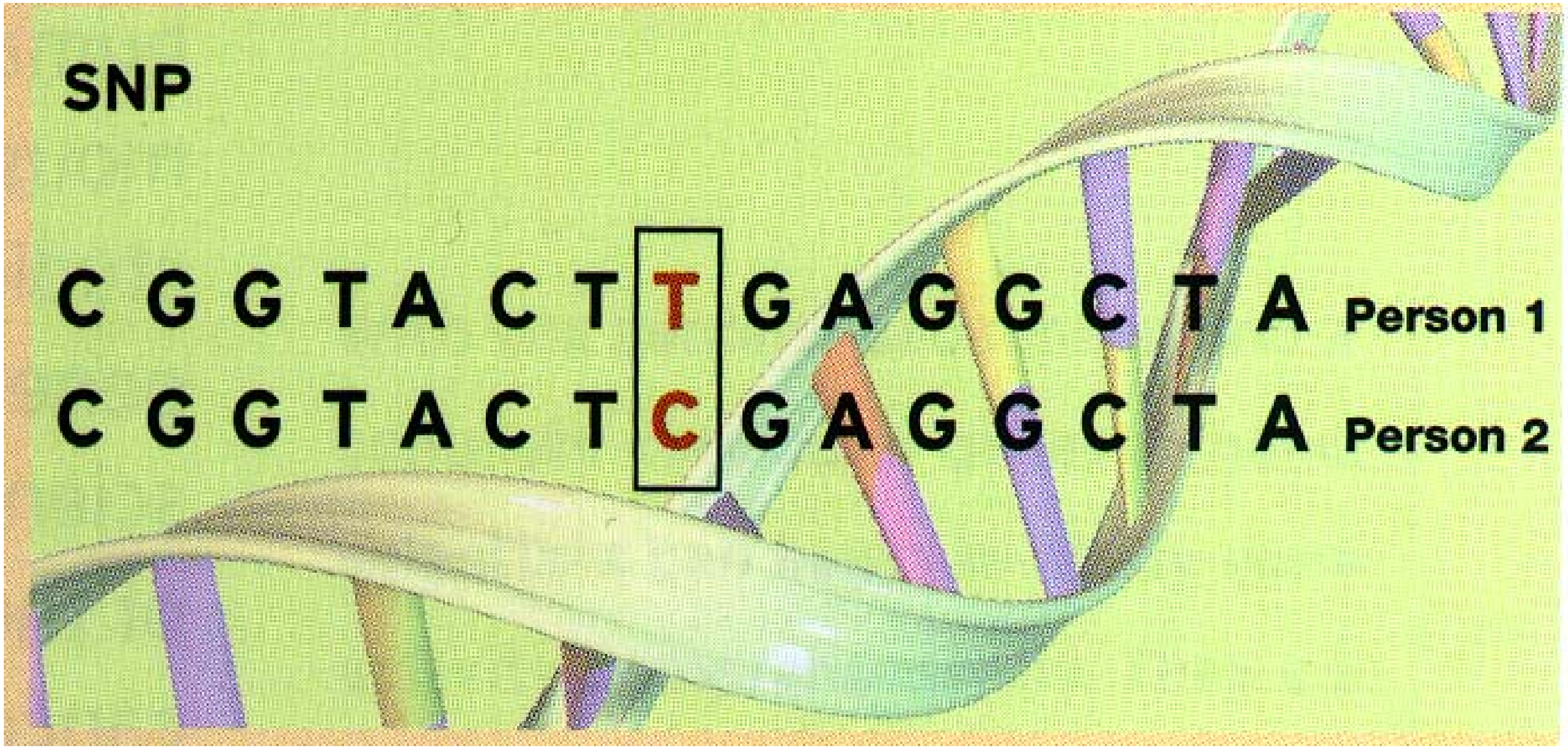
Genetic Association Analysis

Jurg Ott
Rockefeller University, New York

ott@rockefeller.edu

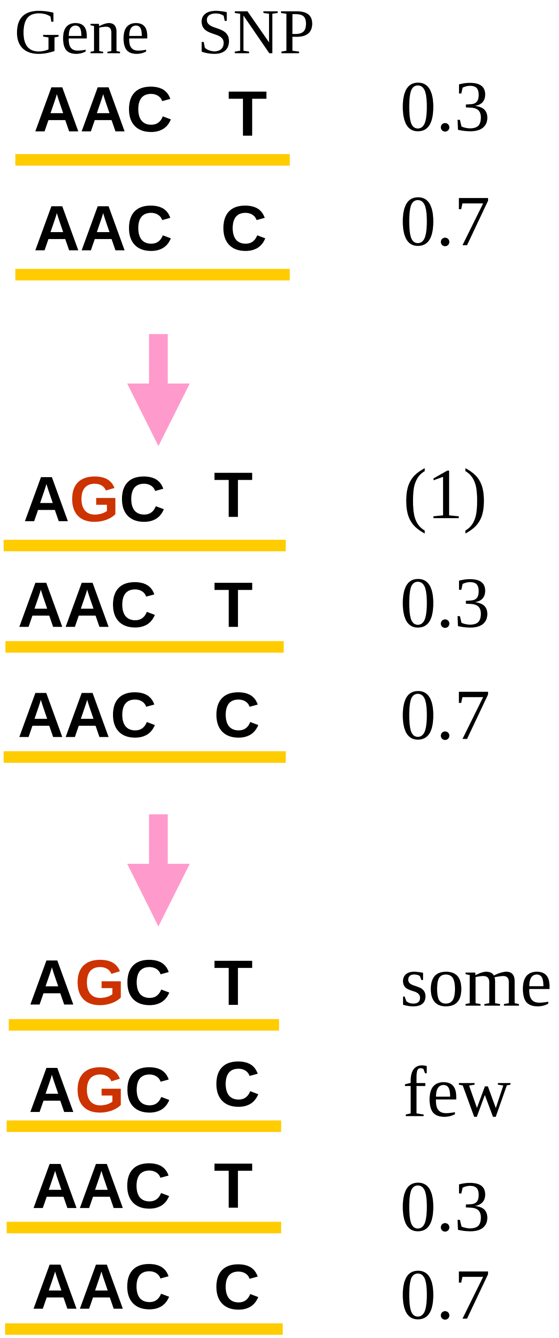
Single Nucleotide Polymorphism


SNP
C G G T A C T T G A G G C T A Person 1
C G G T A C T C G A G G C T A Person 2

Some people have a different base at a given location

Linkage Disequilibrium (LD)

Origin in single mutation



Gene SNP

AAC T 0.3

AAC C 0.7

AGC T (1)

AAC T 0.3

AAC C 0.7

AGC T some

AGC C few

AAC T 0.3

AAC C 0.7

	T	C
G	some	few
A	0.3	0.7

Measuring the Amount of LD

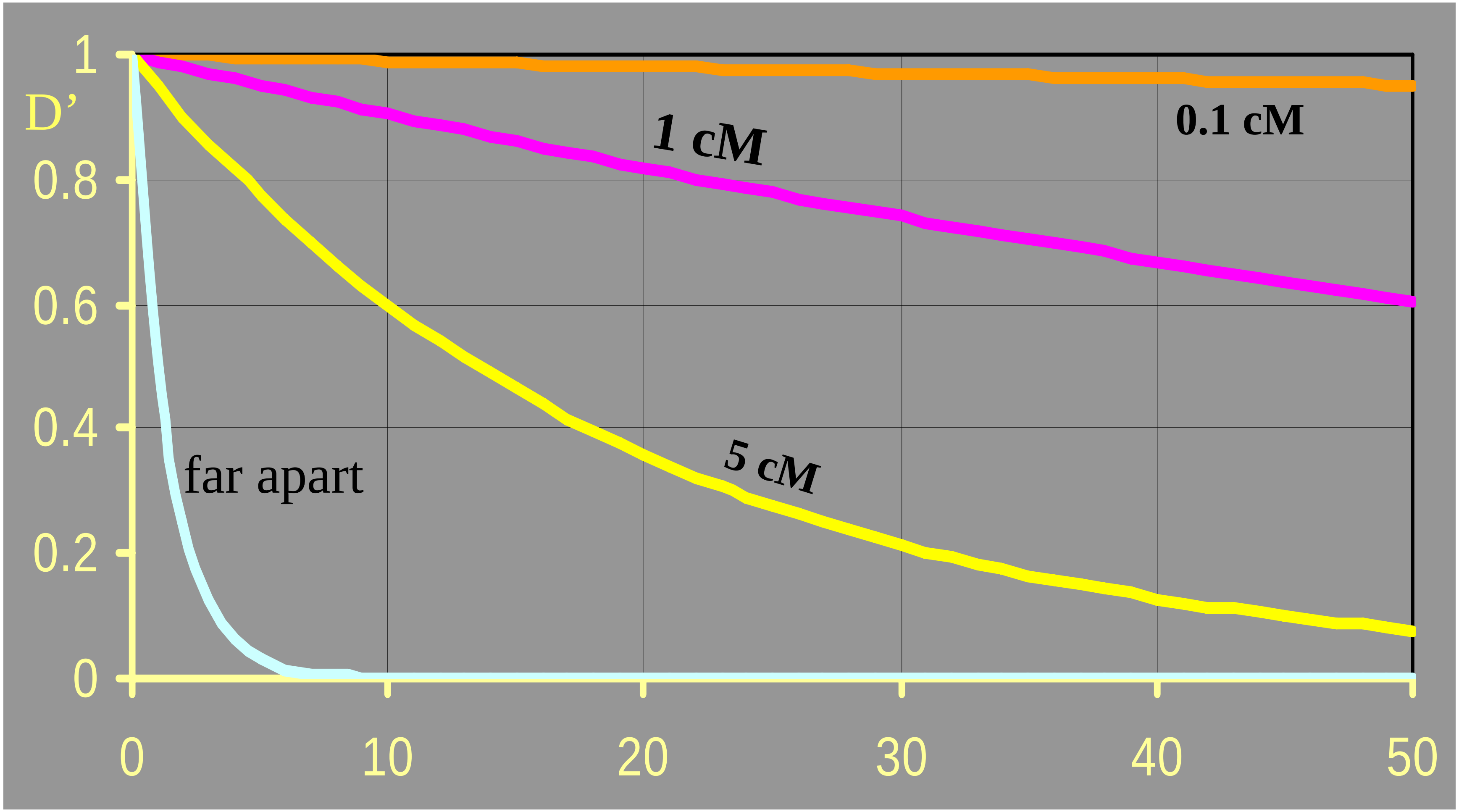
SNP 1	SNP 2	
	G	C
A	P(AG) = P(A)P(G) + D	...
T	...	...

$$D = \mathrm{P}(AG) - \mathrm{P}(A)\,\mathrm{P}(G)$$

$$D' = \begin{cases} D / D_{\mathrm{max}} & \text{if } D > 0 \\ D / D_{\mathrm{min}} & \text{if } D < 0 \end{cases}$$

Many other measures used. Correlation coefficient, *r*, with alleles numbered 0 and 1. All measures ± depend on allele frequencies.

Decay of linkage disequilibrium (LD) over time, single ancestral mutation



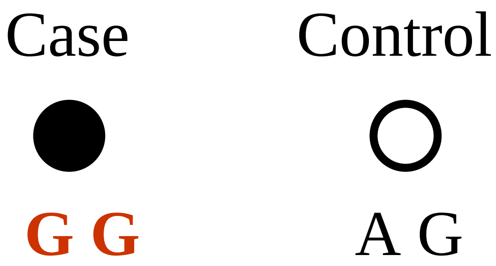
Number of Generations

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.
- LD is the basis for **association studies**.

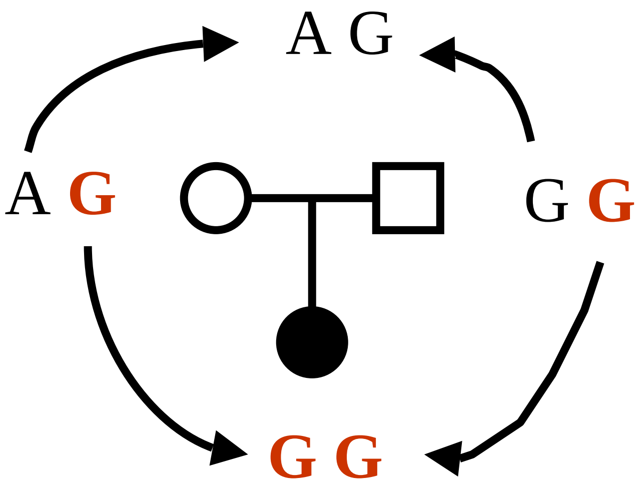
Data Designs for Association Studies

Population controls: Easy to collect, efficient yet prone to population stratification (problems can be overcome). Companies.



versus

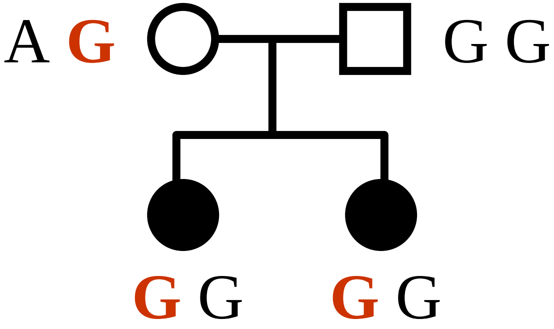
Family based controls: Compare alleles transmitted to affected child with those not transmitted. Academic researchers.

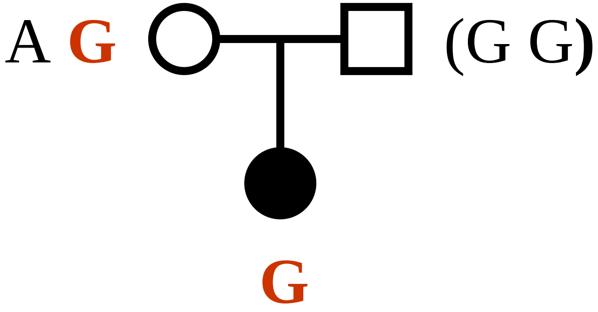


Data Designs: Family-based Association Tests

- Basic layout: Trio family
- More general approaches: Multiple affected and unaffected offspring, with or without parents
    - TDT: Multiple affected offspring
    - Teng and Risch (1999) *Genome Res* **9**, 234
    - Xu, Horvath and Laird: FBAT program (and references) available at http://www.biostat.harvard.edu/~fbat/default.html

Linkage versus Association





- *Linkage*: Excess of proportion of alleles shared (any alleles).
- *Association*: Excess of specific allele transmitted (TDT).

SNP in case-control study: Alleles or Genotypes?

- Compare allele frequencies: 1 df test good, but strictly valid only with HWE. May miss genotype effects.
- Genotypes: 2 df, more information than necessary?
- Test for trend in direction of increasing number of “2” alleles: 1 df, valid with HWD.
- Test each of 2 df

Test for Trend

Armitage P (1971) *Statistical Methods in Medical Research*, Wiley

	1/1	1/2	2/2	Total
cases	r_1	r_2	r_3	R
controls	n_1 - r_1	n_2 - r_2	n_3 - r_3	N - R
	n_1	n_2	n_3	N

$$X^2 = \frac{N[N(r_1 - r_3) - R(n_1 - n_3)]^2}{R(N - R)[N(n_1 + n_3) - (n_1 - n_3)^2]}$$

cases	19	29	24
controls	497	560	269

X^2 = 7.19, 1 df

Two independent df

Snedecor & Cochran (1969) *Statistical Methods*, Iowa State University Press, p. 309

- Test two 2 × 2 tables (LR chi-square)
  - Compare the two homozygotes (allelic effects)
  - Compare het versus sum of hom’s
- Each table yields a chi-square with 1 df, whose sum will be equal to the chi-square with 2 df for the 2 × 3 table of genotypes.

Example Data (X^2 computed as LR chi-square)

Total table

19	29	24
497	560	269

$X^2 = 7.32$, 2 df
$p = 0.026$

Allele effect

19	24
497	269

$X^2 = 7.21$, 1 df
$p = 0.007$

Genotype deviation

29	43
560	766

$X^2 = 0.11$, 1 df
$p = 0.743$

Most of the association is due to alleles

Alleles (not valid!)

67	77
1554	1098

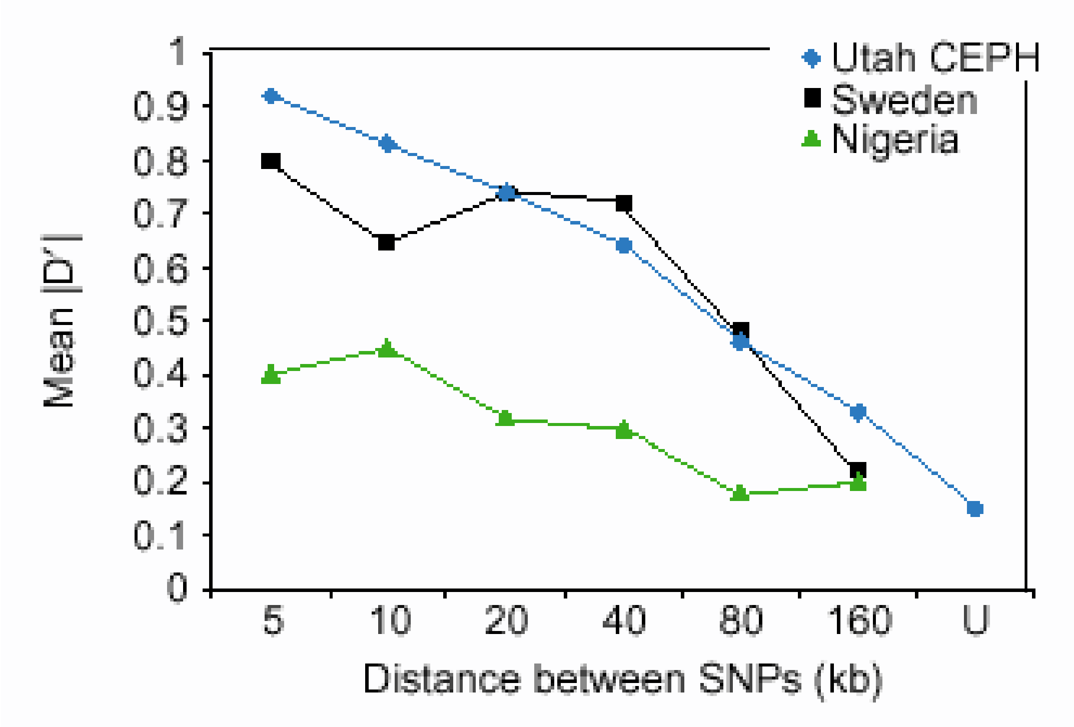
$X^2 = 8.05$, 1 df
$p = 0.005$

Genome Screens for Disease Loci

- Complex trait due to multiple possibly interacting disease genes
- Candidate genes: Focus on specific regions
- Unknown locations: Genome-wide screening with 1000s or 100,000s of SNP markers.

LD Between SNPs Versus Physical Distance

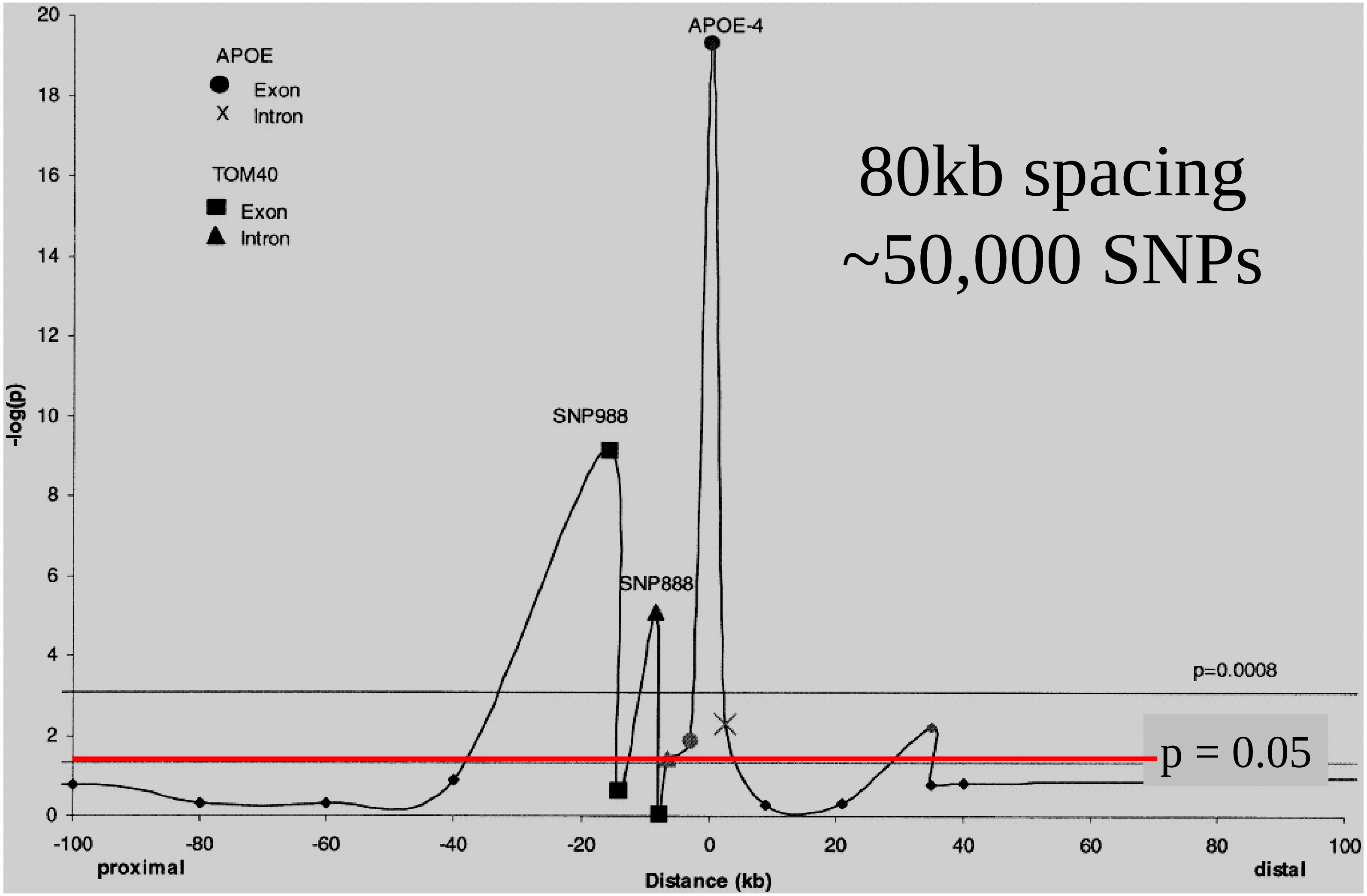
Weiss & Clark (2002) *Trends in Genetics* **18**, 19



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

Example: LD Around Alzheimer Disease Gene

Martin *et al*. (2000) *Am J Hum Genet* **67**, 383



80kb spacing
~50,000 SNPs

p = 0.05

Current Approaches

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

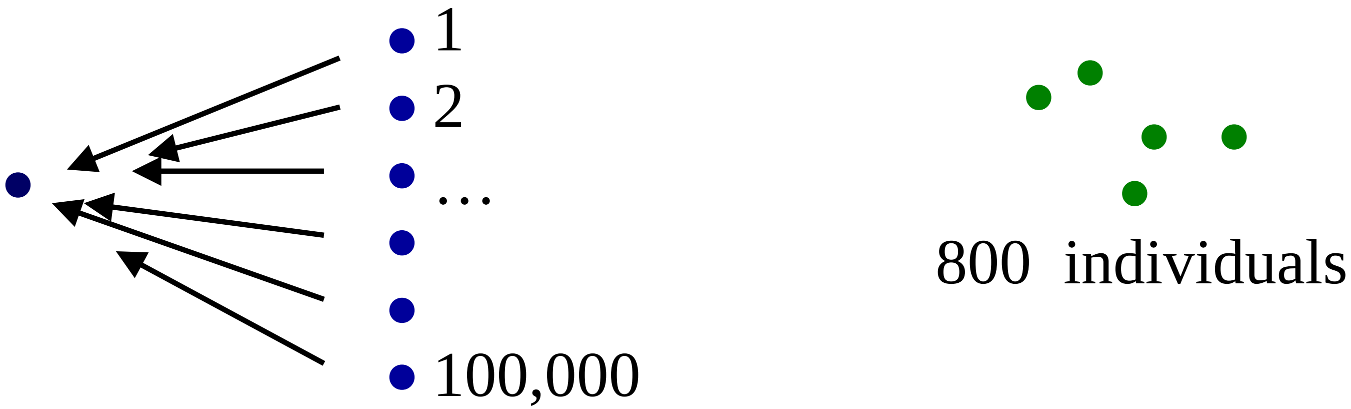
	Allele 1	Allele 2
Cases	…	…
Controls	…	…

Size of χ^2 shows significance of association

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 of approaches*:
  - Combine single-locus statistics over multiple SNPs (wherever they are in genome)
  - Look for patterns of genotypes at SNPs in different genomic locations

Problem

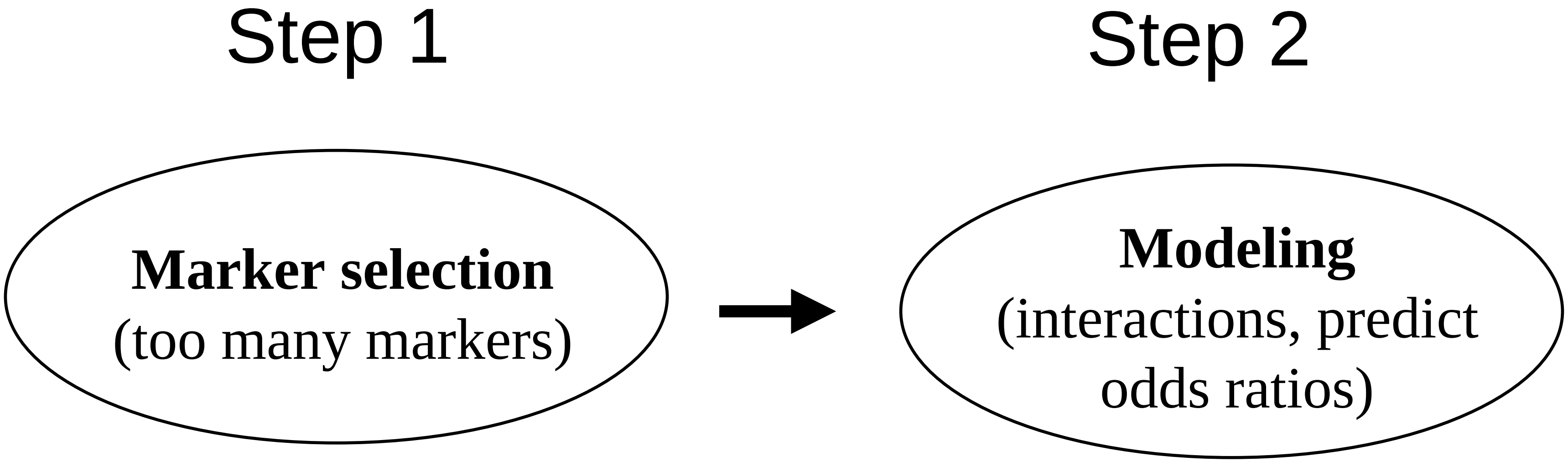


- Want to allow for interactions between susceptibility genes (i.e., marker loci).
- Ideally, analyze all data jointly.
- Number of variables (markers) is much larger than number of observations.

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



Step 1

Marker selection
(too many markers)

Step 2

Modeling
(interactions, predict odds ratios)

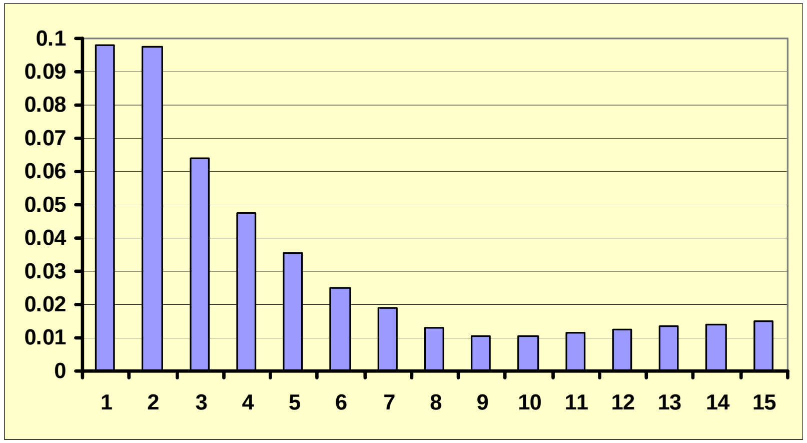
Step 1: Marker Selection Procedures

- 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, *s*
- Build sum over 1, 2, 3, etc. highest *s* values
- Evaluate significance of given sum by permutation test
- Sum with smallest *p*-value → marker selection
- Smallest *p* = single statistic → overall *p*-value



Example: Disease Data Set

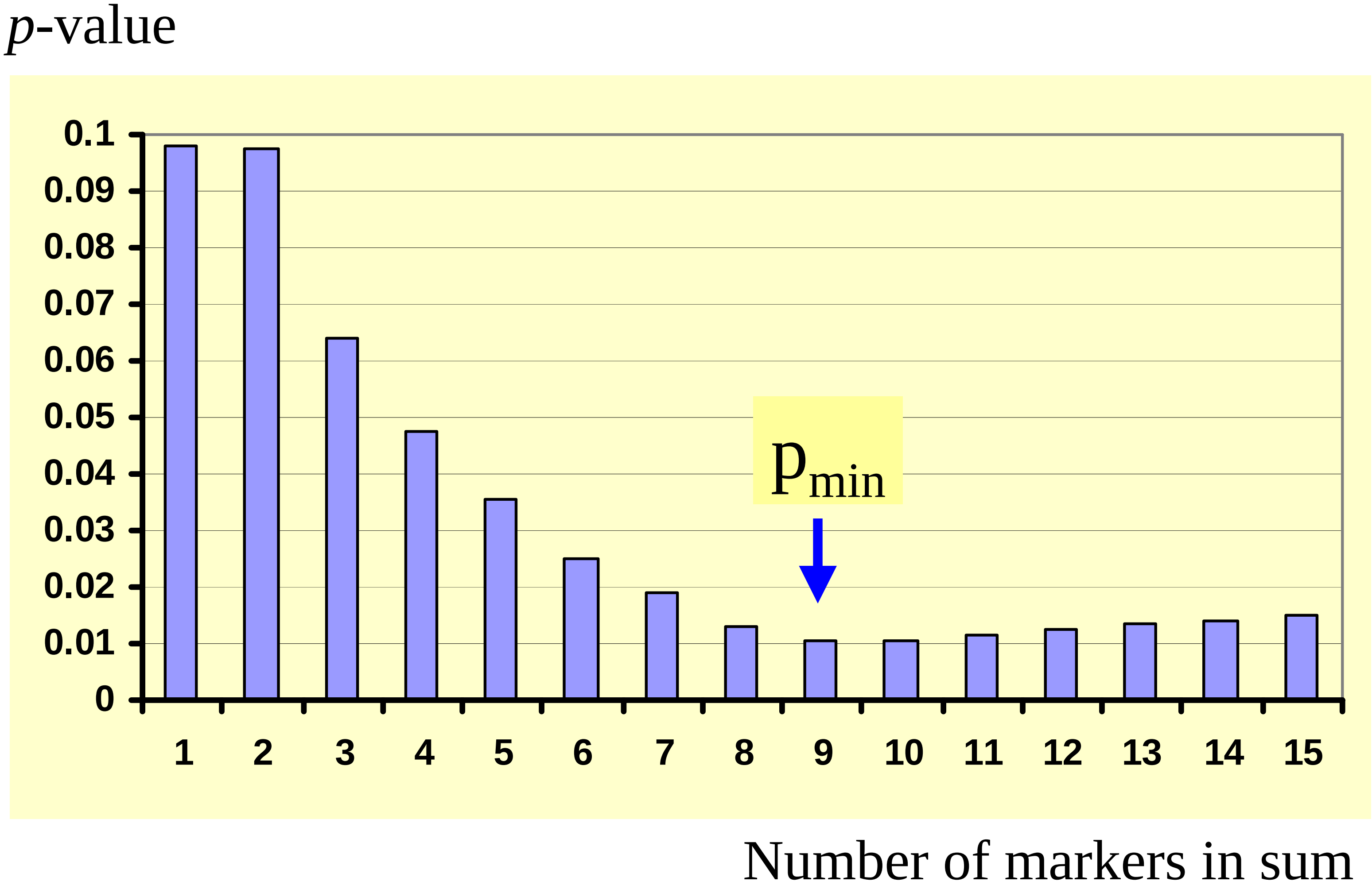
Zee *et al* (2002) *Pharmacogenomics J* **2**, 197-201

(candidate genes, not genome screen)

- **Study subjects**: 779 heart disease patients, angioplasty, 342 with restenosis (cases), 437 without restenosis (controls)
- **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

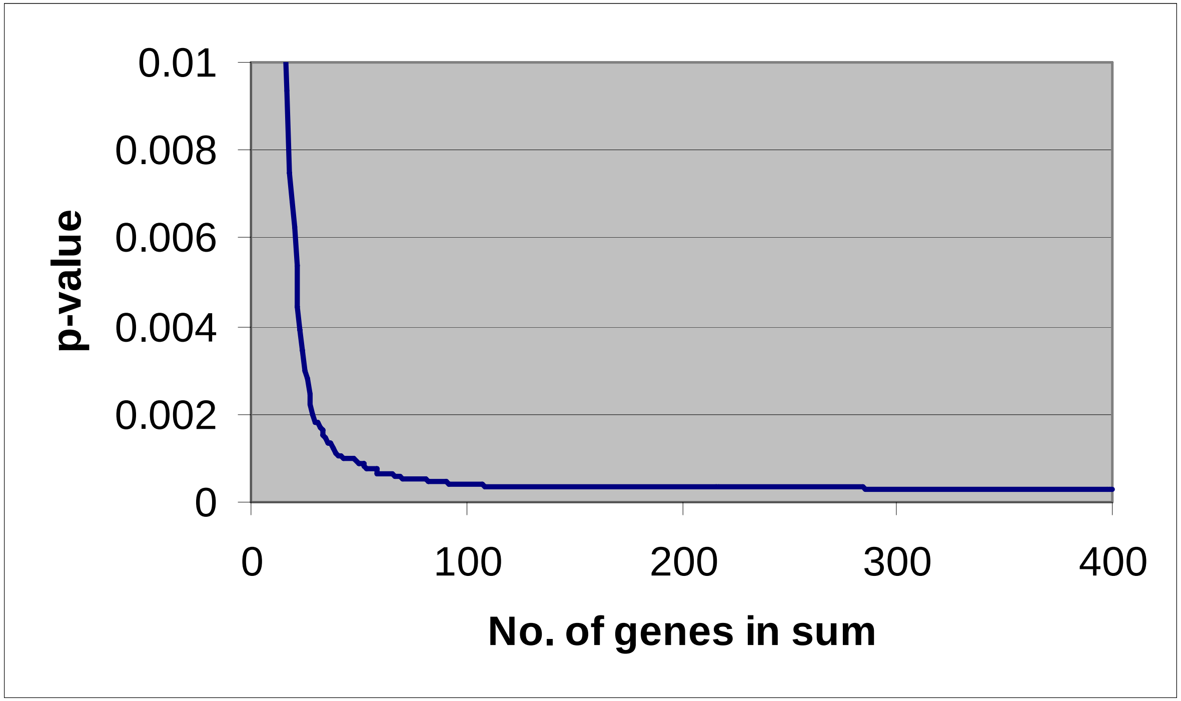
Application to Heart Disease Data

(overall $p = 0.04$)



Example: Microarray Data

- 14 pairs of observations, stomach cancer and its metastasis. Expression levels measured for each of 28,500 genes.



- Analysis: Paired |t|-test for each gene, largest value trimmed.
- Best gene: p = 0.1654. Best 400 genes: p = 0.0008 (overall).
- $p_{\min}$ at ~380 genes

Step 2: Adverse Drug Reactions

Acuña *et al* (2002) *Pharmacogenomics J* **2**, 327-334

- An otherwise good prescription drug (Tasmar®) causes liver toxicity (LT) in a small proportion of users.
- Hypothesis: People with LT are genetically susceptible.
- Biologists identify 14 candidate genes and genotype SNPs in 135 cases (LT) and 274 control individuals (no TS).

Statistical Analysis of Selected Variables

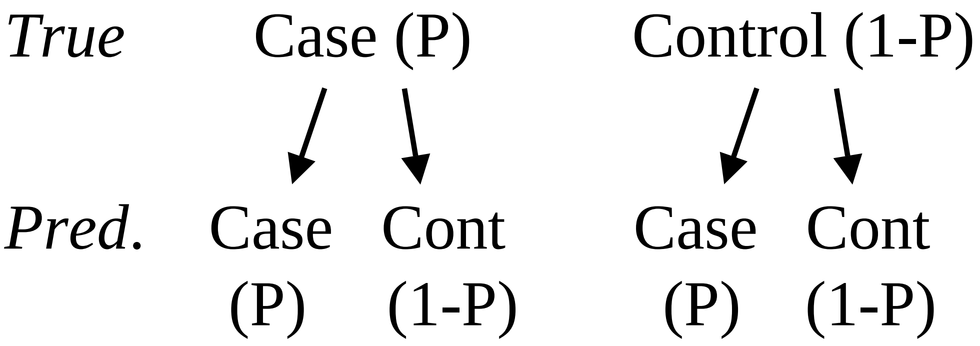
- Logistic regression analysis on the two principal components and non-genetic variables. Genetic effect only in male patients but females also show LT.
- P = probability of being a case

$$\log\left(\frac{P}{1-P}\right)=c_0+c_1x_1+c_2x_2+\ldots$$

Prediction Efficiency

- Prediction on the basis of estimated logistic regression model is correct for 65% of individuals.

- Random prediction: Say “case” with probability P = 135/(135 + 274) = 0.33.



True Case (P) Control (1-P)

Pred. Case (P) Cont (1-P) Case (P) Cont (1-P)

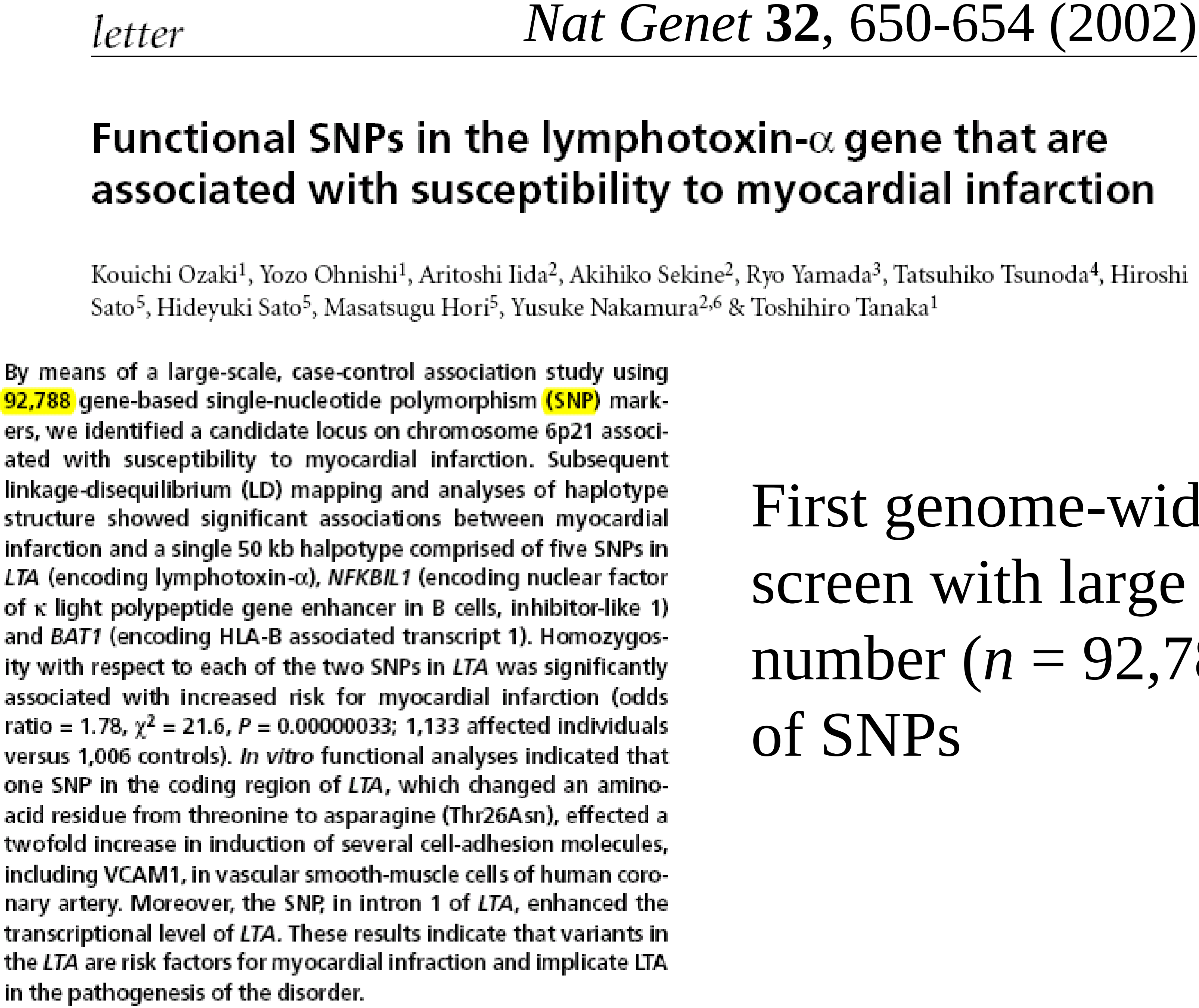
- Random prediction is 56% accurate.

Other Analysis Approaches

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, Vanderbuilt U)
- 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

Nat Genet **32**, 650-654 (2002)



letter

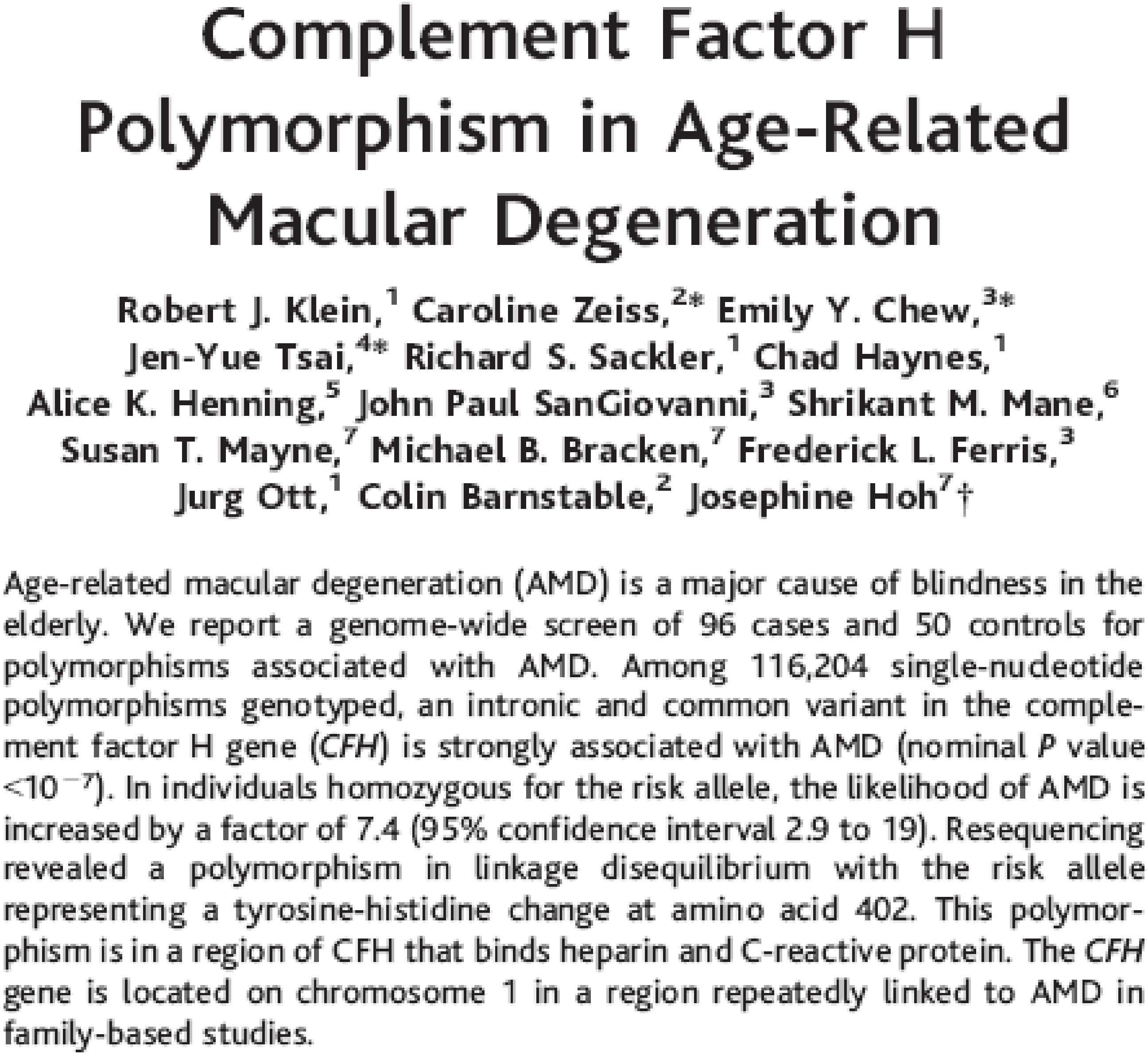
Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction

Kouichi Ozaki[1], Yozo Ohnishi[1], Aritoshi Iida[2], Akihiko Sekine[2], Ryo Yamada[3], Tatsuhiko Tsunoda[4], Hiroshi Sato[5], Hideyuki Sato[5], Masatsugu Hori[5], Yusuke Nakamura[2,6] & Toshihiro Tanaka[1]

By means of a large-scale, case-control association study using 92,788 gene-based single-nucleotide polymorphism (SNP) markers, we identified a candidate locus on chromosome 6p21 associated with susceptibility to myocardial infarction. Subsequent linkage-disequilibrium (LD) mapping and analyses of haplotype structure showed significant associations between myocardial infarction and a single 50 kb halpotype comprised of five SNPs in *LTA* (encoding lymphotoxin-α), *NFKBIL1* (encoding nuclear factor of κ light polypeptide gene enhancer in B cells, inhibitor-like 1) and *BAT1* (encoding HLA-B associated transcript 1). Homozygosity with respect to each of the two SNPs in *LTA* was significantly associated with increased risk for myocardial infarction (odds ratio = 1.78, $\chi^2 = 21.6$, $P = 0.00000033$; 1,133 affected individuals versus 1,006 controls). *In vitro* functional analyses indicated that one SNP in the coding region of *LTA*, which changed an amino-acid residue from threonine to asparagine (Thr26Asn), effected a twofold increase in induction of several cell-adhesion molecules, including VCAM1, in vascular smooth-muscle cells of human coronary artery. Moreover, the SNP, in intron 1 of *LTA*, enhanced the transcriptional level of *LTA*. These results indicate that variants in the *LTA* are risk factors for myocardial infraction and implicate LTA in the pathogenesis of the disorder.

First genome-wide screen with large number (n = 92,788) of SNPs

Science **308**, 358-389   15 Apr 2005



Complement Factor H Polymorphism in Age-Related Macular Degeneration

Robert J. Klein,[1] Caroline Zeiss,[2*] Emily Y. Chew,[3*] Jen-Yue Tsai,[4*] Richard S. Sackler,[1] Chad Haynes,[1] Alice K. Henning,[5] John Paul SanGiovanni,[3] Shrikant M. Mane,[6] Susan T. Mayne,[7] Michael B. Bracken,[7] Frederick L. Ferris,[3] Jurg Ott,[1] Colin Barnstable,[2] Josephine Hoh[7]†

Age-related macular degeneration (AMD) is a major cause of blindness in the elderly. We report a genome-wide screen of 96 cases and 50 controls for polymorphisms associated with AMD. Among 116,204 single-nudeotide polymorphisms genotyped, an intronic and common variant in the complement factor H gene (*CFH*) is strongly associated with AMD (nominal *P* value $<10^{-7}$). In individuals homozygous for the risk allele, the likelihood of AMD is increased by a factor of 7.4 (95% confidence interval 2.9 to 19). Resequencing revealed a polymorphism in linkage disequilibrium with the risk allele representing a tyrosine-histidine change at amino acid 402. This polymorphism is in a region of CFH that binds heparin and C-reactive protein. The *CFH* gene is located on chromosome 1 in a region repeatedly linked to AMD in family-based studies.

New genome-wide screen, n = 116,204 SNPs, strong association of a variant and AMD.

Permutation Tests

- Want to find 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)

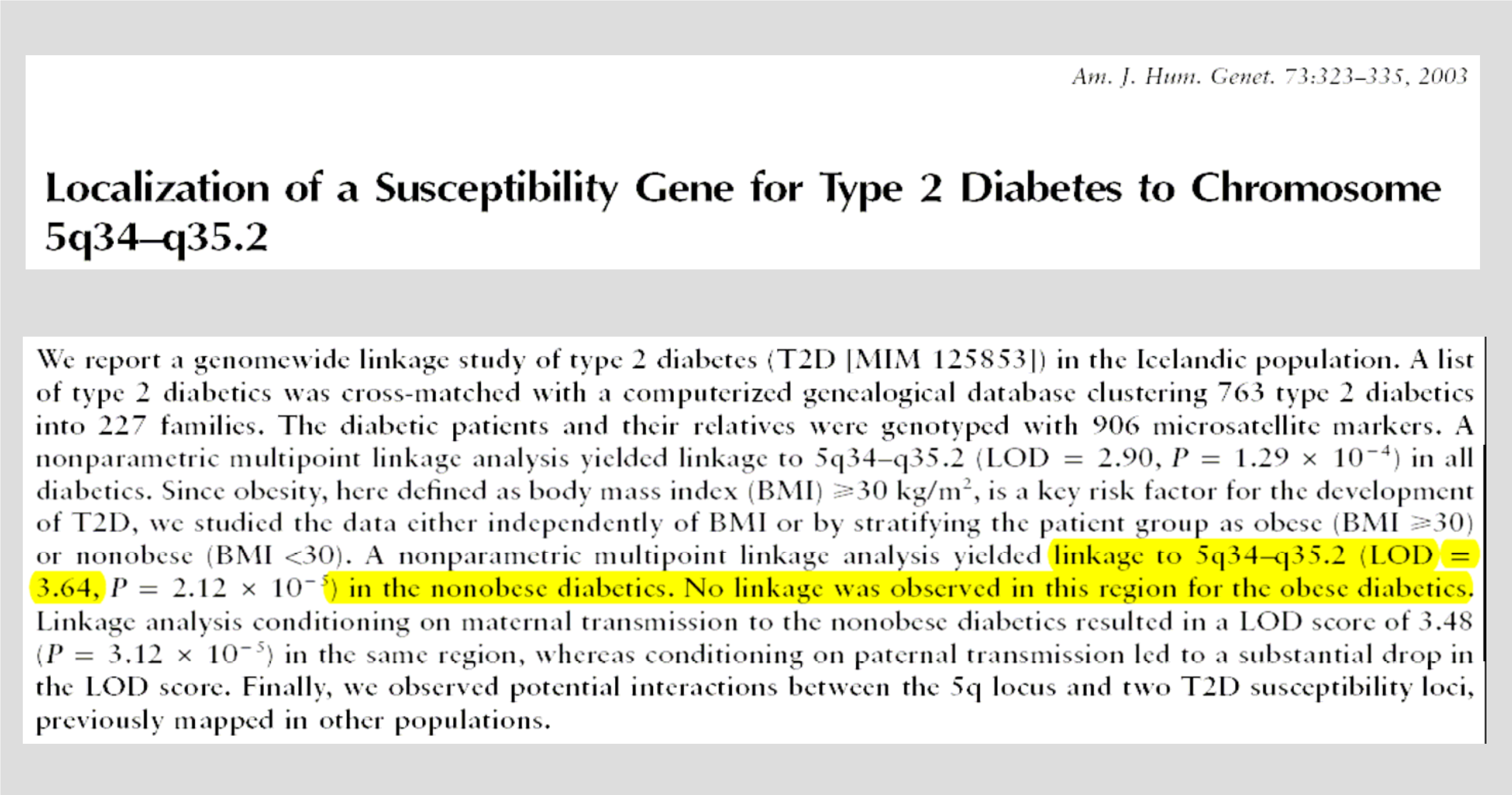
Population Substructure (Heterogeneity)

- *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 χ^2 from the χ^2 in your study.

Data Subdivision

- Pritchard method, based on unassociated SNPs
- Identify groups of individuals with similar non-genetic risk factors, each group → *p*-value.
- Sparse tables in case-control studies:
  - Exact methods (*StatXact* program) rather than table values of χ^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.

Example of successful stratification



Am. J. Hum. Genet. 73:323–335, 2003

Localization of a Susceptibility Gene for Type 2 Diabetes to Chromosome 5q34–q35.2

We report a genomewide linkage study of type 2 diabetes (T2D [MIM 125853]) in the Icelandic population. A list of type 2 diabetics was cross-matched with a computerized genealogical database clustering 763 type 2 diabetics into 227 families. The diabetic patients and their relatives were genotyped with 906 microsatellite markers. A nonparametric multipoint linkage analysis yielded linkage to 5q34–q35.2 (LOD = 2.90, $P = 1.29 \times 10^{-4}$) in all diabetics. Since obesity, here defined as body mass index (BMI) $\geqslant 30$ kg/m^2, is a key risk factor for the development of T2D, we studied the data either independently of BMI or by stratifying the patient group as obese (BMI $\geqslant 30$) or nonobese (BMI <30). A nonparametric multipoint linkage analysis yielded linkage to 5q34–q35.2 (LOD = 3.64, $P = 2.12 \times 10^{-5}$) in the nonobese diabetics. No linkage was observed in this region for the obese diabetics. Linkage analysis conditioning on maternal transmission to the nonobese diabetics resulted in a LOD score of 3.48 ($P = 3.12 \times 10^{-5}$) in the same region, whereas conditioning on paternal transmission led to a substantial drop in the LOD score. Finally, we observed potential interactions between the 5q locus and two T2D susceptibility loci, previously mapped in other populations.

Are results “significant”?

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

- *n* SNPs, each tested for association at significance level α = 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
p19	***RARb***	***0.0001***	1	0.0008
p16	***TIMP3***	***0.0002***	2	0.0015
DAPK	***p21***	***0.0012***	3	0.0023
MGMT	***RARb***	***0.0016***	4	0.0030
RARb	***TIMP3***	***0.0023***	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, λ_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

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, ϕ → 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		***Power***		
α	***Prior***	***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

Purely Epistatic Disease Model

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

L.1	*L.3 = 1/1*			*L.3 = 1/2*			*L.3 = 2/2*		
↓L.2	*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

Assume all allele frequencies = 0.50.

Heritability = 55%, prevalence = 6.25%.

Expected Genotype Patterns

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

Inference

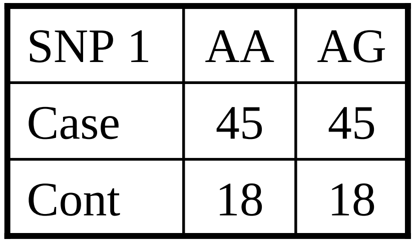
- Given 3 disease SNPs: $\chi^2 = 166.7$ (26 df), $p = 1.76 \times 10^{-22}$.
- 50,000 SNPs → 2.1×10^{13} subsets of size 3.
- Bonferroni-corrected $p = 3.6 \times 10^{-9}$.
- Alternative approach: Test all possible pairs of loci for interaction effects, different in case and control individuals [Hoh & Ott (2003) *Nat Rev Genet* **4**, 701-709].

Pairwise SNP interaction effects

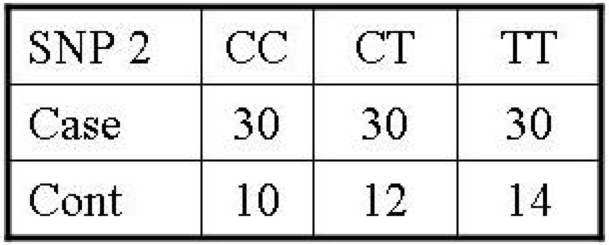
SNP1	AA			AG		
SNP2	CC	CT	TT	CC	CT	TT
Cases	10	15	20	20	15	10
Controls	5	6	7	5	6	7

Assumed data.
Analyze in two different ways.

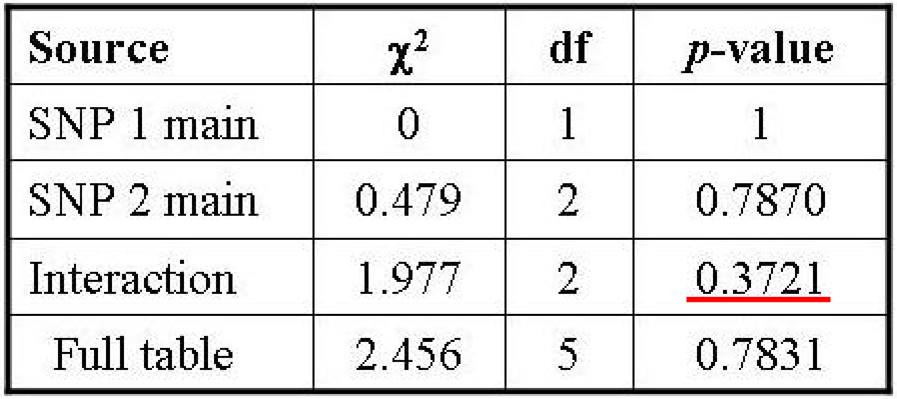
(1) Partitioning LR chi-square



SNP 1	AA	AG
Case	45	45
Cont	18	18

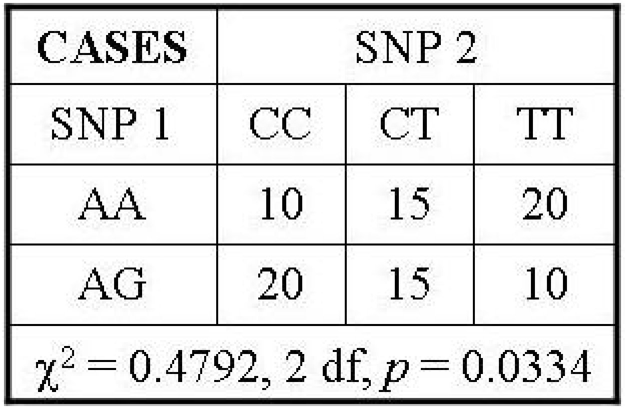


SNP 2	CC	CT	TT
Case	30	30	30
Cont	10	12	14

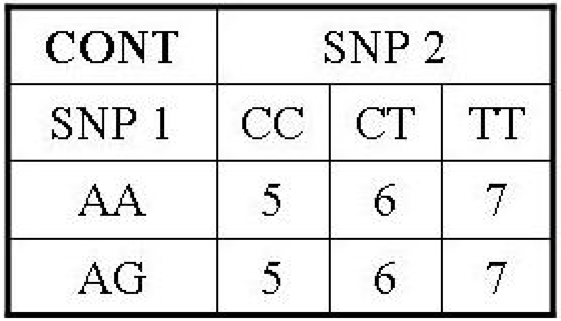


Source	χ^2	**df**	***p*-value**
SNP 1 main	0	1	1
SNP 2 main	0.479	2	0.7870
Interaction	1.977	2	0.3721
Full table	2.456	5	0.7831

(2) Cases vs. controls



CASES	SNP 2		
SNP 1	CC	CT	TT
AA	10	15	20
AG	20	15	10
$\chi^2 = 0.4792$, 2 df, $p = 0.0334$			



CONT	SNP 2		
SNP 1	CC	CT	TT
AA	5	6	7
AG	5	6	7