kurtosis)
  - D.F. estimator: $\hat{df} = 12/\hat{\gamma}$
  - Null distribution

\[
(Q_{SKAT} - \sum \lambda_j^2) \frac{\sqrt{2\hat{df}}}{\sqrt{\sum \lambda_j \lambda_k c_{jk}}} + \hat{df} \sim \chi_{\hat{df}}^2
\]
Small sample adjustment

Figure: ARDS data (89 samples)
General SKAT

- General SKAT Model:

\[ \frac{\mu_i}{\text{logit}(\mu_i)} = \alpha_0 + X_i \alpha + h_i \]

where \( h_i \sim GP(0, \tau K) \).

- Kernel \( K(G_i, G_i') \) measures genetic similarity between two subjects.
General SKAT

- Examples:
  - Linear kernel: linear effect
    \[ K(Z_i, Z_{i'}) = w_1^2 Z_{i1} Z_{i'1} + \cdots + w_p^2 Z_{ip} Z_{i'p} \]
  - IBS Kernel (Epistatic Effect: SNP-SNP interactions)
    \[ K(Z_i, Z_j) = \frac{\sum_{k=1}^{p} w_k^2 IBS(Z_{ik}, Z_{jk})}{2p} \]
Omnibus Tests

- Questions:
  - Which group of variants test? I.e. what is the threshold for “rare”?  
  - Which type of test should I use? Variance component or burden?

- Truth is unknown: depends on the situation
- Omnibus tests: work well across situation
Variable threshold (VT) test

- Most methods use a fixed threshold for rare variants: \( \leq 0.5\% \), \( \leq 1\% \), ..., \( \leq 5\% \)?

- Choosing an appropriate threshold can have a huge impact on power: prefer to restrict analysis to meaningful variants
Variable threshold (VT) test

Price AL, Kryukov GV, et al. (2010) AJHG

- Find the **optimal threshold** to increase the power.
  - Weight:
    
    \[
    w_j(t) = \begin{cases} 
    1 & \text{if } maf_j \leq t \\
    0 & \text{if } maf_j > t 
    \end{cases}
    \]

  - \( C_i(t) = \sum w_j(t)g_{ij} \)
  - Test statistics:
    
    \[ Z_{max} = \max_t Z(t) \]

where \( Z(t) \) is a Z-score of \( C_i \).
P-value Calculations of Variable threshold (VT) test

▶ Price *et al.* proposed to use *permutation* to get a p-value

▶ Lin and Tang (2011) showed that the p-values can be calculated through *numerical integration using normal approximation*
Variable threshold (VT) test

- More robust than using a fixed threshold.
- Provide information on the MAF ranges of the causal variants.
- Lose power if there exist variants with opposite association directions.
SKAT vs. Collapsing

- Collapsing tests are more powerful when a large % of variants are causal and effects are in the same direction.
- SKAT is more powerful when a small % of variants are causal, or the effects have mixed directions.
- Both scenarios can happen when scanning the genome.
- Best test to use depends on the underlying biology.
  → Difficult to choose which test to use in practice.

We want to develop a unified test that works well in both situations. → Omnibus tests
Combine p-values of Burden and SKAT

Derkach A et al. (2013) Genetic Epi, 37:110-121

- Fisher method:

\[ Q_{Fisher} = -2 \log(P_{Burden}) - 2 \log(P_{SKAT}) \]

- \( Q_{Fisher} \) follows \( \chi^2 \) with 4 d.f when these two p-values are independent
- Since they are not independent, p-values are calculated using resampling
- Mist (Sun et al. 2013) modified the SKAT test statistics to make them independent
Combine Test Statistics: Unified Test Statistics

Lee et al. (2012). *Biostatistics*

- Combined Test of Burden tests and SKAT

\[ Q_\rho = (1 - \rho)Q_{SKAT} + \rho Q_{Burden}, \quad 0 \leq \rho \leq 1. \]

- \( Q_\rho \) includes SKAT and burden tests.
  - \( \rho = 0 \): SKAT
  - \( \rho = 1 \): Burden
Derivation of the Unified Test Statistics

- **Model:**

\[ g(\mu_i) = X_i\alpha + G_i\beta \]

where \( \beta_j/w_j \) follows any arbitrary distribution with mean 0 and variance \( \tau \) and the correlation among \( \beta_j \)'s is \( \rho \).

- **Special cases:**
  - SKAT: \( \rho = 0 \)
  - Burden: \( \rho = 1 \)
  - Combined: \( 0 \leq \rho \leq 1 \)
Derivation of the Unified Test Statistics

$Q_{\rho}$ is a test statistic of the SKAT with $\text{corr}(\beta) = R(\rho)$:

- $R(\rho) = (1 - \rho)I + \rho \mathbb{1}\mathbb{1}'$ (compound symmetric)
- $K_{\rho} = GWR(\rho)WG'$.

\[
Q_{\rho} = (y - \hat{\mu})'K_{\rho}(y - \hat{\mu}) = (1 - \rho)Q_{SKAT} + \rho Q_{Burden}
\]
Adaptive Test (SKAT-O)

- Use the smallest p-value from different $\rho$s:

$$T = \inf_{0 \leq \rho \leq 1} P_\rho.$$

where $P_\rho$ is the p-value of $Q_\rho$ for given $\rho$.

- Test statistic:

$$T = \min P_{\rho_b}, \quad 0 = \rho_1 < \ldots < \rho_B = 1.$$
Adaptive Test (SKAT-O)

- $Q_\rho$ is a mixture of two quadratic forms.

\[
Q_\rho = (1 - \rho)(y - \hat{\mu})'GWWG'(y - \hat{\mu})' + \rho(y - \hat{\mu})'GW11'WG'(y - \hat{\mu})
= (1 - \rho)(y - \hat{\mu})'K_1(y - \hat{\mu})' + \rho(y - \hat{\mu})'K_2(y - \hat{\mu})
\]

- $Q_\rho$ is asymptotically equivalent to

\[
(1 - \rho)\kappa + a(\rho)\eta_0,
\]

where and $\eta_0 \sim \chi_1^2$, $\kappa$ approximately follows a mixture of $\chi^2$. 
SKAT-O

- $Q_\rho$ is the asymptotically same as the sum of two independent random variables.
  \[ (1 - \rho) \kappa + a(\rho) \eta_0 \]

- $\eta_0 \sim \chi^2_1$
- Approximate $\kappa$ via moments matching.

- P-value of $T$:
  \[
  1 - Pr \left\{ Q_{\rho_1} < q_{\rho_1}(T), \ldots, Q_{\rho_b} < q_{\rho_b}(T) \right\} \\
  = 1 - E \left[ Pr \left\{ (1 - \rho_1) \kappa + a(\rho_1) \eta_0 < q_{\rho_1}(T), \ldots | \eta_0 \right\} \right] \\
  = 1 - E \left[ P \left\{ \kappa < min\{(q_{\rho_V}(T)) - a(\rho_V) \eta_0)/(1 - \rho_V)\} | \eta_0 \right\} \right],
  \]

where $q_\rho(T) = \text{quantile function of } Q_\rho$
Simulation

- Simulate sequencing data using COSI
- 3kb randomly selected regions.
- Percentages of causal variants = 10%, 20%, or 50%.
- \((\beta_j > 0)\%\) among causal variants = 100% or 80%.
- Three methods
  - Burden test with beta(1,25) weight
  - SKAT
  - SKAT-O
Simulation

\[ \beta^{+/-}=100/0 \]

Causal = 10 %

Causal = 20 %

Causal = 50 %

\[ \beta^{+/-}=80/20 \]

Causal = 10 %

Causal = 20 %

Causal = 50 %
Simulation

- SKAT is more powerful than Burden test (Collapsing) when
  - Existence of $+/-\beta$s
  - Small percentage of variants are causal variants
- Burden test is more powerful than SKAT when
  - All $\beta$s were positive and a large proportion of variants were casual variants
- SKAT-O is robustly powerful under different scenarios.
Summary

- Region based tests can increase the power of rare variants analysis.
- Relative performance of rare variant tests depends on underlying disease models.
- The combined test (omnibus test), e.g., SKAT-O, is robust and powerful in different scenarios.
MAF based weighting

- It is generally assumed that rarer variants are more likely to be causal variants with larger effect sizes.
- Simple thresholding is widely used.

\[
w(MAF_j) = \begin{cases} 
1 & \text{if } MAF_j < c \\
0 & \text{if } MAF_j \geq c 
\end{cases}
\]
MAF based weighting

- Instead of thresholding, **continuous weighting** can be used to upweight rarer variants.
- Ex: Flexible beta density function.

\[ w(MAF_j) = (MAF_j)^{\alpha-1}(1 - MAF_j)^{\beta-1} \]

- \((\alpha = 0.5, \beta = 0.5)\) : Madsen and Browning weight
- \((\alpha = 1, \beta = 1)\) : Flat weight
MAF based weighting- beta weight

(A)

Weight

MAF

(B)

Weight

MAF

- Beta(1,1)
- Beta(1,25)
- Beta(1,50)
- Beta(0.5,0.5)
MAF based weighting- logistic weight

- Soft-thresholding.

\[ w(maf_j) = \frac{\exp((\alpha - maf_j)\beta)}{1 + \exp((\alpha - maf_j)\beta)} \]
Weighting Using Functional information

- Variants have different functionalities.
  - Non-synonymous mutations (e.g. missense and nonsense mutations) change the amino-acid (AA) sequence.
  - Synonymous mutations do not change AA sequence.
Weighting Using Functional information

- Bioinformatic tools to predict the functionality of mutations.
  - Polyphen2 (http://genetics.bwh.harvard.edu/pph2/)
  - SIFT (http://sift.jcvi.org/)
- Test only functional mutations can increase the power.