Genetic Mapping of Complex Trait Genes in Association Studies

o

Presentation at the

NIA Alzheimer's Genetics Symposium

New York, October 12, 2002

o

Jurg Ott

Rockefeller University, New York

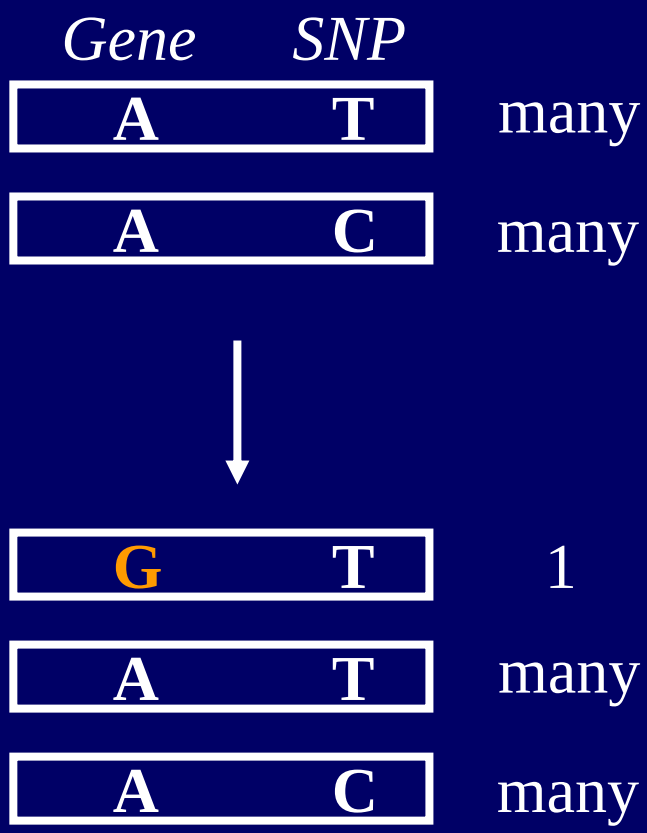
o

ott@rockefeller.edu

1. Background: Linkage disequilibrium, case-control association studies
2. Genome screens I: Evaluating effects of multiple disease genes simultaneously
3. Genome screens II: Patterns of genotypes

Linkage Disequilibrium (LD)

Origin in single mutation



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

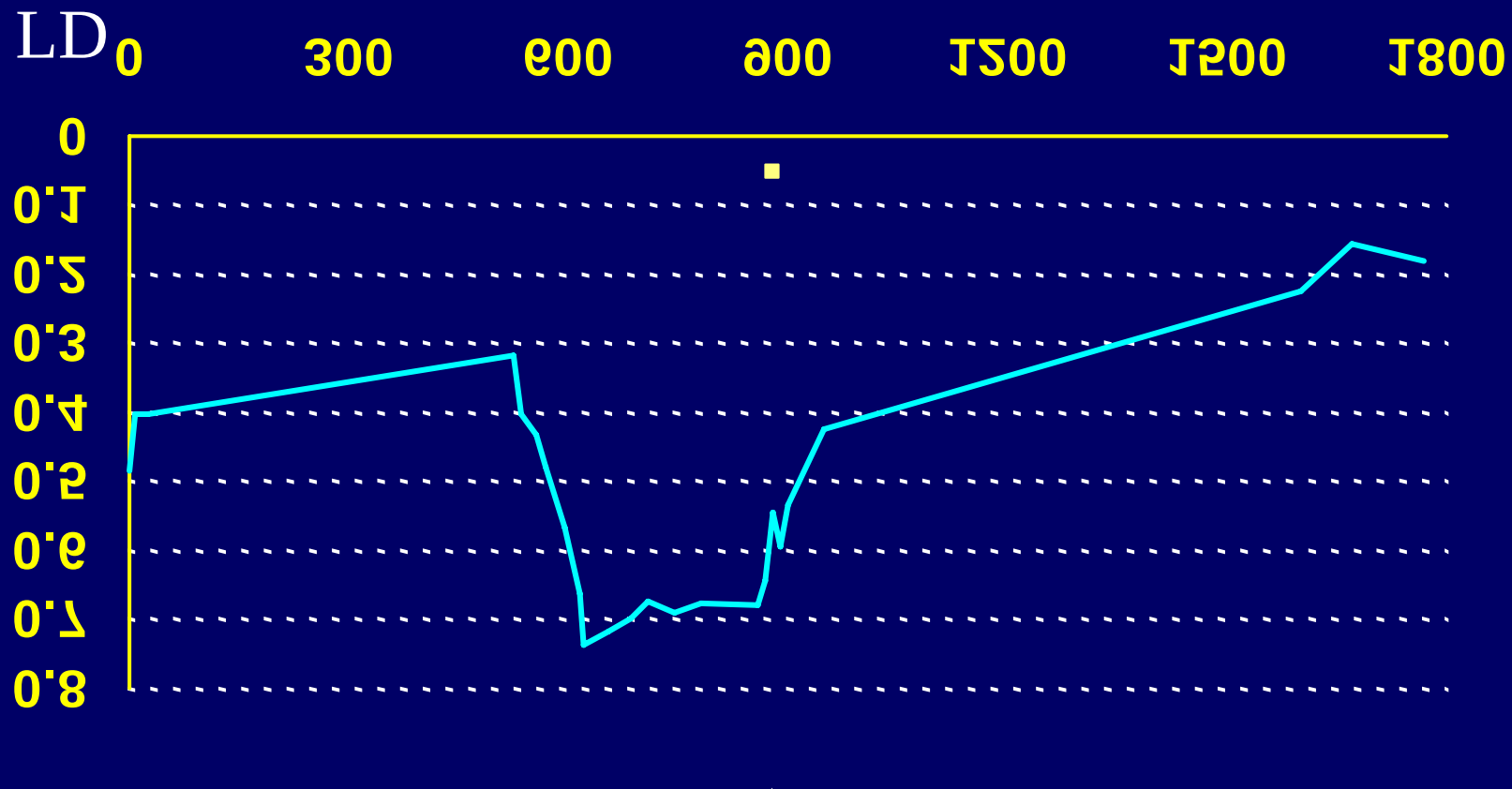
	T	C
G	1	0
A	many	many

Origin of LD: Idealized Situation!

- Population with small number of founder individuals, rapidly expanding → strong LD.
- Most disease genes show multiple mutations (alleles), having occurred at different times → strength of LD (measured by D') reduced. *Cystic fibrosis*: >500 mutations.
- LD is the basis for **association studies**.

Linkage Disequilibrium at CF locus

Kerem et al (1989) *Science* 245:1073

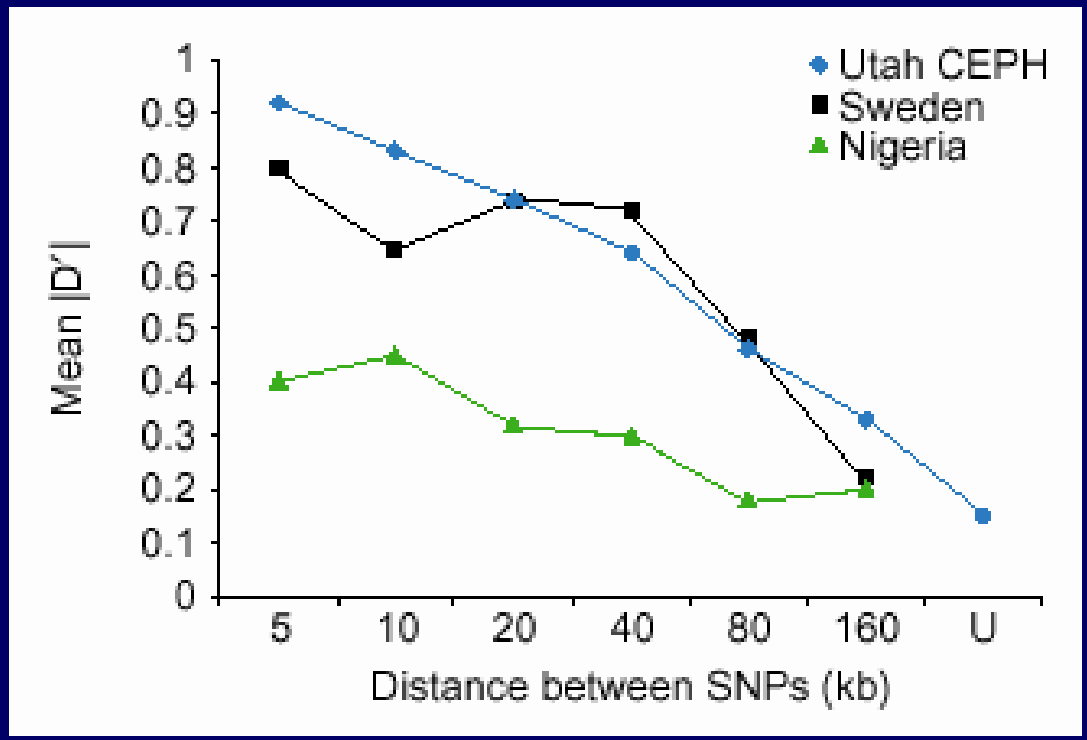


LD mapping for Complex Traits

- Complex traits: Diabetes, schizophrenia, ...: Heritable, no mendelian inheritance
- Due to multiple interacting genes on different chromosomes
- Genome-wide association studies with large numbers of SNP markers
- Difference in allele or genotype frequencies between case and control individuals $\rightarrow$ evidence for LD.

How Many SNPs?

Weiss & Clark (2002) *Trends in Genetics* **18**, 19



Mean LD (n = 48 Utah and Sweden; n = 96 Nigeria; U = unlinked). Data from Reich et al. (2001) *Nature* **411**, 199.

3200 Mb/(2 × 80 kb) = 20,000 SNPs.

Daly et al. (2001) *Nat Genet* **29**, 229: "Blocks" of DNA with strong LD, low LD segments.

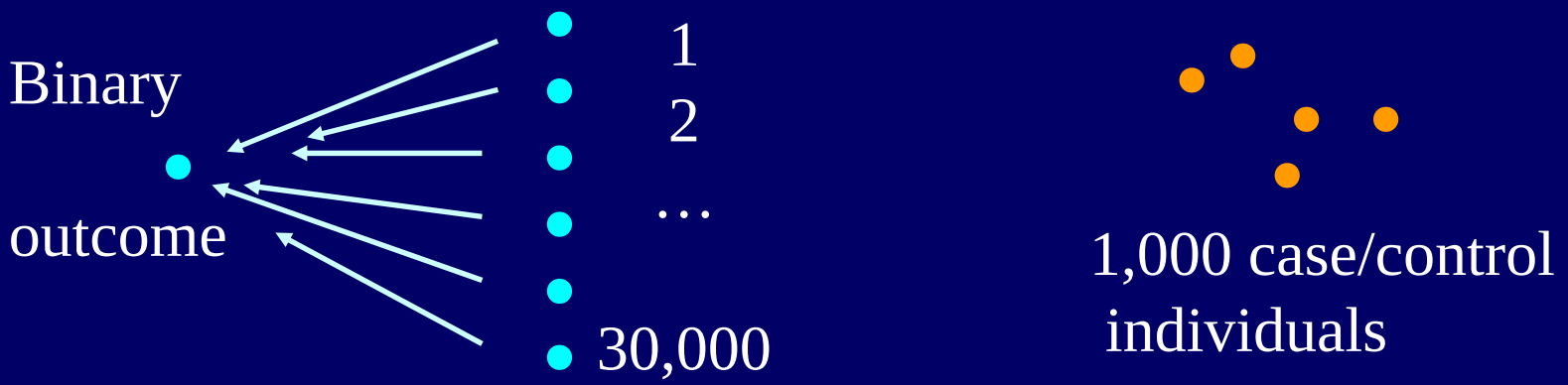
Current Approaches

Most genome screens evaluated on a marker-by-marker basis.

	SNP Genotype		
	1/1	1/2	2/2
Cases	…	…	…
Controls	…	…	…

Size of χ^2 shows significance of association

Problem



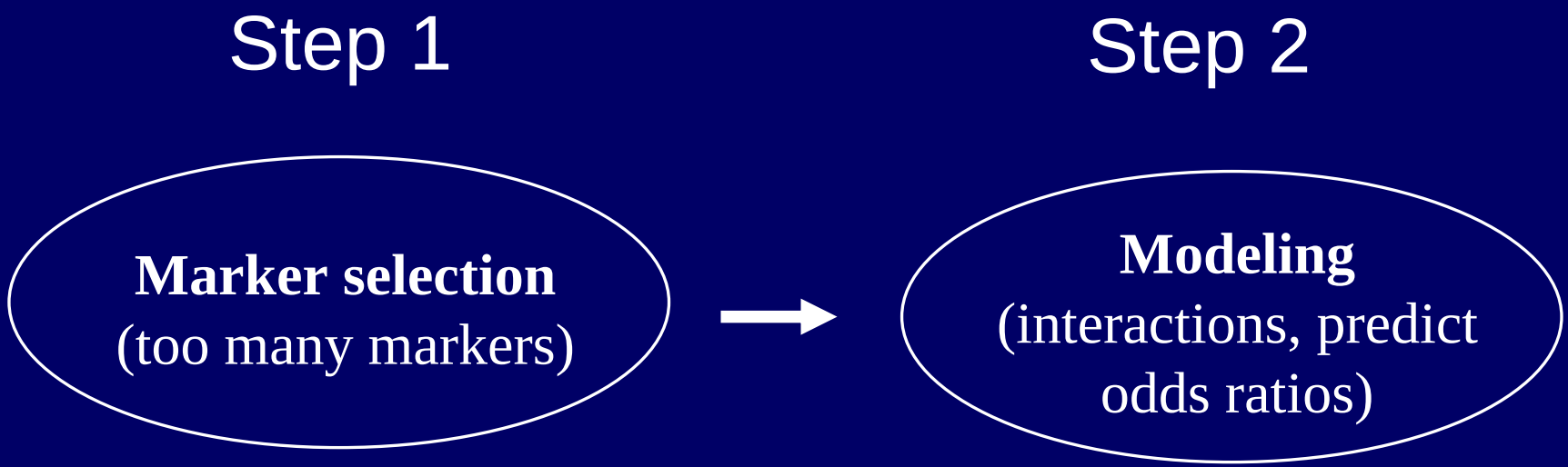
- Want to allow for interactions between susceptibility genes (i.e., marker loci).
- Ideally, analyze all data jointly.
- Number of variables >> number of observations: “Curse of dimensionality”

1. Background: Linkage disequilibrium, case-control association studies
2. Genome screens I: Evaluating effects of multiple disease genes simultaneously
3. Genome screens II: Patterns of genotypes

Proposed Analysis Strategy

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

- **Aim**: To find a *set* of SNP loci with significant association to disease
- **General principle**: 2-step analysis



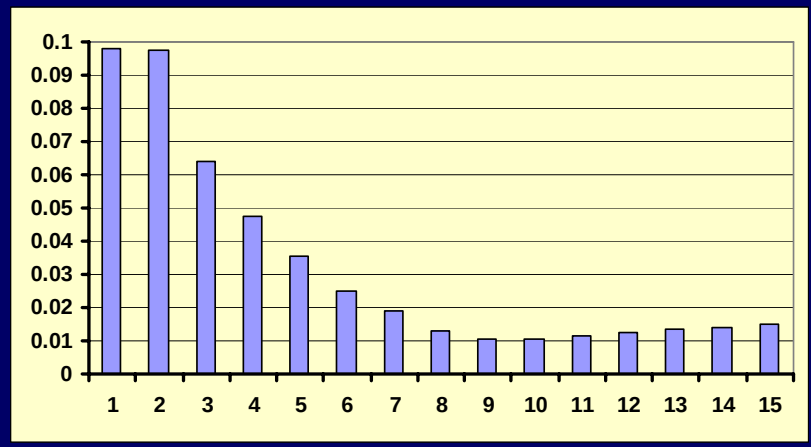
Marker Selection Procedures

- Pick markers with individually significant association. More sophisticated approaches?
- For a specific tissue, compare genes expressed in cases and controls (psoriasis, disease vs. normal skin)
- Nested bootstrap approach. Hoh et al. (2000) *Ann Hum Genet* **64**, 413
- *Set Association* approach (see below)

Set Association Approach

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

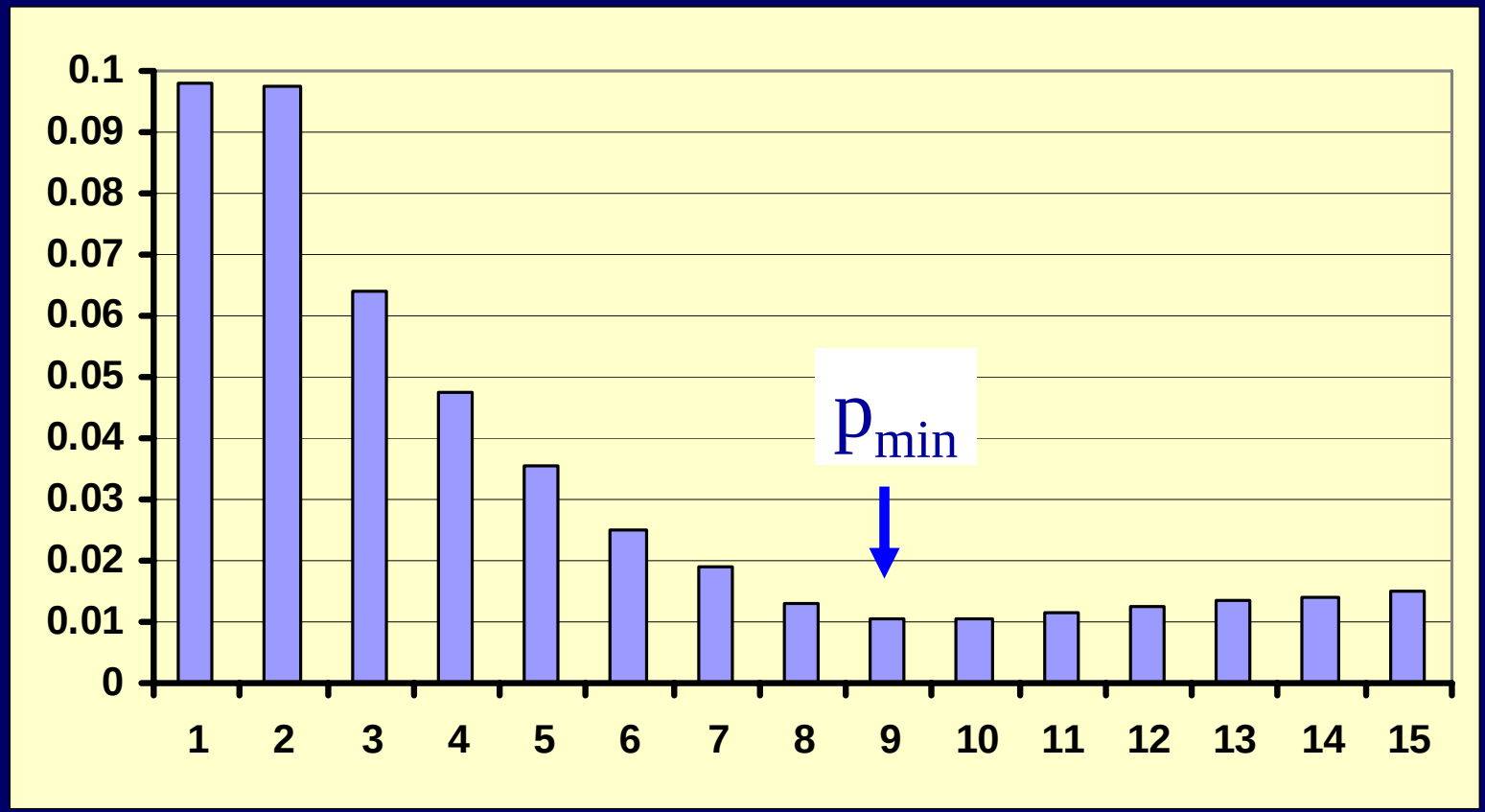
- At each SNP, compute association statistic, *t* (chi-square, or product $\chi^2_{ASSOC} \times \chi^2_{HWD}$)
- Combine information over multiple SNPs: sum up *t*'s
- Which SNPs? Order SNPs by *t* values, build sums, e.g. $s_2 = t_1 + t_2$, $s_3 = t_1 + t_2 + t_3$.
- Sums larger than expected? Permutation tests, p-values
- Smallest *p*-value $\rightarrow$ select
- Smallest *p* = single experiment-wise statistic $\rightarrow$ overall significance level



Example: Heart Disease
(candidate genes, no genome screen)
Zee et al. (2002) *Pharmacogenomics J* **2**:197

- **Subjects**: 779 heart disease patients, angioplasty. After 6 months: 342 cases (restenosis), 437 controls (no restenosis).
- **Marker data**: 89 SNPs in 63 candidate genes
- **Complex trait**: multiple genes controlling candidate pathway. Each marker is in one of the underlying genes.
- Conventional approach: $p > 0.20$, genome-wide

Application to heart disease data
(overall $p = 0.04$)



1. Background: Linkage disequilibrium, case-control association studies
2. Genome screens I: Evaluating effects of multiple disease genes simultaneously
3. Genome screens II: Patterns of genotypes

Genotype Patterns

- Why have LD studies not been more successful?
  - Small sample sizes
  - Analysis methods do not combine information from multiple SNPs at different locations. Exception: Nelson et al. (2001) *Genome Res* 11:458-470
  - Pool “main” effect, most heritability via interactions?
- Reported cases of disease through interactions only: Savage et al. (2002) *Nat Genet* 31:379-384

Purely Epistatic Disease Models

Culverhouse et al. (2002) *Am J Hum Genet* 70:461-471

L3→	*1/1*			*1/2*			*2/2*		
L2→ L1↴	*1/1*	*1/2*	*2/2*	*1/1*	*1/2*	*2/2*	*1/1*	*1/2*	*2/2*
1/1	0	0	1	0	0	0	0	0	0
1/2	0	0	0	0	0.25	0	0	0	0
2/2	0	0	0	0	0	0	1	0	0

$p_1 = p_2 = 0.5$, heritability = 60%, $\lambda_{sib} = 2.1$, prevalence = 0.06

Purely Epistatic Disease: Genotype Patterns

Genotype at locus			*Prob*	*Pene-trance*	*Expected no.*	
L1	L2	L3			*cases*	*controls*
1/1	*2/2*	*1/1*	0.0156	1	25	0
2/2	*1/1*	*2/2*	0.0156	1	25	0
1/2	*1/2*	*1/2*	0.1250	0.25	50	10
other			0.8438	0	0	90
		Sum	1	-	100	100

Pattern search

- Assume 1,000 SNPs with SNPs 1, 2, 3 being the disease loci.
- Total number of subsets of 3 loci: $n = (1000)_3/3! = 166{,}167{,}000$.
- Loci 1, 2, 3: Chi-square = 166.67 (26 df), $p = 1.76 \times 10^{-22}$.
- Bonferroni: $np = 3 \times 10^{-14}$ (conservative)