

# Gene-gene Interaction and Other Topics

Jurg Ott

Rockefeller University, New York

[ott@rockefeller.edu](mailto:ott@rockefeller.edu)



# Topics

- Statistical significance
- Population substructure
- Multilocus approaches
- Purely epistatic traits

# Are results "significant"?

Benjamini et al (2001) *Behav Brain Res* 125, 279-284

- $n$  SNPs, each tested for association at significance level  $\alpha$  = probability of false positive result.
- Prob(any SNP is significant) =  $1 - (1 - \alpha)^n \approx n\alpha$ .
- Bonferroni correction:  
 $p \rightarrow p \times n$ , or  $\alpha \rightarrow \alpha/n$
- Number of SNPs with false discovery rate,  $FDR < 0.05$ .

*Example for FDR calculation (Benjamini-Hochberg method)*

12 genes, all  $n = 66$  pairwise tests for correlation in methylation status in colon cancer. 5 pairs are significant. Bonferroni criterion =  $0.05/66 = 0.0008$ : only 2 pairs are significant.

gene1	gene2	pi	rank, i	$i * 0.05/66$
<b>p19</b>	<b>RARb</b>	<b>0.0001</b>	1	0.0008
<b>p16</b>	<b>TIMP3</b>	<b>0.0002</b>	2	0.0015
<b>DAPK</b>	<b>p21</b>	<b>0.0012</b>	3	0.0023
<b>MGMT</b>	<b>RARb</b>	<b>0.0016</b>	4	0.0030
<b>RARb</b>	<b>TIMP3</b>	<b>0.0023</b>	5	0.0038
DAPK	GSTP1	0.0053	6	0.0045
GSTP1	p21	0.0053	7	0.0053
ECAD	GSTP1	0.0108	8	0.0061
...	...	...	...	...
GSTP1	MGMT	0.9364	65	0.0492
APC	RARb	0.9878	66	0.0500

# Significance of Results

Cheverud (2001) *Heredity* 87, 52-58

- Bonferroni and FDR criteria are valid for dependent data but are conservative, low power.
- Cheverud method computes an effective number,  $n_{\text{eff}} < n$ , of independent SNPs and uses this in the Bonferroni correction:
  1. Compute correlation matrix for genotype codes (AA = -1, AG = 0, GG = 1) of  $n$  SNPs
  2. Compute  $n$  eigenvalues,  $\lambda_i$  (principal components) and their variance,  $v = \Sigma(\lambda_i - 1)^2 / (n - 1)$ .
  3.  $n_{\text{eff}} = n[1 - (n - 1)v/n^2]$
- Permutation testing is more reliable

# Permutation Tests

- Need distribution of test statistic under no association
- Create non-association data sets by permuting *case* and *control* labels.
- Most useful for (1) unknown null distribution of test statistic and (2) dependent tests (dense SNPs)

# Replication

- Example of a non-replication:
  - Siddiqui *et al* (2003, NEJM): Association of SNP to multidrug resistance in epilepsy; 200 cases, 115 controls
  - Tan et al (2004, Neurology): Twice as many observations, no confirmation.
- Correcting for multiple testing → experiment-wise (overall) significance level,  $\alpha = 0.05$ , or FDR = 0.05
- Low prior probability,  $\phi$  → low posterior probability that association is true (low power)
- Thomas & Clayton (2004) *J Natl Cancer Inst* 96, 421:  
 $\phi = 1:1000 = 0.001$

# Is a Significant Result a True Positive Result? 🤔

Ott (2004) *Neurology* 63, 955-958 (editorial)

- Even though a disease association is statistically significant with proper correction for multiple testing, it might still be a false positive result.
- Replication has been advocated as a check whether a significant result is “real”.
- Many published “significant” results cannot be replicated.

# Posterior Probability that Significant Result is Real

Overall		<i>Power</i>		
$\alpha$	<i>Prior</i>	90%	50%	20%
0.05	0.100	0.67	0.53	0.31
	0.010	0.15	0.09	0.04
	0.001	0.02	0.01	0
0.01	0.100	0.91	0.85	0.69
	0.010	0.48	0.34	0.17
	0.001	0.08	0.05	0.02
0.005	0.100	0.95	0.92	0.82
	0.010	0.65	0.5	0.29
	0.001	0.15	0.09	0.04

*Recommendation:* Significance level, corrected for multiple testing, should be no more than 0.005

# Population Substructure (Heterogeneity)

- *Pritchard method*: Based on unassociated SNPs, identify more homogeneous portions of data. Analyze each of these separately.  
<http://pritch.bsd.uchicago.edu/software.html>
- *Genomic Control* (B. Devlin): Heterogeneity leads to apparent association with unassociated SNPs. Subtract resulting  $\chi^2$  from the  $\chi^2$  in your study.

# Data Subdivision

- Pritchard method, based on unassociated SNPs
- Identify groups of individuals with similar non-genetic risk factors, each group  $\rightarrow$   $p$ -value.
- Sparse tables in case-control studies:
  - Exact methods (*StatXact* program) rather than table values of  $\chi^2$
  - Permutation tests
- Combine  $p$ -values via Fisher's method. Analogous to blocked design in ANOVA; efficient if blocks have an effect. Example: Low education = risk factor for obesity (OR = 3.8; *Eur J Epidemiol* **19**:33, 2004)
- Extreme grouping: Matched case-control data. Not generally analyzed under this design.

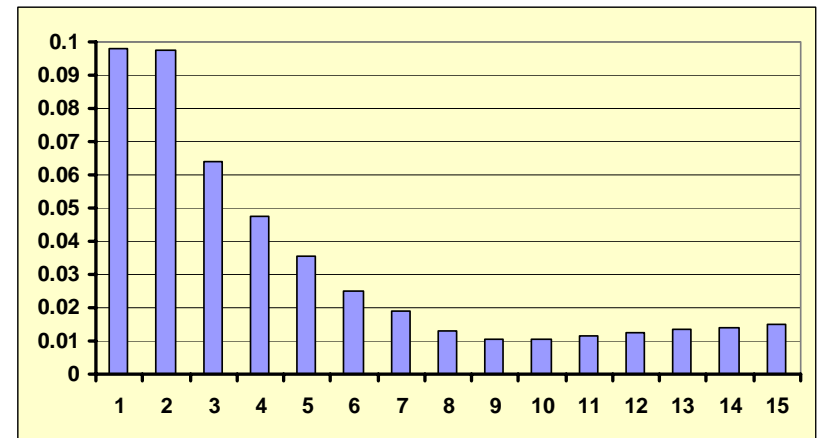
# Multi-Locus Analysis Methods

- Most case-control studies do not take into account the multi-locus nature of complex traits.
- Aim: Analyze multiple SNPs/genes jointly.  
*Two classes:*
  - 1. Combine single-locus statistics over multiple SNPs (wherever they are in genome)
  - 2. Look for patterns of genotypes at SNPs in different genomic locations

# Sums of single-marker statistics: *Set Association* method

Hoh et al. (2001) *Genome Res* **11**, 2115

- Let  $t_i$  = association statistic for  $i$ -th marker, ordered by size.
- Build sums, e.g.  $s_2 = t_1 + t_2$ ,  $s_3 = t_1 + t_2 + t_3$ .
- Sums larger than expected? Permutation tests,  $p$ -values
- Smallest  $p$ -value  $\rightarrow$  select
- Smallest  $p$  = single experiment-wise statistic  $\rightarrow$  overall significance level



# Application: Restenosis Data

Zee et al. (2002) *Pharmacogenomics J* 2:197

- Conventional approach:  $p > \mathbf{0.20}$ , corrected for multiple testing.
- Set association method: Smallest  $p = 0.011$  for sum containing 9 SNPs.
- Significance level associated with smallest  $p$  is **0.04**.
- *sumstat* computer program available at <http://linkage.rockefeller.edu/register/>

# Pattern Recognition Methods

Hoh & Ott (2003) *Nat Rev Genet* **4**, 701-709

- Neural networks (Lucek & Ott)
- CPM = combinatorial partitioning method (Charlie Sing, U Michigan)
- MDR = multifactor-dimensionality reduction method (Jason Moore, Vanderbilt U)
- LAD = logical analysis of data (P. Hammer, Rutgers U)
- Mining association rules, *Apriori* algorithm (R. Agrawal)
- Pairs of SNPs: Is association different in cases than in controls?

# Association Rules

<http://fuzzy.cs.uni-magdeburg.de/~borgelt/software.html>

- Developed by Agrawal, published in conference reports.
- Pattern recognition method to search for sets of articles purchased by consumers. Market basket analysis of large databases compiled from scanner data at cash registers, *Apriori* algorithm.
- Very fast. Few applications so far to genetic data (Toivonen et al [2000] *Am J Hum Genet* **67**, 133).

# Purely Epistatic Disease Model

Culverhouse et al. (2002) *Am J Hum Genet* **70**, 461

L.1 ↓L.2	<i>L.3 = 1/1</i>			<i>L.3 = 1/2</i>			<i>L.3 = 2/2</i>		
	<i>1/1</i>	<i>1/2</i>	<i>2/2</i>	<i>1/1</i>	<i>1/2</i>	<i>2/2</i>	<i>1/1</i>	<i>1/2</i>	<i>2/2</i>
<i>1/1</i>	0	0	<b>1</b>	0	0	0	0	0	0
<i>1/2</i>	0	0	0	0	<b>0.25</b>	0	0	0	0
<i>2/2</i>	0	0	0	0	0	0	<b>1</b>	0	0

Assume all allele frequencies = 0.50.

Heritability = 55%, prevalence = 6.25%.

# Expected Genotype Patterns

<i>L.1</i>	<i>L.2</i>	<i>L.3</i>	P(g)	E(#aff)	E(#unaff)
<i>1/1</i>	<i>2/2</i>	<i>1/1</i>	0.0156	25	0
<i>2/2</i>	<i>1/1</i>	<i>2/2</i>	0.0156	25	0
<i>1/2</i>	<i>1/2</i>	<i>1/2</i>	0.1250	50	10
other			0.8438	0	90
Sum			1	100	100

# Inference

- Given 3 disease SNPs:  $\chi^2 = 166.7$  (26 df),  $p = 1.76 \times 10^{-22}$ .
- 50,000 SNPs  $\rightarrow 2.1 \times 10^{13}$  subsets of size 3.
- Bonferroni-corrected  $p = 3.6 \times 10^{-9}$ .
- Alternative approach: Test all possible pairs of loci for interaction effects with suitable statistic, e.g.  $|\chi^2_{\text{cases}} - \chi^2_{\text{controls}}|$ , and find associated  $p$ -value via permutation samples (Hoh & Ott (2003) *Nat Rev Genet* **4**, 701-709).