Lecture 4: Gene and Pathway Level Analysis of Genetic Association Studies

Timothy Thornton and Michael Wu

Summer Institute in Statistical Genetics 2015

Lecture Overview

Standard Analysis Strategy

Individual Variant Analysis:

1. For each SNP, compute a statistic measuring association
2. Compute a *p*-value for significance
3. Adjust for multiple comparisons:
   - FWER
   - FDR
4. Follow-up
   - Directly report results
   - Meta-analyze
5. Auxiliary analyses

Focus of traditional analyses is on a handful of SNPs that meet criteria for significance.

Limitations of the traditional approach:

Biggest problem: What if we don't find anything???

1. Genome Wide Significance: Stringent and difficult to reach. After correcting for multiple hypotheses testing, no SNPs are statistically significant.

2. An untyped causal SNP is in LD with multiple typed SNPs: Typed SNPs may only show moderate effects.

3. Most common diseases are complex: multi-SNP effects
   ▶ Most individual SNPs have only modest effects
   ▶ Joint effect of several, individually moderate, SNPs is important.

4. Reproducibility: Without strict thresholds: a large number of false positives!

5. Who Cares?: What's the biological or mechanistic interpretation of what you've found?

Alternative: Multi-SNP Analysis

Operationally Equivalent Terms: multi-SNP testing, multi-locus testing, gene based analysis, pathway analysis

Multi-SNP Analysis

▶ Idea: Group SNPs to form SNP sets and test them as a unit
▶ Forming SNP sets:
  1. Genes
  2. Pathways (many SNPs)
  3. Evolutionarily conserved regions
  4. Moving window
  5. Any group of SNPs selected w/o using outcome data

Advantages to Gene and Pathway Level Analysis

- ▶ Reduced multiple testing burden
    - ▶ Millions of SNPs $\rightarrow$ 20,000 genes
    - ▶ A few candidate pathways and candidate genes
- ▶ Capture multi-SNP effects:
    - ▶ Aggregate modest signals
    - ▶ Capture effects of untyped SNPs
    - ▶ Possibly capture complex (e.g. interactive) effects
- ▶ Biologically meaningful unit

Example: ASAH1 Gene

LD plot (correlation structure)
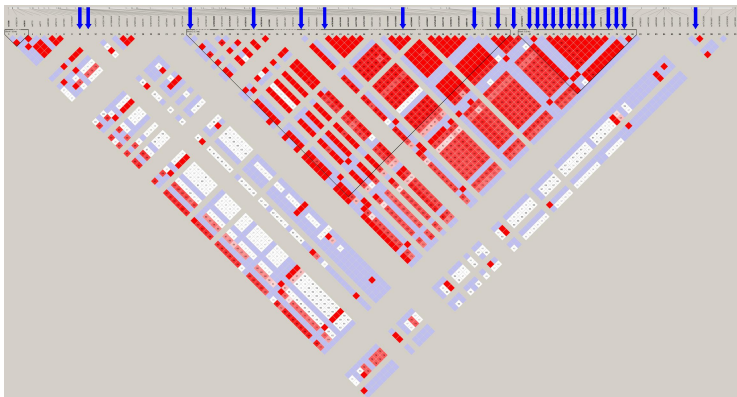
Pathways and Gene Sets

Beyond gene level (or a single region) analysis:

- ▶ Most biological phenomena occur through the concerted expression of multiple genes (signaling pathways or functional relationships)

- ▶ Use our prior knowledge of what SNPs belong to various genes which in turn belong to pathways or functional groups

- ▶ Numerous databases organizing genes into groups exist:
    1. Ingenuity
    2. MetaCore
    3. KEGG
    4. Gene Ontology (GO) Consortium
    5. etc...

- ▶ Note: Gene sets and functional groupings are NOT the same as Pathways.

Gene Ontology Consortium Taxonomy

- ▶ Three principal ontologies: Biological Processes, Cellular Components, and Molecular Function
- ▶ Each ontology is a directed acyclic graph
- ▶ The graph has a hierarchy of terms (GO terms) from very broad (metabolism) down to more narrow levels (GTP biosynthesis)
- ▶ Each ontology and GO term has a comprehensive list of genes previously demonstrated to be associated with that ontology or GO term.
- ▶ Contains a lot of JUNK! Filtering is necessary.
- ▶ A wide variety of packages in R can provide many basic tools for mining gene ontology information

Gene Ontology
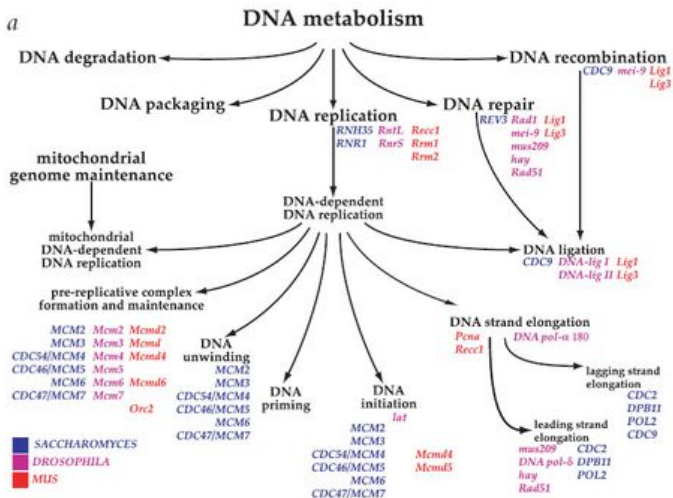
KEGG

- ▶ "Collection of online databases dealing with genomes, enzymatic pathways, and biological chemicals" - Wiki
- ▶ KEGG Pathways is network of gene pathways
- ▶ Cleaner set of pathways than GO, but much smaller: emphasis on metabolic pathways though there are also disease and other trait related pathways.
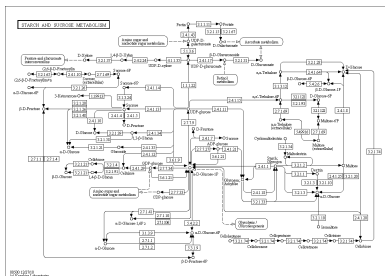
Question...

Suppose we know that a bunch of SNPs are inside of (genes in) a pathway. How do we test if the pathway is associated with the phenotype?

Statistical Methods:

Gene Level Analysis

- ▶ Minimum p-value Tests (minP)
- ▶ Combined p-value approaches
- ▶ Averaging/Collapsing Tests
- ▶ Variance Component (VC) Tests

Pathway Level Analysis

- ▶ Over-representation Analysis (ORA)
- ▶ Gene Set Enrichment Analysis (GSEA)
- ▶ minP, Collapsing, Combined p-value, VC Tests
- ▶ Graphical methods $\leftarrow$ not covered (usually like ORA)

Many tools can (technically) be used interchangeably

Minimum p-value

- ▶ Idea: let the smallest individual SNP p-value be the p-value for the entire pathway.
- ▶ Easy to run individual SNP analysis.
- ▶ How do we correct for having taken the smallest p-value?
  - ▶ Bonferroni correction.
  - ▶ Compute the effective number of tests.
  - ▶ Permutation.

Combined p-value Approaches

- ▶ Idea: combine the p-values across the SNPs in the gene
- ▶ Operationally:
    1. Test each individual SNP for association
    2. Combine the p-value for top SNPs, e.g. via Fisher's method
- ▶ Challenge: Most p-value combination approaches require independent p-value (i.e., no LD)
    - ▶ Permutation
- ▶ Variations include taking only top few p-values

Averaging/Collapsing

- ▶ Idea: can we collapse the SNP values down to a single value?
- ▶ We can construct a weighted average:

$$C_i = \sum_{j=1}^{p} w_j x_{ij}$$

  such that C_i is a "super-SNP". Then we can test for association between C and y.

- ▶ Common approaches to get the w_j
  - ▶ Simple average
  - ▶ Inverse of MAF
  - ▶ p-values from previous studies
  - ▶ PCA (1st or many)
  - ▶ Supervised approaches.
- ▶ Test effect of gene by regressing outcome on C_i

Similarity Based/Variance Component Methods: "Global Test"

- ► Build a regression model to predict the phenotype based on the SNPs:

$$g(\mathcal{E}(y_i)) = \alpha' \mathbf{Z}_i + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}$$

  Where x_{ij} is the genotype value for the jth SNP of the ith sample, $\mathbf{Z}_i$ are covariates, and g is some link function (e.g. logit).

- ► Testing for the joint effect of the SNPs is equivalent to:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_{N_s} = 0$$

- ► Assuming β's are iid with mean 0 and variance τ^2, then our null hypothesis is simply $H_0 : \tau^2 = 0$
- ► Can either use permutation or asymptotics to get the p-values.

Similarity Based/Variance Component Methods: Kernel Machine Methods

- ▶ Generalize the variance component testing to nonparametric regression setting:

$$g(\mathcal{E}(y_i)) = \alpha' \mathbf{Z}_i + h(\mathbf{X}_i)$$

  where the effect of the SNPs are modeled non-parametrically.

- ▶ Allows for "complex" effects of SNPs on outcome: interactions, nonlinearity, etc.

- ▶ More on this when we talk about rare variants.

Over-representation Analysis (ORA)

▶ Start from the list of "significant" SNPs
  ▶ Can be based on multiple comparisons criterion as mentioned earlier
  ▶ 100 SNPs with smallest p-value
  ▶ Top 5% of SNPs with smallest p-value
  ▶ Many other ways...

▶ Look for an over-representation of the SNPs in the pathway among "most significant" SNPs (or over-representation of "most significant" SNPs in the pathway)

ORA - 2x2 Contingency Tables

With the list of "significant" SNPs (D) and the list of SNPs in the pathway (S), we can build a 2x2 table:

	Significant	Not Significant	
In pathway	N_{SD}	N_{SD^c}	N_S
Not in pathway	$N_{S^c D}$	$N_{S^c D^c}$	N_{S^c}
total	N_D	N_{D^c}	N

Generate a p-value for representation by using a test for independence:

- Fisher's Exact Test
- χ^2-test
- Hypergeometric Test
- Binomial proportions z-test
- Choice of test is unimportant in practice.

ORA - Criticism

- ▶ All of the tests on the previous slide require independence among SNPs.
- ▶ Length Bias.
- ▶ Alternative approach:
  - ▶ Conduct a gene level analysis (multiple regression) to get a p-value for all SNPs in the gene
  - ▶ Apply ORA at the gene (instead of SNP) level.
- ▶ LD and length bias are NOT the biggest problem: more on this later.

Gene Set Enrichment Analysis (GSEA)

Original GSEA Approach:

1. Rank all N SNPs (or genes) based on their p-values to obtain L, the SNP/gene list

2. Calculate an Enrichment Score (ES) for the data set:
   For G_i (the i-th gene in L), let:

   $$X_i = \begin{cases} \sqrt{\frac{N_{S^c}}{N_S}} & \text{if } G_i \text{ is in } S \\ -\sqrt{\frac{N_S}{N_{S^c}}} & \text{if } G_i \text{ is NOT in } S \end{cases}$$

   $\text{ES}(S) = max_{1 \le j \le N} |\sum_{i=1}^{j} X_i|$

3. Evaluate Significance:
   3.1 Randomly permute the class labels
   3.2 Re-rank the SNPs
   3.3 Calculate $\text{ES}(S)$ based on the new ranked gene list
   3.4 Repeat the above for a bunch of times

Gene Set Enrichment Analysis (GSEA)

- ▶ Start from the full list of genes
- ▶ Order the genes according to association p-value to obtain L
- ▶ Look to see of genes in the gene set S are randomly distributed throughout L or primarily at the top or bottom.
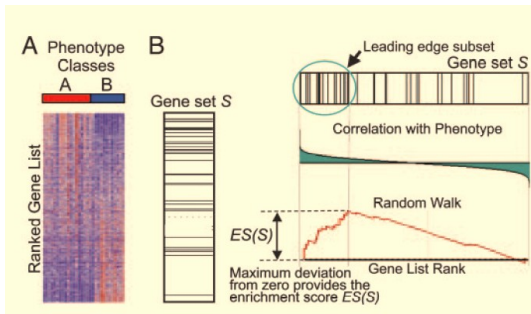
Statistical Considerations

Goals...

Goal: Test the null hypothesis that my pathway is not associated
with the outcome...

What does this even mean???

What's my Null?

Two different possible null hypotheses:

Competitive Null Hypothesis:

H_0^{comp} : The SNPs in S are at most as often associated with the outcome as the SNPs in S^c

- ▶ Over-representation analysis (2x2 contingency table methods)
- ▶ GSEA

Self-contained Null Hypothesis:

H_0^{self} : No SNPs in S are associated with the outcome

- ▶ Variance Component Tests
- ▶ Minimum P-value
- ▶ Collapsing

Competitive Null Hypotheses

- ▶ Pits one pathway against another
- ▶ Competitive tests cannot compare all of the SNPs on the chip.
- ▶ In the competitive testing framework, significant SNPs in one pathway will generally lead to larger p-values for other pathway. Thus, p-values tend to be negatively correlated which is problematic if we want to control for the FDR.

Self Contained Null Hypotheses

- ▶ Self-contained tests theoretically have more power since truth of H_0^{self} generally implies H_0^{comp}. Under the competitive setup significance is penalized in experiments with many disease associated SNPs.

- ▶ Self-contained tests are direct generalizations of individual SNP tests (they are equivalent for pathways with only a single SNP).

- ▶ Testing the global null sometimes violates the spirit of pathway analysis.

- ▶ Note: outside of SNPs, self-contained tests may be *too* powerful in data sets where many features appear to be important

What's my sampling unit?

Subject Sampling:

- ▶ GSEA
- ▶ Variance Component Tests
- ▶ Averaging/Collapsing
- ▶ MinP and Combined p-value tests

SNP Sampling:

- ▶ Over-representation analysis (2x2 contingency table methods)

SNP vs. Subject Sampling

- *Classical tests are based on experiments that sample subjects:* draw a sample of subjects, each with the same fixed set of SNPs (sample size is number of subjects)

- *SNP sampling flips the classical setup:* draw a new sample of SNPs coming from a fixed set of subjects (sample size is number of SNPs)

- Interpretation of *p*-value's depends on the sampling scheme:
  - *Subject Sampling:* significant *p*-value gives confidence that the associations found between SNPs and the outcome will be found for a new sample of subjects
  - *SNP Sampling:* significant *p*-value gives confidence that a for a new set of SNPs from the same subjects, there will be a similar association between being in the gene/pathway and being called "significant"

SNP vs. Subject Sampling (continued)

- ▶ *SNP sampling fails to mimic the biological experiment performed* which always take a new sample of subjects rather than a new sample of genes.
- ▶ Both sampling schemes assume sampling units are *independent and identically distributed*. That SNPs are independent is extremely unrealistic. – this is minor relative to the interpretation of the *p*-value.
- ▶ Broadly speaking, SNP sampling is wrong!
- ▶ How to look out for SNP sampling:
  - ▶ Words: "enrichment", "over-representation", "fisher's exact test", "hypergeometric test"
  - ▶ Software: DAVID, Ingenuity (IPA), GeneGo, most commercial stuff.
  - ▶ Tiny, tiny p-values
  - ▶ Any method that only uses individual *p*-values.
  - ▶ Fancy pictures.

Remarks

- ▶ Different methods give different results
- ▶ Different methods operate under different assumptions
- ▶ SNP sampling is generally not reasonable for most practical settings: "invalid"
  - ▶ Invalid statistics does not mean biology is wrong
  - ▶ Can still be useful for "interpretation" (though then the p-value calculation is a waste of time)
- ▶ Self contained testing is in some ways more natural, but can be difficult to interpret as a pathway result.

Skepticism Regarding Pathway Analysis

A quote from a well known statistician regarding pathway analysis:

... at best the authors believe it to be true.

Some Issues:

▶ Inappropriate or invalid methods used

▶ Applied when no marginal significance (i.e. run when there really isn't much going on in the data)

▶ Cherry-picking results: inappropriate control for multiple testing

References

▶ Wang, Kai, Mingyao Li, and Maja Bucan. "Pathway-based approaches for analysis of genomewide association studies." *The American Journal of Human Genetics* (2007): 1278-1283.

▶ Yu, Kai, et al. "Pathway analysis by adaptive combination of Pvalues." *Genetic epidemiology* (2009): 700-709.

▶ Wu, Michael C., et al. "Powerful SNP-set analysis for case-control genome-wide association studies." The American Journal of Human Genetics 86.6 (2010): 929-942.

▶ Wu, Michael C., and Xihong Lin. "Prior biological knowledge-based approaches for the analysis of genome-wide expression profiles using gene sets and pathways." *Statistical methods in medical research* (2009): 577-593.

▶ Goeman, Jelle J., and Peter Bhlmann. "Analyzing gene expression data in terms of gene sets: methodological issues." *Bioinformatics* (2007): 980-987.