Overall goal: to understand the necessity of identifying rare variants and the proposed statistical methods for rare-variant association testing.

- Rationale for studying rare variants (complex trait)
- Sequencing and study designs
- Rare-variant association tests
- Summary
Common vs Rare variant

- MAF: frequency at which the least common allele occurs in population
- Common variants: MAF $\geq 5\%$
- Low frequency variants: $0.5\% \leq$ MAF $< 5\%$
- Rare variants: MAF $\leq 0.5\%$
Genome-wide association studies (GWASs)

- GWASs focused on identification of common variants
- Only limited heritability explained
- Recall that heritability refers to the proportion of total trait variation that is attributable to (considered) genetic component.
Most of human variants are rare
Why rare variants?

- Functional variants tend to be rare
Further, since common variants explained limited variation in the trait . . .

Some argued rare variants could explain additional trait variability

Advancement of sequencing technology (NGS), reduction in cost
Challenges

- Require cost-effective study designs to genotype many individuals

It can be shown that at least $\log(1 - \theta)/[2 \log(1 - MAF)]$ individuals are needed to observe a variant with no less than $\theta$ chance. For $\theta = 99.9\%$, we have

<table>
<thead>
<tr>
<th>MAF(%)</th>
<th>Minimum sample size</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>33</td>
</tr>
<tr>
<td>1</td>
<td>344</td>
</tr>
<tr>
<td>0.1</td>
<td>3453</td>
</tr>
<tr>
<td>0.01</td>
<td>34537</td>
</tr>
</tbody>
</table>

- Classical single-variant tests, developed for common variants detection, are underpowered - focus of the paper

- Multiple testing
Design alternatives to deep sequencing

- Low-depth whole-genome sequencing: sequencing depth refers to the average number of reads that cover each base; limited accuracy
- Exome sequencing: limited to exome
- High-priority region sequencing: limited to the target region
- ...

In summary, either the sequenced range or accuracy is compromised
How to test for rare-variant association? The single-variant test?

Assume an additive genetic model (additive effects depending on the number of allele copies), and fit linear/logistic regression

Test for association with the genome-wide significance threshold $5 \times 10^{-8}$

Much less powerful! More stringent threshold (larger number for rare variants)!
Aggregation tests for multiple variants

- Region-based: gene, regulatory region
- Identify multiple genetic variants within a region
- Evaluate the joint effects of these variants while adjusting for covariates
- **Caution:** These tests rely on assumptions for genetic model (e.g.: mode of inheritance), and the power depends on the true disease model $h(\mu(Y))$
Regression model

- $n$ subjects ($i = 1, \ldots, n$)
- $m$ variants in a region
- Allele counts in a region $\mathbf{G}_i = (G_{i1}, \ldots, G_{im})'$, ($G_{ij} = 0, 1, 2$)
- $q$-dimensional covariates $\mathbf{X}_i$ (age, gender, PC scores etc.)
- The disease model is given by a GLM

$$h(\mu(Y_i)) = \alpha_0 + \alpha' \mathbf{X}_i + \beta' \mathbf{G}_i$$  \hspace{1cm} (1)

- Now the interest is in the null of no genetic-region effect:

$$H_0 : \beta = (\beta_1, \ldots, \beta_m)' = \mathbf{0}_{m \times 1}$$
The score statistic

- All the tests are in some sense a modification of the score test for the previous $H_0$.
- Under $H_0$, the score statistic for a single variant $j$ (marginally)

$$S_j = \sum_{i=1}^{n} G_{ij} (Y_i - \hat{\mu}_i),$$

where $\hat{\mu}_i$ is estimated under the null model with $\beta$ set to zero vector.
Burden tests

- Collapse information on multiple genetic variants into a single genetic score
- Essentially an association test between the score and trait
- Define a weight for each variant $w = (w_1, \ldots, w_m)'$, the score is developed as

$$h(\mu(Y_i)) = \alpha_0 + \alpha'X_i + \beta'G_i$$

$$= \alpha_0 + \alpha'X_i + \tilde{\beta} \underbrace{w'G_i}_{\text{scalar } C_i} \quad (2)$$

- Under $H_0 : \tilde{\beta} = 0$, the score statistic is
$$Q_{\text{burden}} = (\sum_{j=1}^m w_j S_j)^2$$
If $w = 1$, we can collapse rare variants the following way

<table>
<thead>
<tr>
<th>Y</th>
<th>$G_1$</th>
<th>$G_2$</th>
<th>$G_3$</th>
<th>$G_4$</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Choice of \( w \) accommodates different assumptions about disease mechanism

e.g., the cohort allelic sums test (CAST)

\[
C_i = \begin{cases} 
1 & \text{when } 1'G_i > 0, \\
0 & \text{otherwise.}
\end{cases}
\]

Limitation: strong assumption about the same direction/magnitude of effect (post to weight adjustment); loss of power
Adaptive burden tests

- To obtain tests that are robust to null variants and allow for different effect directions
- Let the data speak!
- e.g. the data-adaptive sum test (aSum)
  - Estimate direction of each variant in marginal models
  - Use the burden test framework with $w_j = 1$ if the coef is likely to be positive and $w_j = -1$ otherwise
  - Require permutation (How?) to obtain the null distribution
  - Further modification based on model-selection allowing for zero weight

Limitations: although more robust, marginal models are unstable; permutation requires extensive calculation
Variance-component tests

- Is there another way to pool/group the rare variants in a region?
- Yes, resort to random-effects models
- To evaluate the distribution of genetic effects for a group of variants
- Suppose $\beta_j \sim N(0, w_j^2 \tau)$, $\text{corr}(\beta_j, \beta_k) = \rho$
  - e.g., the widely-used sequence kernel association test (SKAT, $\rho = 0$) tests $H_0 : \tau = 0$
  - $Q_{SKAT} = \sum_{j=1}^{m} w_j^2 S_j^2$, a weighted sum of squares of single variant scores, approx follows a mixture of Chi-squared dist
  - Robust to different directions of effects, but ...
  - Can lead to inflated test size in small effective sample size
Omnibus tests

- To achieve robust power
- Often referred to as “combined tests”
- How to combine different tests?
- Fisher’s combination method

\[
Fisher = -2 \log(p_{SKAT}) - 2 \log(p_{burden})
\]

- Combining test statistic

\[
Q_\rho = (1 - \rho) Q_{SKAT} + \rho Q_{burden}, \quad \rho \in [0, 1]
\]
Limitation: it might have lower power than SKAT or burden tests if the assumption underlying one of these tests are largely true

For unknown genetic architecture, this is an attractive choice
General comments on aggregation tests

- The tests are designed to boost power assuming the rare variants can be grouped together.
- This point is shown by simulation work by Li and Neal, 2008.
- Power loss occurs (relative to single-variant tests) when only a very few of the variants are associated with the trait and when many variants have no effects.
- e.g., Liu et al. studied the association between blood lipids and BCAM and CD300LG, but found weaker signal using gene-level test than single-variant test.
Meta-analysis

- Combine data from multiple studies
  - Rare variants association detection requires large sample
- Popular frameworks combine score statistic from different studies instead of combing p-values
  - only requires summary statistics
  - allows study-specific covariates
- Methods should account for heterogeneity of genetic effects (how? see Lee et al 2013 AJHG) across studies to increase power (diff in ancestries)
Summary - Which variants to use?

- Can use all the variants
- Obtain a refined subset on the basis of MAF, impact of amino acid sequence
- A subset based on the predicted functional role of variants (with bioinformatics tool)
First acknowledge that relative performance depends on the unknown disease architecture

Use available prior information
- the region has a large fraction of causal rare variants, majority increase disease risk – burden tests
- exist both risk-increasing and risk-decreasing variants – variance-component tests

If no prior information, one can try multiple methods or use the omnibus test