

# Gene-gene and gene-environment interactions in genetic case-control association studies

Jurg Ott<sup>1</sup> & Josephine Hoh<sup>1,2</sup>

<sup>1</sup>Rockefeller University, New York

<sup>2</sup>Yale University, New Haven

ott@rockefeller.edu

# Rationale

- Modern technology allows for the creation of more and more experimental results, ie. data.
- Examples:
  - Microarray expression studies with 1000s of genes
  - Genetic linkage or association studies with large numbers of genetic marker loci.
- “Curse of dimensionality”: More variables (parameters to estimate) than observations.

# Heritable Diseases

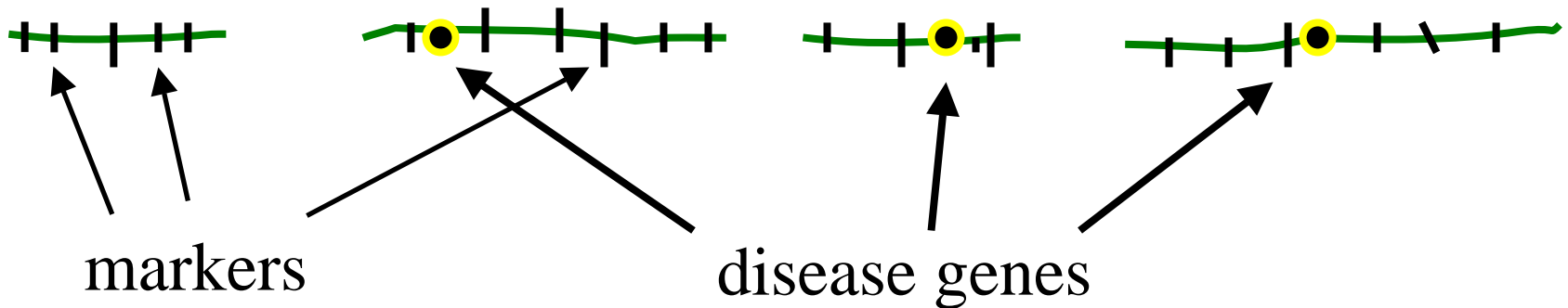
- **Rare Diseases**

- Mendelian inheritance
- Examples: Huntington disease, cystic fibrosis

- **Common Diseases**

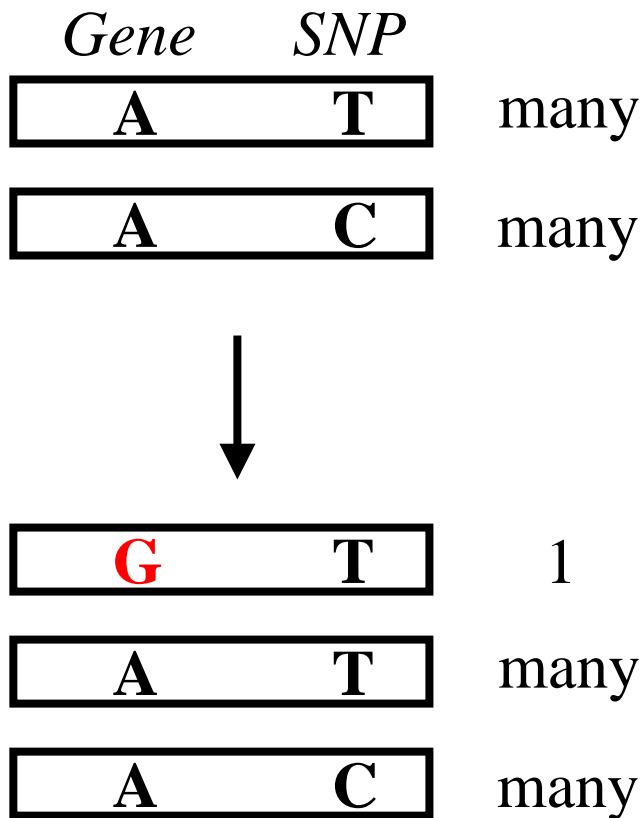
- Non-mendelian (“complex”) mode of inheritance. Examples: Diabetes, schizophrenia.
- Genetically relevant phenotype often unclear
- Multiple underlying susceptibility genes

# Genome Screens for Disease Loci



- Candidate genes: Focus on specific regions
- Unknown locations: Genome-wide screening with up to 800 microsatellites, or 1000s if not 100,000s of SNP markers.

# Linkage Disequilibrium (LD) Genetic Association



- Population expands  
→ >1 disease allele, **G**
- Crossovers → chromosomes with **G** - **C** alleles
- Motivates case-control studies

	T	C
G	1	0
A	many	many

# Establishing Association

	Marker Genotypes		
	G/G	G/T	T/T
cases	...	...	...
controls	...	...	...

Size of  $\chi^2$  shows significance of association.  
Effects of association within short range of a locus, in contrast to linkage analysis.

# One-by-One Approach

- Need to correct for multiple testing.
- **Linkage analysis:** For dense map of markers, testing each marker at  $\alpha = 0.00005$  ( $\text{lod} = 3.3$ ) leads to genome-wide sig. level of 0.05 (Lander & Kruglyak, *Nat Genet* **11**:241, 1995). Neighboring markers yield similar results; not so for association analysis.
- **Association analysis:** Independent data. Strong effects of multiple testing (loss of power).

# Two Classes of Approaches

Devlin et al (2003) *Genet Epidemiol* **25**, 36

- Model selection
  - Stepwise (logistic) regression
  - Main effects first, then model interactions
  - Aim: Prediction of response variable. May be non-sig.
- Significance testing
  - Aim: Control the number of falsely included genes or SNP markers
  - Bonferroni correction
  - Controlling False Discovery Rate (FDR)  
(Benjamini et al [2001] *Behav Brain Res* **125**, 279)

# FDR versus Significance Level

Devlin et al. (2003); Storey & Tibshirani (2003) *PNAS* **100**, 9440

	Test not signif.	Test significant	# tests
$H_0$ true	U	V	$m_0$
$H_0$ false	T	S	$m_1$
	$m - R$	R	$m$

- Avg. significance level =  $V/m_0$  (false pos.)
- Avg. FDR =  $V/R$  (need estimate)

# Complex Traits

- ... are due to interacting effects of environmental agents and multiple underlying susceptibility genes, each with small effect.
- Essentially none of the current methods address the multi-locus nature of complex diseases.
- Do they exist?

# Multiple Hits ... Digenic Diseases

Ming & Muenke (2002) *Am J Hum Genet* 71:1017 (review)

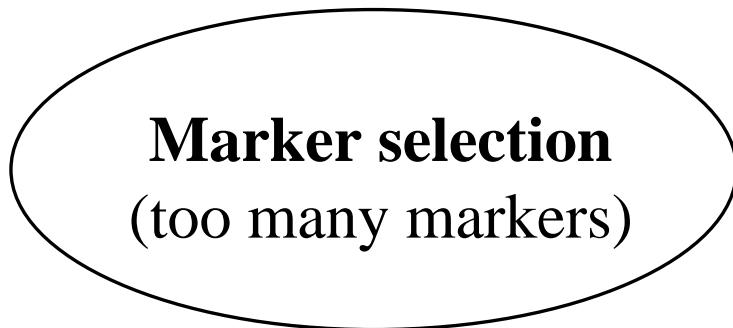
EFFECT AND PHENOTYPE	GENE 1		GENE 2	
	Mutation	Phenotype	Mutation	Phenotype
<b>Synergistic:</b>				
RP	<i>ROM1</i> <sup>+/G80insG</sup>	Normal	<i>RDS</i> <sup>+/L185P</sup>	Normal
RP	<i>ROM1</i> <sup>+/L114insG</sup>	Normal	<i>RDS</i> <sup>+/L185P</sup>	Normal
Bardet-Biedl	<i>BBS2</i> <sup>Y24X/Q59X</sup>	Normal	<i>BBS6</i> <sup>+/Q147X</sup>	Normal
Deafness	<i>GJB2</i> <sup>+/35delG</sup>	Normal	<i>GJB6</i> <sup>+/-</sup>	Normal
Deafness	<i>GJB2</i> <sup>+/L67delT</sup>	Normal	<i>GJB6</i> <sup>+/-</sup>	Normal
Hirschsprung	<i>RET</i> <sup>+/L647I</sup>	Normal	<i>EDNRB</i> <sup>+/S305N</sup>	Normal
Severe insulin resistance	<i>PPARG</i> <sup>+/A553delAAAAT</sup>	Normal	<i>PPP1R3A</i> <sup>+/C1984delAG</sup>	Normal
<b>Modifier:</b>				
Juvenile-onset glaucoma	<i>MYOC</i> <sup>+/G399V</sup>	Adult-onset glaucoma	<i>CYP11B1</i> <sup>+/R368H</sup>	Normal
Usher 1	<i>USH3</i> <sup>mut/mut</sup>	Usher 3	<i>MYO7A</i> <sup>+delG (exon 25)</sup>	Normal
Congenital nonlethal JEB	<i>COL17A1</i> <sup>R1226X/L855X</sup>	Juvenile JEB	<i>LAMB3</i> <sup>+/R635X</sup>	Normal
More severe ADPKD	<i>PKD1</i> <sup>+/mut</sup>	Less severe ADPKD	<i>PKD2</i> <sup>+/2152delA</sup>	Less severe ADPKD
More severe hearing loss	<i>DFNA1</i>	Mild hearing loss	<i>DFNA2</i>	Mild hearing loss
WS2/OA	<i>MITF</i> <sup>+/894delA</sup>	?WS2	<i>TYR</i> <sup>+/R402Q</sup>	Normal
More severe WS2/OA	<i>MITF</i> <sup>+/894delA</sup>	?WS2	<i>TYR</i> <sup>R402Q/R402Q</sup>	Normal

# Proposed Analysis Strategy

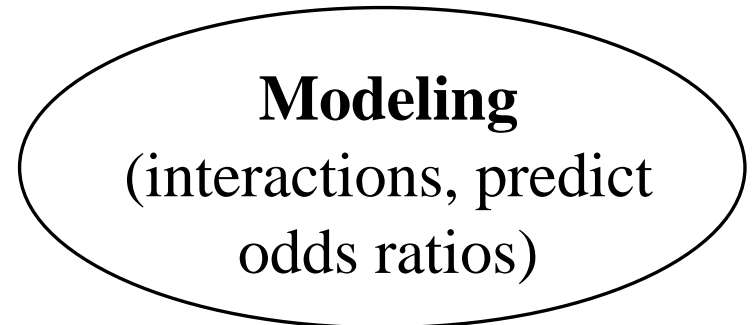
Hoh et al. (2000) *Ann Hum Genet* **64**, 413

- **Aim:** To find a set of genes or SNP loci with significant effect, e.g. disease association
- **General principle:** 2-step analysis

Step 1



Step 2



# Approaches

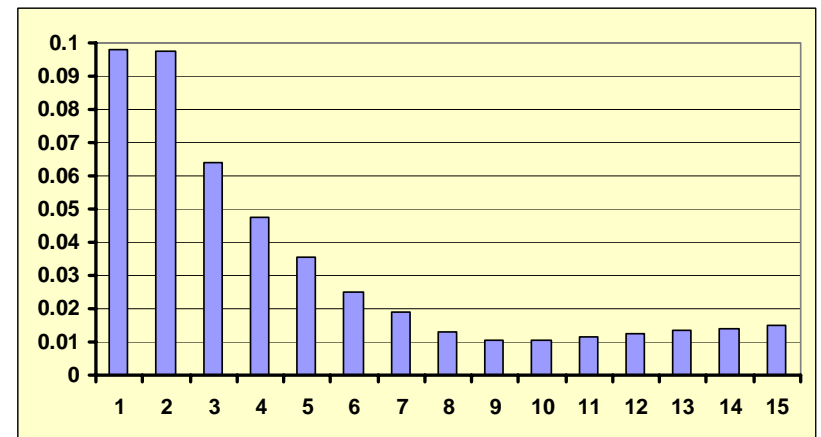
Hoh & Ott (2003) *Nat Rev Genet* **4**, 701-709

- Neural networks (Lucek & Ott)
- Sums of single-marker statistics (Hoh and Ott)
- CPM = combinatorial partitioning method (Charlie Sing, U Michigan)
- MDR = multifactor-dimensionality reduction method (Jason Moore, Vanderbilt U)
- Bump Hunting (Friedman)
- LAD = logical analysis of data (P. Hammer, Rutgers U)
- Mining association rules, *Apriori* algorithm (R. Agrawal)
- Special approaches for microarray data
- All pairs of genes

# Sums of marker statistics: *Set Association* method

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

- Let  $t_i$  = statistic of  $i$ -th gene, 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:  $p > 0.20$ , corrected for multiple testing
- Set association method: Smallest  $p = 0.011$  for sum containing 10 SNPs in 9 different genes.
- Significance level associated with smallest  $p$  is 0.04.

# Association Rules

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

- Developed by Agrawal, published in conference reports, implemented in *Apriori* algorithm.
- 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.
- Very fast. Few applications so far to genetic data (Toivonen et al [2000] *Am J Hum Genet* **67**, 133).

# Purely Epistatic Traits

- “Complex traits due to multiple interacting genes”
- No main effects (single gene effects), only interactions causing disease → set association analysis (based on single-gene statistics) not useful unless modified.

# 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  $p = 3.6 \times 10^{-9}$ .
- More manageable approach: Test all possible pairs of loci for interaction effects whether they are different in case and control individuals (Hoh & Ott (2003) *Nat Rev Genet* **4**, 701-709).