Review

We have covered so far:

- Single variant association analysis and effect size estimation
- GxE interaction and higher order $>$2 interaction
- Measurement error in dietary variables (nutritional epidemiology)
- Today's lecture: set-based association

http://research.fhcrc.org/hsu/en/edu.html

Set-based association analysis

- $i = 1, \cdots, n$
- m variants on a certain region
- Genotype $G_i = (G_{i1}, G_{i2}, \cdots, G_{im})', g_{ij} = 0, 1, 2$
- Covariates X_i : intercept, age, sex, principal components for population structure.
- Model:

$$g\{E(y_i)\} = X_i'\alpha + \sum_{j=1}^{m} G_{ij}\beta_j$$

  where $g(\cdot)$ is a link function.

- No association for a set means $\beta = (\beta_1, \cdots, \beta_m) = 0$

Why?

- In gene expression studies, a list of differentially expressed genes fails to provide mechanistic insights into the underlying biology to many investigators
- Pathway analysis extracts meaning by grouping genes into small sets of related genes
- Function databases are curated to help with this task, e.g., biological processes in which genes are involved in or interact with each other.
- Analyzing high-throughput molecular measurements at the functional or pathway level is very appealing
  - Reduce complexity
  - Improve explanatory power than a simple list of differentially expressed genes

Why?

- Variants in a defined set (e.g. genes, pathways) may act concordantly on phenotypes. Combining these variants aggregates the signals; as a result, may improve power
- Power is particularly an issue when variants are rare.
- The challenge is that not all variants in a set are associated with phenotypes and those who are associated may have either positive or negative effects

Outline

- Burden test
- Variance component test
- Mixed effects score test

Burden Test

- If m is large, multivariate test $\beta = 0$ is not very powerful
- Population genetics evolutionary theory suggests that most rare missense alleles are deleterious, and the effect is therefore generally considered one-sided (Kryukov et al., 2007)
- Collapsing: Suppose $\beta_1 = \cdots = \beta_m = \eta$

$$g\{E(y_i)\} = X_i'\alpha + B_i\eta$$

- $B_i = \sum_{j=1}^{m} g_{ij}$: genetic burden/score.
- With weight (adaptive burden test)

$$B_i = \sum_{j=1}^{m} w_j G_{ij}.$$

- Test $H_0 : \eta = 0$ (d.f.=1).

Weight

- Threshold-based method

$$w_j(t) = \begin{cases} 1 & \text{if } MAF_j \leq t \\ 0 & \text{if } MAF_j > t \end{cases}$$

- Burden score $B_i(t) = \sum_{j=1}^{m} w_j(t) g_{ij}$, and the corresponding Z-statistic is denoted by $Z(t)$
- Variable threshold (VT) test

$$Z_{\text{max}} = \text{max}_t Z(t)$$

- P-value can be calculated by permutation (Price et al 2010) or numerical integration using normal approximation.

$$\begin{aligned} P(Z_{\text{max}} \geq z) &= 1 - P(Z_{\text{max}} < z) \\ &= 1 - P(Z(t_1) < z, \cdots, Z(t_b) < z) \end{aligned}$$

where $\{Z(t_1), \cdots, Z(t_b)\}$ follows a multivariate normal distribution MVN$(0, \Sigma)$.

Weight

- Variant effects can be positive or negative and the strength can be different too.
- Fit the marginal model for each variant

$$g\{E(y_i)\} = X_i'\alpha + G_{ij}\gamma_j$$

Weight

- Adaptive sum test (Han & Pan, *Hum Hered* 2010)

$$w_j = \begin{cases} -1 & \text{if } \widehat{\gamma}_j < 0 \text{ and } p\text{-value} < \alpha_0; \\ 1 & \text{otherwise} \end{cases}$$

- If $\alpha_0 = 1$, the weight is the sign of $\widehat{\gamma}_j$, but the corresponding weighted burden test has low power because the null distribution has heavy tails.
- α_0 is chosen such that only when H_0 likely does not hold, the sign is changed if $\widehat{\gamma}_j$ is negative.
- The authors suggest $\alpha_0 = 0.1$, but it is data dependent

Weight

- Estimated regression coefficient (EREC) test (Lin & Tang *Am J Hum Genet* 2011)

$$w_j = \widehat{\gamma}_j + c, \qquad \text{for } c \neq 0$$

- Score statistic

$$T_{EREC} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{ (\sum_{j=1}^{m} (\widehat{\gamma}_j + c) G_{ij}) (y_i - \mu_i(\widehat{\alpha})) \} \longrightarrow N(0, \Sigma)$$

  - $\mu_i(\widehat{\alpha})$ is estimated under the null of no association
  - If $c = 0$, $T_{EREC} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{ (\sum_{j=1}^{m} \widehat{\gamma}_j G_{ij}) (y_i - \mu_i(\widehat{\alpha})) \}$ is not asymptotically normal

- $c = 1$ for binary traits, $c = 2$ for standardized quantitative traits

- Compute *p*-values using permutation.

Burden Tests

- Burden tests lose a significant amount of power if there are variants with different association directions or a large # of variants are neutral.
- Adaptive burdent tests have robust power, but they rely on resampling to compute *p*-values.
  - Computationally intensive, not suitable for genome-wide discovery.

Variance Component Test

- Model

$$g\{E(y_i)\} = X_i'\alpha + \sum_{j=1}^{m} G_{ij}\beta_j$$

- Burden tests are derived assuming $\beta_1 = \cdots = \beta_m$.
- Variance component test
  - Assume $\beta_j \sim F(0, \tau^2)$, where $F(\cdot)$ is an arbitrary distribution, and the mean of $\beta_j' s = 0$.
  - $H_0 : \beta_1 = \cdots = \beta_p = 0 \Leftrightarrow H_0 : \tau^2 = 0$.

Derivation of Variance Component Test

- Suppose $g(\cdot)$ is linear and $Y|X, G \sim$ Normal. That is,

$$y_i = X_i\alpha + \sum_{j=1}^{m} G_{ij}\beta_j + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

- Suppose $\beta_j \sim$ Normal $(0, \tau^2)$, $j = 1, \ldots, m$
- Marginal model:

$$Y_{n\times 1} \sim MVN(X_{n\times p}\alpha, \tau^2 G_{n\times m}G'_{m\times n} + \sigma^2 \mathrm{I})$$

Derivation of Variance Component Test

- Log likelihood

$$\ell = -\frac{(Y - X\alpha)'(\tau^2 GG' + \sigma^2 \mathrm{I})^{-1}(Y - X\alpha)}{2} - \frac{1}{2}\log|\tau^2 GG' + \sigma^2 \mathrm{I}| - \frac{n}{2}\log(2\pi)$$

- Let $V(\tau^2) = \tau^2 GG' + \sigma^2 \mathrm{I}$
- Score function

$$\frac{\partial \ell}{\partial \tau^2} = \frac{(Y - X\alpha)' V(\tau^2)^{-1} GG' V(\tau^2)^{-1}(Y - X\alpha)}{2} - \frac{tr(V(\tau^2)^{-1}(GG'))}{2}$$

Derivation of Variance Component Test

- Score test statistic

$$
\begin{aligned}
Q &= \left.\frac{\partial \ell}{\partial \tau^2}\right|_{\tau=0} \\
&= \frac{1}{2}(Y - X\alpha)' V^{-1} GG' V^{-1} (Y - X\alpha) \\
&\quad - \frac{1}{2} tr(V^{-1}(GG')) \\
&= \frac{1}{2}(Y - X\alpha)' M (Y - X\alpha) - \frac{1}{2} tr(V^{\frac{1}{2}} M V^{\frac{1}{2}})
\end{aligned}
$$

- $M = V^{-1} GG' V^{-1}$, $V = \sigma^2 \mathrm{I}$

Derivation of Variance Component Test

- Q is not asymptotically normal

$$\begin{aligned} Q &= \frac{1}{2}(Y - X\alpha)'M(Y - X\alpha) - \frac{1}{2}tr(V^{\frac{1}{2}}MV^{\frac{1}{2}}) \\ &= \frac{1}{2}\widetilde{Y}'(V^{\frac{1}{2}}MV^{\frac{1}{2}})\widetilde{Y} - \frac{1}{2}tr(V^{\frac{1}{2}}MV^{\frac{1}{2}}) \end{aligned}$$

where $\widetilde{Y} = V^{-\frac{1}{2}}(Y - X\alpha) \sim N(0, \mathrm{I})$

- Let $\{\lambda_j, u_j, j = 1, \ldots, m\}$ be the eigenvalues and eigenvectors of $V^{\frac{1}{2}}MV^{\frac{1}{2}}$. Then

$$Q = \sum_{j=1}^{m} \lambda_j((u_j'\widetilde{Y})^2 - 1) = \sum_{j=1}^{m} \lambda_j(Z_j^2 - 1)$$

  - Q is not asymptotically normal

- Zhang and Lin (2003) show that

$$\widetilde{Y}'(V^{\frac{1}{2}}MV^{\frac{1}{2}})\widetilde{Y} \sim \sum_{j=1}^{m} \lambda_j \chi^2_{1,j}$$

Variance Component Test

- The exact probability associated with a mixture of χ^2 distributions is difficult to calculate.

- Satterthwaite method to approximate the distribution by a scaled χ^2 distribution, $\kappa\chi^2_\nu$, where κ and ν are calculated by matching the first and second moments of the two distributions.

- To adjust for $\widehat{\alpha}$, replace V^{-1} by projection matrix $P = V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}$.

General Form of Variance Component Test

- Linear model

$$y_i = X_i\alpha + h(G_i) + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

- $h(\cdot)$ is a centered unknown smooth function $\in \mathcal{H}$ generated by a positive definite kernel function $K(\cdot, \cdot)$.
- $K(\cdot, \cdot)$ implicitly specifies a unique function space spanned by a set of orthogonal basis functions $\{\phi_j(G), j = 1, \dots, J\}$ and any $h(\cdot)$ can be represented by linear combination of these basis $h(G) = \sum_{j=1}^{J} \zeta_j \phi_j(G)$(**the primal representation**)

General Form

- Equivalently, $h(\cdot)$ can also be represented using $K(\cdot,\cdot)$ as $h(G_i) = \sum_{j=1}^{n} \omega_j K(G_i, G_j)$ (**the dual representation**)
- For a multi-dimensional G, it is more convenient to specify $h(G)$ using the dual representation, because explicit basis functions might be complicated to specify, and the number might be high

Estimation

- Penalized likelihood function (Kimeldorf and Wahba, 1970)

$$l = -\frac{1}{2}\sum_{i=1}^{n}\left\{y_i - X_i'\alpha - \sum_{j=1}^{n}\omega_j K(G_i, G_j)\right\}^2 - \frac{1}{2}\lambda\omega' K\omega$$

where λ is a tuning parameter which controls the tradeoff between goodness of fit and complexity of the model

$$\widehat{\alpha} = \left\{X'(\mathrm{I} + \lambda^{-1}K)^{-1}X\right\}^{-1} X'(\mathrm{I} + \lambda^{-1}K)^{-1}y$$

and

$$\widehat{\omega} = \lambda^{-1}(\mathrm{I} + \lambda^{-1}K)^{-1}(y - X'\widehat{\alpha})$$

$$\widehat{h} = K\widehat{\omega}$$

Connection with Linear Mixed Models

- The same estimators can be re-written as

  $$\begin{bmatrix} X'V^{-1}X & X'V^{-1} \\ V^{-1}X & V^{-1} + (\tau K)^{-1} \end{bmatrix} \begin{bmatrix} \alpha \\ h \end{bmatrix} = \begin{bmatrix} X'V^{-1}y \\ V^{-1}y \end{bmatrix}$$

  where $\tau = \lambda^{-1}\sigma^2$ and $V = \sigma^2 \mathrm{I}$
- Estimators $\widehat{\alpha}$ and $\widehat{h}$ are best linear unbiased predictors under the linear mixed model

  $$y = X'\alpha + h + \varepsilon$$

  where h is a $n \times 1$ vector of random effects with distribution $N(0, \tau K)$ and $\varepsilon \sim N(0, V)$

General Form of Variance Component Test

- Testing $H_0 : h = 0$ is equivalent to testing the variance component τ as $H_0 : \tau = 0$ versus $H_1 : \tau > 0$
- The REML under the linear mixed model is

$$
\begin{aligned}
l &= -\frac{1}{2}\log|V(\tau^2)| - \frac{1}{2}|X'V^{-1}(\tau^2)X| \\
&\quad -\frac{1}{2}(y - X'\alpha)'V(\tau^2)^{-1}(y - X'\alpha)
\end{aligned}
$$

- Score statistic for $H_0 : \tau^2 = 0$ is

$$Q = (Y - X\widehat{\alpha})'K(Y - X\widehat{\alpha}) - tr(KP),$$

which follows a mixture of χ_1^2 distribution.

Kernel

- Kernel function $K(\cdot,\cdot)$ measures similarity for pairs of subjects
  - Linear kernel: $K(G_i, G_k) = \sum_{j=1}^{m} G_{ij} G_{kj}$

- Something about $K(\cdot,\cdot)$
  - Ability to incorporate high-dimension and different types of features (e.g., SNPs, expression, environmental factors)
  - $K(\cdot,\cdot)$ is a symmetric semipositive definite matrix
    - Eigenvalues are interpreted as % of the variation explained by the corresponding eigenvectors, but a negative eigenvalue implying negative variance is not sensible.
    - No guarantee that the optimization algorithms that work for positive semidefinite kernels will work when there are negative eigenvalues
    - Mathematical foundation moves from real numbers to complex numbers

Some Kernels

- Some kernels
  - $K(G_i, G_k) = \sum_{j=1}^{m} G_{ij} G_{kj} =< G_i, G_k >$
  - $K(G_i, G_k) = \frac{1}{2m} \sum_{j=1}^{m} \text{IBS}(G_{ij}, G_{kj})$, where IBS is identity-by-state
  - $K(G_i, G_k) = (< G_i, G_k >)^p$: polynomial kernel, $p > 0$
    - Modeling higher-order interaction

$$(< G_i, G_k >)^2 = (\sum_{j=1}^{m} G_{ij} G_{kj})^2 = \sum_{j=1}^{m} \sum_{j'=1}^{m} (G_{ij} G_{ij'})(G_{kj} G_{kj'})$$

  - $K(G_i, G_k) = \exp(-\|G_i - G_j\|^2/\sigma^2)$: Gaussian kernel

- Schaid DJ. (2010) Genomic similarity and kernel methods I: advancements by building on mathematical and statistical foundations. *Hum Hered* 70:109–31.
- Schaid DJ. (2010) Genomic similarity and kernel methods II: methods for genomic information. *Hum Hered* 70:132–140.

Choice of Kernels

- An advantage of the kernel method is its expressive power to capture domain knowledge in a general manner.
- Generally difficult to construct a good kernel for a specific problem
- Basic operations to create new kernels from existing kernels:
  - multiplying by a positive scalar
  - adding kernels
  - multiplying kernels (element-wise).

Generalized Linear Model

- Observations for the linear model apply to the generalized linear model
- Penalized log-likelihood function

$$l = \sum_{i=1}^{n} \left[y_i(X_i'\alpha + \sum_{j=1}^{n} \omega_j K(G_i, G_j)) \right.$$

$$\left. - \log\{1 + \exp(X_i'\alpha + \sum_{j=1}^{n} \omega_j K(G_i, G_j))\} \right] - \frac{1}{2}\lambda\omega' K\omega$$

- The logistic kernel machine estimator

$$\begin{bmatrix} X'DX & X'D \\ DX & D + (\tau K)^{-1} \end{bmatrix} \begin{bmatrix} \alpha \\ h \end{bmatrix} = \begin{bmatrix} X'D\tilde{y} \\ D\tilde{y} \end{bmatrix}$$

where $\tau = \lambda^{-1}\sigma^2$, $D = diag\{E(y_i)(1 - E(y_i))\}$, and $\tilde{y} = X\alpha + K\omega + var(y)^{-1}(y - \mu)$

Generalized Linear Model

- The same estimators can be obtained from maximizing the penalized quasilikelihood from a logistic mixed model

  $$logit E(y_i) = X_i'\alpha + h_i$$

  where $h = (h_1, \ldots, h_n)$ is a $n \times 1$ vector of random effects following $h \sim N(0, \tau K)$ with $\tau = 1/\lambda$
- The score statistic for τ is

  $$Q = (y - X'\widehat{\alpha})'K(y - X'\widehat{\alpha}),$$

  which follows a mixture of χ^2 distributions

Exponential Family

- Suppose y_i follows a distribution in the exponential family with density

$$p(y_i; \theta_i, \phi) = \exp\{\frac{y_i\theta_i - a(\theta_i)}{\phi} + c(y_i, \phi)\},$$

  where $\theta_i = X_i'\alpha + h(G_i)$ is the canonical parameter, $a(\cdot)$ and $c(\cdot)$ are known functions, ϕ is a dispersion parameter

- $\mu_i = E(y_i) = a'(\theta_i)$ and $var(y_i) = \phi a''(\theta_i)$
- Gaussian: $\phi = \sigma^2$, $a(\theta_i) = \theta_i^2/2$, and $a'(\theta_i) = \theta_i$
- Logistic: $\phi = 1$, $a(\theta_i) = \log(1 + \exp(\theta_i))$, $a'(\theta_i) = \frac{\exp(\theta_i)}{1+\exp(\theta_i)}$
- Other distributions: log-normal, Poisson, etc.

Summary

- Burden tests are more powerful when a large number of variants are causal and all causal variants are harmful or protective.
- Variance component test is more powerful when a small number of variants are causal, or mixed effects exist.
- Both scenarios can happen across the genome and the underlying biology is unknown in advance.

Combined Test

- SKAT (SNP-set/Sequence Kernel Association Test): variance component test
- Combine the SKAT variance component and burden test statistics (Lee et al. 2012)

$$Q_\rho = (1-\rho)Q_{\text{SKAT}} + \rho Q_{\text{burden}}, \qquad 0 \le \rho \le 1$$

  - $\rho = 0$: SKAT
  - $\rho = 1$: Burden

- Instead of assuming $\{\beta_j\}$ are iid from $F(0, \tau^2)$, assume

$$\begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \end{pmatrix} \sim F\left(\underline{0}, \quad \tau^2 \begin{pmatrix} 1 & \rho \dots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \dots & 1 \end{pmatrix} \right)$$

SKAT-O

- $Q_\rho = (1-\rho)Q_{\text{SKAT}} + \rho Q_{\text{burden}}$, which is asymptotically equivalent to
$$(1-\rho)\kappa + a(\rho)\eta_0,$$
where κ follows a mixture of χ_1^2 and $\eta_0 \sim \chi_1^2$.
- Use the smallest P-value from different ρs:
$$T = \inf_{0\le\rho\le 1} P_\rho$$
- In practice, evaluate Q_ρ on a set of pre-selected grid points,
$$0 = \rho_1 < \cdots < \rho_B = 1$$
$$T = \min_{\rho\in\{\rho_1,\cdots,\rho_B\}} P_\rho$$

Summary

- Have robust power under a wide range of models
- Q_{SKAT} and Q_{burden} are not independent.
- The underlying model for SKAT-O is not natural.

Mixed Effects Model

- Model

$$g\{E(y_i)\} = X_i'\alpha + \sum_{j=1}^{m} G_{ij}\beta_j \tag{1}$$

  - Burden: $\beta_1 = \cdots = \beta_m$
  - SKAT: $\beta_j \sim F(0, \tau^2)$ independently
  - SKAT-O: $\beta_j \sim F(0, \tau^2)$ with pairwise correlation ρ

- Hierarchical model of β

$$\beta_j = w_j\eta + \delta_j \tag{2}$$

  - w_j: known features for the *j*th variant (e.g., $w_j = 1$ for all j's)
  - $\delta_j \sim F(0, \tau^2)$

Mixed Effects Model

- Plug (2) into (1)

$$g\{E(y_i)\} = X_i'\alpha + (\sum_{j=1}^{m} w_j G_{ij})\eta + \sum_{j=1}^{m} G_{ij}\delta_j$$

- Some examples:
  - If $w = 0$, $\delta_j = \beta_j$, the model becomes

$$g\{E(y_i)\} = X_i'\alpha + \sum_{j=1}^{m} G_{ij}\beta_j, \qquad \beta_j \sim F(0, \tau^2)$$

  - If $w = \mathbf{1}$ and $\delta_j = 0$, the model becomes

$$g\{E(y_i)\} = X_i'\alpha + (\sum_{j=1}^{m} G_{ij})\eta$$

Mixed Effects Model

- Some examples:
  - $w_j = (w_{j1}, w_{j2})$ where

$$
\begin{aligned}
w_{j1} &= 1 \text{ for } j = 1, \cdots, m \\
w_{j2} &= \begin{cases} 1 & \text{if } j\text{th variant is a missense} \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
$$

  - $\sum_{j=1}^{m} w_j G_{ij} = (\sum_{j=1}^{m} G_{ij}, \sum_{j=1}^{m} w_{j2} G_{ij})$
  - η_1 : average effect of m variants

    η_2 : effect of missense variants relative to the average
  - δ_j: residual variant specific effect $\sim F(0, \tau^2)$

Mixed Effects Model-based Test

- Mixed effects model

$$g\{E(y_i)\} = X_i'\alpha + (\sum_{j=1}^{m} w_j G_{ij})\eta + \sum_{j=1}^{m} G_{ij}\delta_j$$

- Null hypothesis is $H_0 : \eta = 0$ and $\tau^2 = 0$
  - η: fixed effects; τ^2: variance component
- The score test statistic for τ^2 and η is

$$S_\eta = (Y - X\widetilde{\alpha})'(GW)(GW)'(Y - X\widetilde{\alpha}),$$

and

$$S_{\tau^2} = \left(Y - X\widetilde{\alpha}\right)' GG' \left(Y - X\widetilde{\alpha}\right),$$

where $\widetilde{\alpha}$ is MLE of α under H_0.

- However, S_{τ^2} and S_η are not independent.

Independence of score test statistics

- We made a minor (but important) modification

$$S^*_{\tau^2} = \Big(Y - X\widehat{\alpha} - GW\widehat{\eta}\Big)' GG' \Big(Y - X\widehat{\alpha} - GW\widehat{\eta}\Big),$$

  where $(\widehat{\alpha}^T, \widehat{\pi}^T)$ are obtained under $\tau^2 = 0$.

- We can show that $S^*_{\tau^2}$ and S_η are independent.

$$\begin{aligned} & E\{(GW)'(Y - X\tilde{\alpha})((Y - X\widehat{\alpha} - GW\widehat{\eta}\Big)' G \\ = \;& \sigma^2 E\{(GW)'(I - P_1)(I - P_2)G\} \\ = \;& 0, \end{aligned}$$

  where P_1 is the projection onto X and P_2 is the projection onto $(X, (GW))$.

Combining independent statistics

MiST (Mixed effects Score Test)

- P-value combination
  - Fisher's combination: reject H_0 at significance level α if $-2\log(P_{\tau^2}) - 2\log(P_\eta) \geq \chi^2_{4,\alpha}$
  - Tippitt's combination: reject H_0 at significance level α if $\min(P_{\tau^2}, P_\eta) \leq 1 - (1-\alpha)^{1/2}$
- Other combinations, e.g., linear combination

$$S = \rho S_\eta + (1-\rho) S^*_{\tau^2}$$

- Jianping Sun, Yingye Zheng, and Li Hsu (2013). A Unified Mixed-Effects Model for Rare-Variant Association in Sequencing Studies. Genetic Epidemiology, 37: 334-44.

Power Comparison

- m=10 variants, n=1000 subjects, $\alpha = 0.01$

Burden	SKAT	SKAT-O	MiST_F	MiST_T
$\beta_j = c/\{p_j(1-p_j)\}^{1/2}, j = 1, \ldots, 10$				
0.866	0.435	0.818	0.780	0.811
$\beta_3 = 1.5c, \beta_4 = -1.5c, \beta_5 = c, \beta_6 = -c;$				
0.014	0.507	0.397	0.417	0.455
$\beta_1 = \beta_4 = \beta_7 = c$				
0.283	0.578	0.551	0.652	0.515
$\beta_1 = c, \beta_4 = 0.5c, \beta_7 = 0.25c$				
0.288	0.415	0.427	0.583	0.429

Dallas Heart Study

- Dallas Heart Study (Victor et al. 2004). n=3409 subjects, 3 genes (ANGPTL3, ANGPTL4 and ANGPTL5) were sequencied.
- We analyzed these genes in association with log(triglyceride).

	ANGPTL3	ANGPTL4	ANGPTL5
Burden	0.83	0.76	0.001
SKAT	0.40	0.31	0.38
SKAT-O	0.57	0.47	0.35
EREC	0.36	0.38	0.09
MiST_F	0.36	0.06	0.05
MiST_T	0.40	0.06	0.06
$\text{MiST}_\text{F}(Z)$	0.25	0.77	0.00005
$\text{MiST}_\text{T}(Z)$	0.27	0.32	0.0001

- The component p-values of *ANGPTL5*: $p_\pi = 5x10^{-6}$ and p_{τ^2} = 0.53. Furthermore, $p = 0.004$ for nonsense variants and p=0.24 for frame shift variants.

Summary

- MiST (Mixed effects Score Test) is based on hierarchical models for a set of variants
- The model includes the usual appealing features for regression models such as adjusting for confounders and being able to accommodate different types of outcomes by using appropriate link functions.
- It models the variant effects as a function of (known) variant characteristics to leverage information across loci while still allowing for individual variant effects.

Combining K studies

We have discussed for single variant analysis:

- Pooling the data from K studies. Since all score statistics are derived from regression models, it is easy to account for the differences between studies by adjusting for study and/or study $\times$ covariates
  - Pooling the data ensures consistency in data QC and model fitting
  - Pooling can be logistically difficult and time consuming
  - Sometimes protection of human subjects prohibit sharing the data

- Meta-analysis of combining summary statistics from K studies is still a viable alternative

Revisit Score Statistics

- Weighted burden test

$$U_{\text{burden}} = \sum_{i=1}^{n} \left(\sum_{j=1}^{m} w_j G_{ij} \right) (y_i - X_i' \widehat{\alpha})$$
$$= \sum_{j=1}^{m} w_j \underbrace{\sum_{i=1}^{n} G_{ij}(y_i - X_i' \widehat{\alpha})}_{U_j \text{ : Score of single variant model}}$$

- $U_j = \sum_{i=1}^{n} G_{ij}(Y_i - X_i \widehat{\alpha})$ is a score function of a single variant model.

$$y_i = X_i \alpha + G_{ij} \beta_j + \varepsilon_i, \qquad \varepsilon \sim N(0, \sigma^2)$$

Variance Component test

- Q_{SKAT} is a weighted sum of squared score statistics of the single SNP marginal model.

$$
\begin{aligned}
Q_{\text{SKAT}} &= (Y - X\widehat{\alpha})' GG'_{m\times n}(Y - X\widehat{\alpha})_{n\times 1} \\
&= \sum_{j=1}^{m}\{\sum_{i=1}^{n} G_{ij}(Y_i - X_i\widehat{\alpha})\}^2 \\
&= \sum_{j=1}^{m} U_j^2
\end{aligned}
$$

Key Elements

- A vector of single variant score statistics, $U' = (U_1, \ldots, U_m)$ with covariance $V = cov(U)$
- Burden score statistic

$$U_{\text{burden}} = W'U \quad , \quad \text{var}(U_{\text{burden}}) = W'VW$$

- Variance component score statistic

$$Q_{\text{SKAT}} = U'U,$$

which follows a mixture of chi^2 distribution with weights being the eigenvalues of V

Fixed effects model

- For $k = 1, \cdots, K$, let U_k and V_k denote the score statistics and covariance for the kth study.
- Score statistic over K studies is

$$U = \sum_{k=1}^{k} U_k \quad V = \sum_{k=1}^{k} V_k$$

- Burden test

$$U_{\text{burden}} = W'U \qquad \text{var}(U_{\text{burden}}) = W'VW$$

$$U'_{\text{burden}} \text{var}(U_{\text{burden}})^{-1} U_{\text{burden}} \sim \chi^2_p$$

Fixed effects model

- Variance component test

  $$Q_{\text{SKAT}} = U'U \sim \sum_{j=1}^{m} \lambda_j \chi^2_{1,j}$$

  where λ_j is the jth eigenvalue of $V = \sum_{k=1}^{K} V_k$

- Combination of burden and score statistics

  $$Q_\rho = (1-\rho) Q_{\text{SKAT}} + \rho Q_{\text{burden}}$$

  where ρ is adaptively chosen and the p-value can be obtained by one-dimensional numerical integration

Fixed effects model

- Summary of single variant score statistic may not enough for MiST score statistics

$$S_\eta = (Y - X\widetilde{\alpha})'(GW)(GW)'(Y - X\widetilde{\alpha}),$$

$$S^*_{\tau^2} = \left(Y - X\widehat{\alpha} - GW\widehat{\eta}\right)' GG' \left(Y - X\widehat{\alpha} - GW\widehat{\eta}\right),$$

where $(\widehat{\alpha}^T, \widehat{\eta}^T)$ are obtained under $\tau^2 = 0$.

Random Effects Model

- For $k = 1, \cdots, K, \quad \beta'_k = (\beta_{k1}, \cdots, \beta_{km})$ is the effect of m variants for the kth study.
- Random effects model

$$\beta_k = \beta_0 + \xi_k$$

where $\beta_0 = (\beta_{01}, \cdots, \beta_{0m})$ represents the average effect among the studies, ξ_k is a set of random effects representing the deviation of the kth study from the average effect $\xi_k \sim N(0, \Sigma)$

Heterogeneity

- Assume $\Sigma = \sigma^2 B$, where B is a pre-specified matrix to constrain the potential many parameters in Σ.
- A choice of B is

$$B = \begin{pmatrix} b_1^2 & b_1 b_2 r & \cdots & b_1 b_m r \\ b_2 b_1 r & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ b_m b_1 r & \ldots\ldots & \ldots\ldots & b_m^2 \end{pmatrix}$$

- $(b_1, \cdots b_m)$ controls the relative degrees of heterogeneity for the m variates (e.g., MAF), and r specifies the correlation of heterogeneity.
- Choice of B has no effect on the type I error but may affect the power.

New Random Effects Burden Test

- The null hypothesis $H_0 : \beta_0 = 0, \sigma^2 = 0$
- For $k = 1, \ldots, K$, $\widehat{\beta}_k \sim N(\beta_0, \Omega_k = V_k^{-1} + \sigma^2 B)$. The log-likelihood function is

$$l = -\frac{1}{2}\sum_{k=1}^{K}(\widehat{\beta}_k - \beta_0)'\Omega_k^{-1}(\widehat{\beta}_k - \beta_0) - \frac{1}{2}\sum_{k=1}^{K}\log|\Omega_k|$$

- Let $\widehat{\beta}_k \approx V_k^{-1}U_k$, the random effects (RE) test for fixed effects

$$U_{\text{burden}}^{\text{RE}} = U'V^{-1}U + \frac{U_\sigma^2}{V_\sigma}$$

where $U_\sigma = \frac{1}{2}\sum_{k=1}^{K} U_k' B U_k - \frac{1}{2}\text{tr}(VB)$, $V_\sigma = \frac{1}{2}\text{tr}(\sum_{k=1}^{K} V_k B V_k B)$

- For burden test, replace U by $W'U$, and V by $W'VW$.

New Random Effects Variance Component Test

- $\beta_0 \sim N(0, \tau^2 W)$, where W is a pre-specified matrix, e.g.,

$$W = \begin{pmatrix} w_1^2 & w_1 w_2 \rho & \cdots \\ w_2 w_1 \rho & \ddots & \\ \vdots & & w_d^2 \end{pmatrix}$$

  where $(w_1, \cdots, w_d)$ controls the relative magnitude of the d average genetic effects, and ρ indicates the correlation.

- The null hypothesis $H_0 : \tau^2 = 0, \sigma^2 = 0$
- Let $\widehat{\beta}' = (\widehat{\beta}_1, \ldots, \widehat{\beta}_K)$, then

$$\widehat{\beta} \sim MVN\left(0, \tau^2(J_K \otimes W) + \sigma^2(I_K \otimes B) + \mathrm{diag}(V_1^{-1}, \cdots, V_K^{-1})\right)$$

  where $\otimes$ denotes Kronecker product

- $\{\widehat{\beta}_k \approx V_k^{-1} U_k\}$, the score statistic is a function of U_k, V_k, $k = 1, \ldots, K$.

Summary

- Pooled- vs meta-analysis
- For meta-analysis rare variant association tests can be constructed from multivariate summary statistics, i.e., the score vector U and information matrix V
- Fixed vs random effects model

Set-based gene-environment interaction

- m variants, $G_i = (G_{i1}, \cdots, G_{im})'$
- E_i: environmental covariate
- X_i: covariates
- Gene-environment interaction (GxE) model

$$g\{E(y_i)\} = X_i'\alpha + E_i\beta^E + \sum_{j=1}^{m} G_{ij}\beta_j^G + \sum_{j=1}^{m}(E_i G_{ij})\beta_j^{GE}$$

- No interaction means $\beta = (\beta_1^{GE}, \cdots, \beta_m^{GE}) = 0$

Hierarchical model for β^{GE}

- Model the interaction effect

$$\beta_j^{GE} = w_j \eta + \delta_j$$

  - w_j: a vector of known features
  - $\delta_j \sim F(0, \tau^2)$

- The interaction effect term

$$\begin{aligned}\sum_{j=1}^{m} (E_i G_{ij}) \beta_j^{GE} &= \left(\sum_{j=1}^{m} E_i G_{ij} w_j \right) \eta + \sum_{j=1}^{m} E_i G_{ij} \delta_j \\ &= E_i (\sum_{i=1}^{m} G_{ij} w_j) \eta + \sum_{j=1}^{m} (E_i G_{ij}) \delta_j\end{aligned}$$

- No interaction means $H_0 : \eta = 0, \tau^2 = 0$

Challenges

- Main effects $\{\beta_1^G, \cdots, \beta_m^G\}$ may not be estimated reliably if m is large or variants are rare.
- Assume the main effects $\{\beta_j^G\}$ are random effects such that
$$\beta_j^G \sim F(0, \nu^2)$$
- Need to derive score statistics for the mixed GxE effects (η, τ^2) in the presence of another random effects β_j^G.

Estimation

- β^G can be estimated by maximum posterior approach (or best linear unbiased prediction, in the linear mixed effects model), but the computation is intensive under a generalized linear model due to *m*-dimensional integration with no closed form.
- $\widehat{\beta}_j^G$ minimizes ridge regression

$$\widehat{\beta}^{ridge} = \text{argmin} \left\{ \sum_{i=1}^{n} (y_i - X_i' \alpha - E_i \beta^E - G_i \beta^G)^2 + \lambda \sum_{j=1}^{m} \beta_j^2 \right\}$$

where $\lambda = \sigma^2 / \nu^2$

Some nice properties about ridge

- Knight and Fu (2000) states that if $\lambda = o(\sqrt{n})$ then $\widehat{\beta}^{\lambda}$ is a $\sqrt{n}$ consistent estimator of β_0

- Score statistics for the fixed effects under $H_0 : \eta = 0, \tau^2 = 0$

  $$u_{\eta} = (D - \tilde{\mu})'\big(E(\sum_{j=1}^{m} G_j w_j)\big)' V\big(E(\sum_{j=1}^{m} G_j \cdot w_j)\big)(D - \tilde{\mu})$$

  where $\widetilde{\mu} = \widehat{E}(D|G, E)$ under $\eta = 0, \tau^2 = 0$

- Score statistic for the variance component under $H_0 : \tau^2 = 0$

  $$u_{\tau^2} = (D - \widehat{u})(GE)(GE)'(D - \widehat{\mu})$$

  where $\widehat{u} = \widehat{E}(D|G, E)$ under $\tau^2 = 0$

Combination of score statistics

- P-value based, $Z_\eta = -2\log P_\eta$ and $Z_{\tau^2} = -2\log P_{\tau^2}$

$$T_f = Z_\eta + Z_{\tau^2} \sim \chi^2_4$$

- Grid-search based optimal linear combination

$$T_o = \max_{\rho\in[0,1]} (\rho U_\eta + (1-\rho)U_{\tau^2})$$

  where ρ is restricted on a set of pre-specified grid points $\{0 = \rho_0, \rho_1, \ldots, \rho_d = 1\}$

- Adaptive-weighted linear combination

$$T_a = Z^2_\eta + Z^2_{\tau^2}$$

  - Give more weight to either burden or variance component if the evidence comes mainly from one

- Su YR, Di C and Hsu L (2015). A unified powerful set-based test for sequencing data analysis of GxE interactions. Submitted.

Power comparison

- $m = 25$ variants

T_o	T_a	T_f	Burden	Var Comp
H_a : 30% variants $\beta = c$				
0.541	0.620	0.672	0.473	0.533
H_a : Half $\beta = c$, other half $\beta = -c$				
0.544	0.542	0.516	0.021	0.632
H_a : All $\beta = c$				
0.768	0.770	0.740	0.848	0.050

Weight

- Choices of weight
  - Functioncal characteristics (e.g., missense, nonsense)
  - Screening statistics, M_j and C_j are the Z statistics from marginal association screening and correlation of G and E screening

    $$w_j = \begin{cases} M_j & \text{if } |M_j| > |C_j| \\ C_j & \text{otherwise} \end{cases}$$

    Since the screening statistics are independent of GxE test, no need to use permutation to calculate the p-values
  - Jiao S, Hsu L, et al. (2013, 2015)

Summary

- Set-based association testing
  - Mixed effects model that accounts for both burden genetic risk score and variance component
- Meta-analysis
- GxE interaction between a set of variants and environmental factor

Recommended readings

- Liu D et al. (2007). Semiparametric regression of multidimensional genetic pathway data: Least-Squares Kernel Machines and Linear Mixed Models. Biometrics 63:1079–1988.
- Liu D et al. (2008). Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. BMC Bioinformatics 9:292.
- Schaid DJ (2010) Genomic similarity and kernel methods I: advancements by building on mathematical and statistical foundations. *Hum Hered* 70:109–31.
- Schaid DJ (2010) Genomic similarity and kernel methods II: methods for genomic information. *Hum Hered* 70:132–140.
- Zhang and Lin (2003). Hypothesis testing in semiparametric additive mixed models. *Biostatistics* 4:57–74.

Recommended readings

- Lee S, Lin X and Wu M (2012). Optimal tests for rare variant effects in sequencing association studies. Biostatistics 13: 762–775.
- Lin DY and Tang Z (2011). A general framework for detecting disease associations with rare variants in sequencing studies. AJHG 89: 354–67.
- Jianping Sun, Yingye Zheng, and Li Hsu (2013). A Unified Mixed-Effects Model for Rare-Variant Association in Sequencing Studies. Genetic Epidemiology, 37: 334-44.
- Jiao S, ..., Hsu L (2015) Powerful Set-Based Gene-Environment Interaction Testing Framework for Complex Diseases. Genetic Epidemiology, DOI: 10.1002/gepi.21908
- Tang Z and Lin DY (2015). Meta-analysis for discovering rare-variant associations: Statistical methods and software programs. AJHG 97:35–53
- Wu et al. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. AJHG 89: 82-93.