Banff **I**nternational **R**esearch **S**tation, August 7-12, 2004

Case-Control Association Studies and the Curse of Dimensionality

Jurg Ott

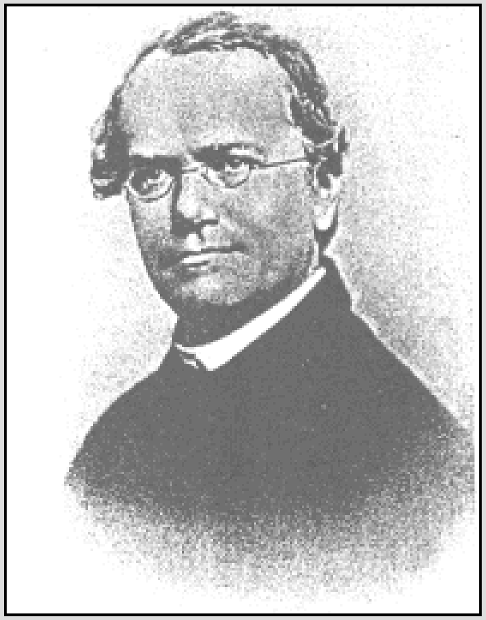
Rockefeller University, New York

Foreword

- Theorists versus experimentalists
- Theory, no mention of data
- Here, start with data, then develop suitable statistical model that can explain data.

- Early experimentalist…

Experiments with garden pea → Inheritance models



Gregor Mendel, monk in a monastery at Brünn (now Brno in Czech Republic)

Versuche

über

Pflanzen-Hybriden,

von

Gregor Mendel.

(Separatabdruck aus dem IV. Bande der Verhandlungen des naturforschenden Vereines.)

Im Verlage des Vereines.

Brünn, 1866.

Aus Georg Gastl's Buchdruckerei, Postgasse Nr. 446.

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. Statistical estimates not unique.

“Curse of Dimensionality“

Bellman R (1961) *Adaptive control processes: A guided tour*. Princeton University Press

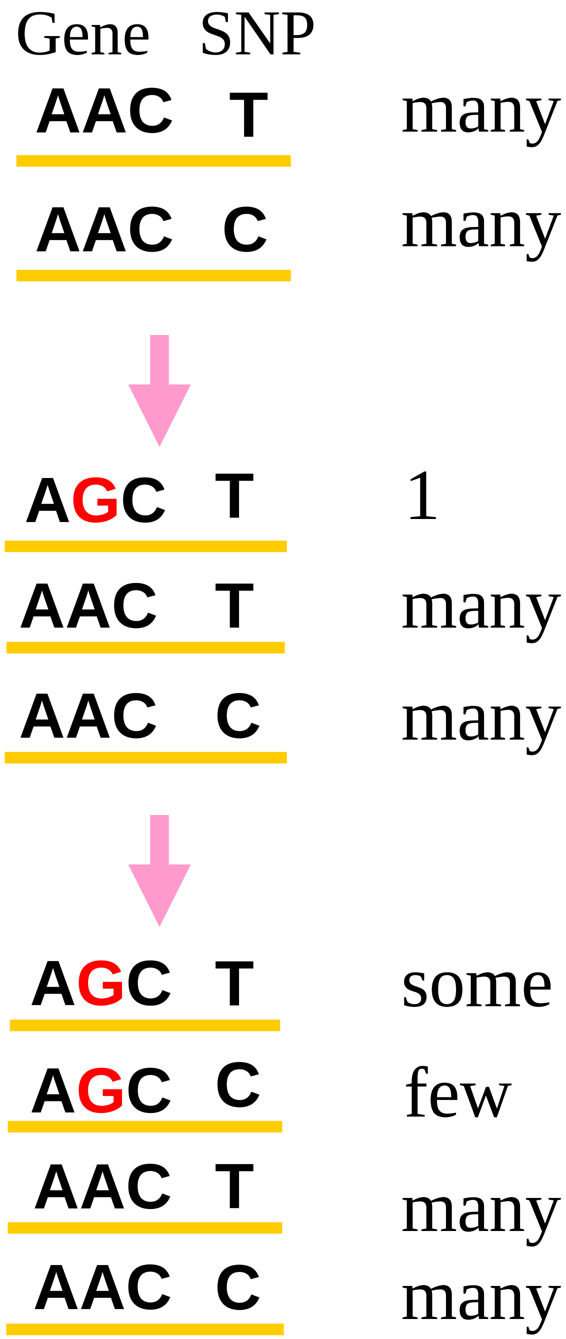
- Section 5.16 **The Curse of Dimensionality**
- “… multidimensional variational problems cannot be solved *routinely* ... . This does not mean that we cannot attack them. It merely means that we must employ some more sophisticated techniques.”

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

Linkage Disequilibrium (LD)



Gene SNP

AAC T many

AAC C many

AGC T 1

AAC T many

AAC C many

AGC T some

AGC C few

AAC T many

AAC C many

Origin in single mutation

	T	C
G	some	few
A	many	many

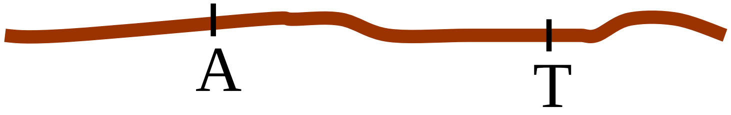
Establishing Association

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

Size of χ^2 shows significance of association.

Effects of association within short range of a locus, in contrast to linkage analysis.

Measuring the Extent of LD

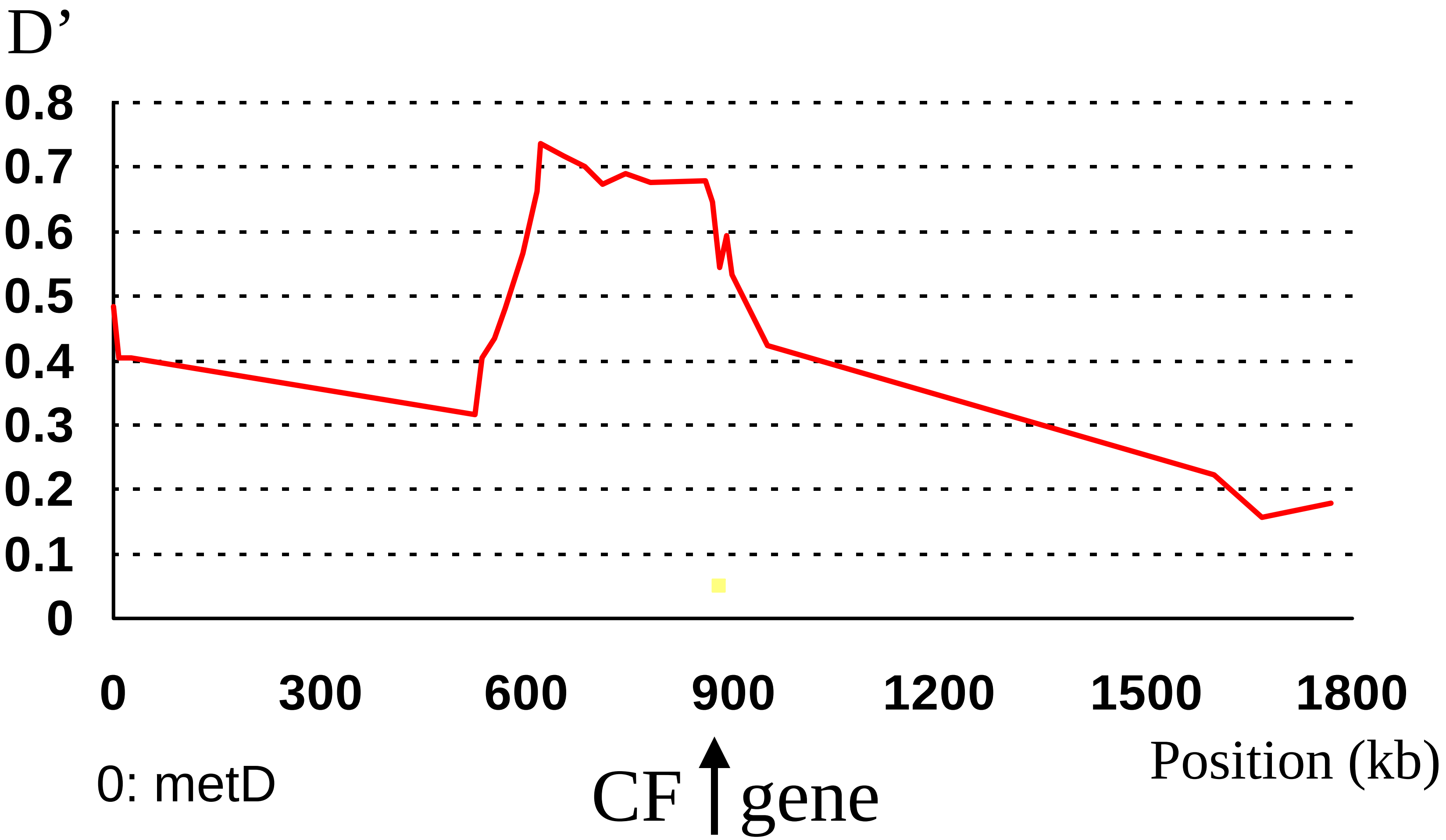


	T	T^c
A	$P(AT)$	…
A^c	…	…

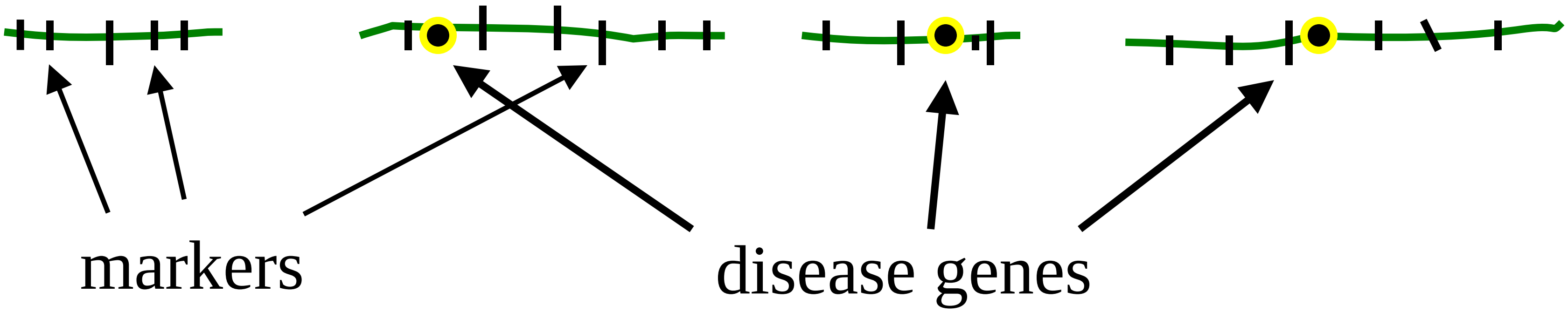
- Two alleles, one each at two loci on same chromosome (same haplotype).
- Independence: $P(AT) = P(A) \times P(T)$
- Dependence: $P(AT) = P(A) \times P(T) + D$
- D = disequilibrium parameter, min. and max. values given by allele frequencies.
- $D' = D/D_{max}$ ranges between 0 and 1.

Linkage Disequilibrium at CF locus

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



Genome Screens for Disease Loci



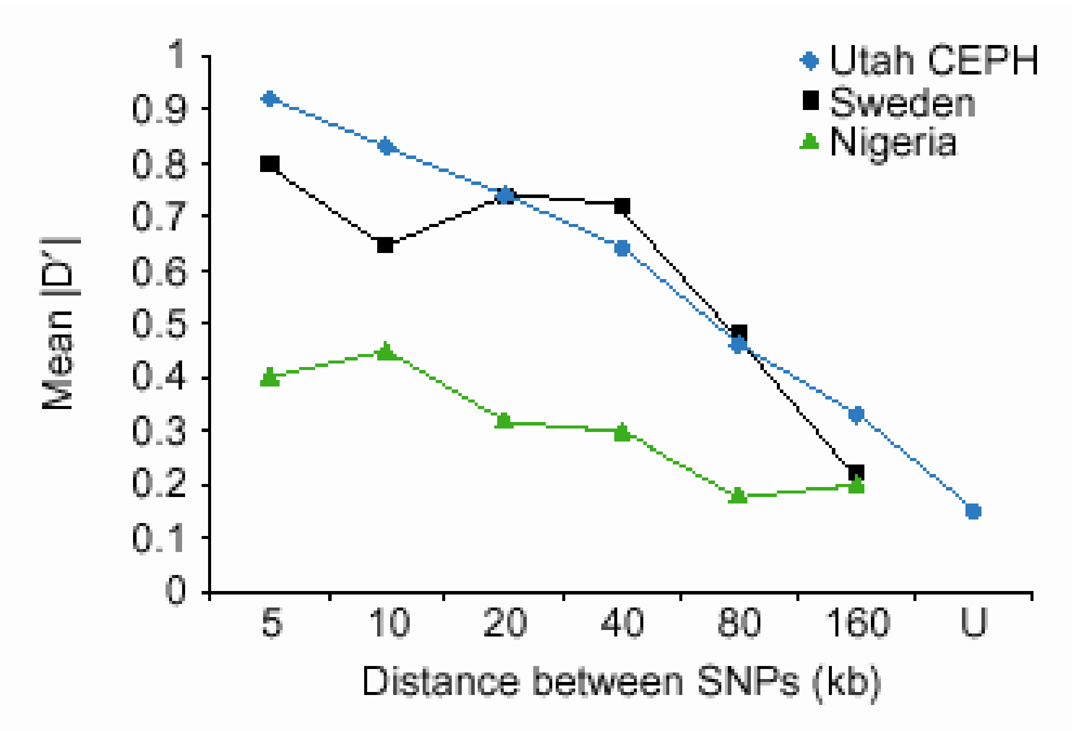
markers

disease genes

- 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.

LD Observed Between Pairs of SNPs

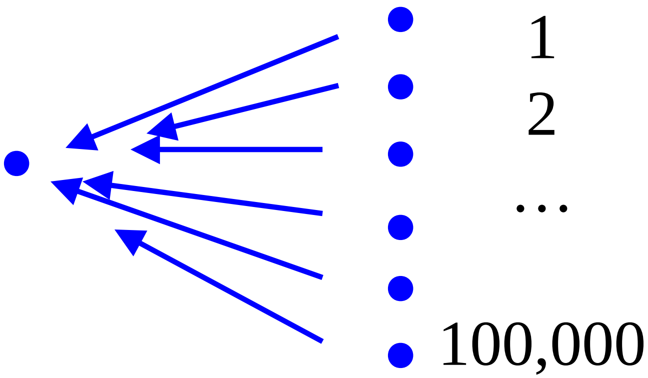
Weiss & Clark (2002) *Trends in Genetics* **18**, 19



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.

Problem

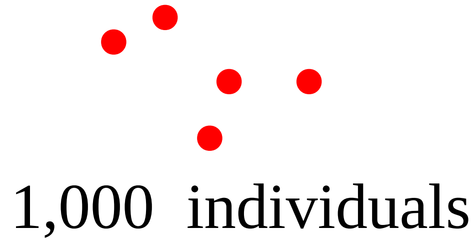


1

2

…

100,000



1,000 individuals

- Want to allow for interactions between susceptibility genes (i.e., *m* marker loci).
- Ideally, analyze all data jointly.
- “Think Big”: 3^m genotype configurations (patterns) → dimensionality enormous.

One-by-One Approach

- Need to correct for multiple testing.
- **Linkage analysis**: For dense map of markers, testing each marker at $\alpha = 0.00005$ (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.

False Discovery Rate, FDR

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

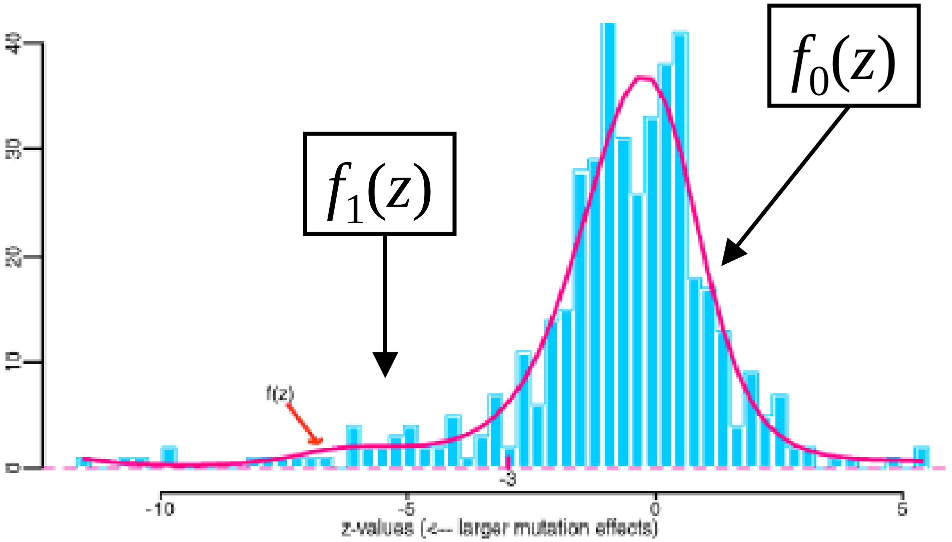
	Test not signif.	Test significant	# tests
H_0 true	U	V	m_0
H_0 false	T	S	m_1
	m - R	R	m

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

Estimating Proportion of True Positives out of All Pos. Results

Efron (2004) *JASA* **99**, 96-104

- Transform *p*-values to normal deviates, *z* (for convenience).
- Define $fdr(z) = f_0(z)/f(z)$



Multilocus 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, Vanderbuilt 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 Single-Marker Statistics

- Idea: Multiple disease genes, each associated with a marker. To capture joint effect, work with sum of association statistic for each marker involved.
- Sum should contain relevant markers: Work with markers that show strong association.

Simple Transformation

$$\begin{pmatrix} s_1 \\ s_2 \\ s_3 \\ \ldots \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} t_{(1)} \\ t_{(2)} \\ t_{(3)} \\ \ldots \end{pmatrix}$$

sums

Ordered
single-marker
statistics

Sums of marker statistics: Nested bootstrap approach

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

- Build sum of all marker statistics and determine its significance level, *p*, via bootstrap samples obtained under H_0 (no association).
- Drop marker with smallest statistic and find *p* of remaining sum.
- Those markers remaining in sum when *p* falls below 0.05
  → **pre-selected**
- Bootstrap copies of original data: Repeat above process in each copy. Marker pre-selected in more than 60% of copies
  → **selected**
  (Diaconis & Efron [1983] *Scientific American* **248**, 116-130).
- Exploratory, no overall signif. test. No model assumptions.

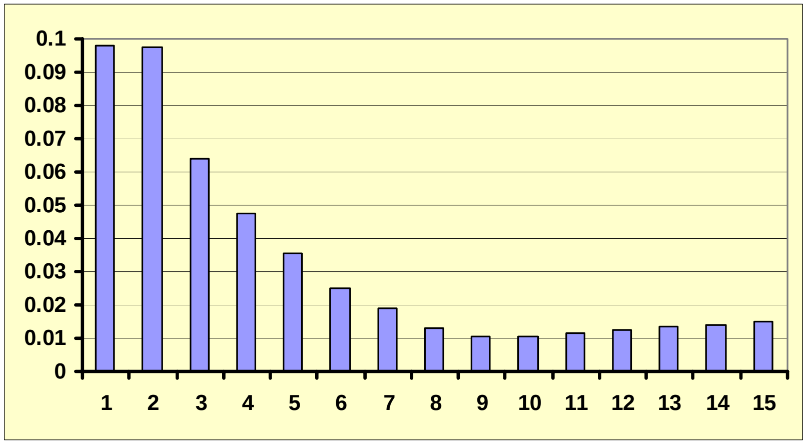
Application of Nested Bootstrap

779 heart disease patients had undergone balloon angioplasty, 342 of whom subsequently experienced coronary artery restenosis (cases), whereas the remainder did not (controls). Genotypes for 94 SNPs representing 62 candidate genes determined. Nested bootstrap procedure selected 11 out of the 94 SNP markers in the following 10 genes: *TNFR1*, *IL4R*α, *TP53*, *CD14*, *APOA*, *CETP*, *TNF*β, *CBS*, *NOS* and *MDM2*.

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 → select
- Smallest *p* = single experiment-wise statistic → overall significance level

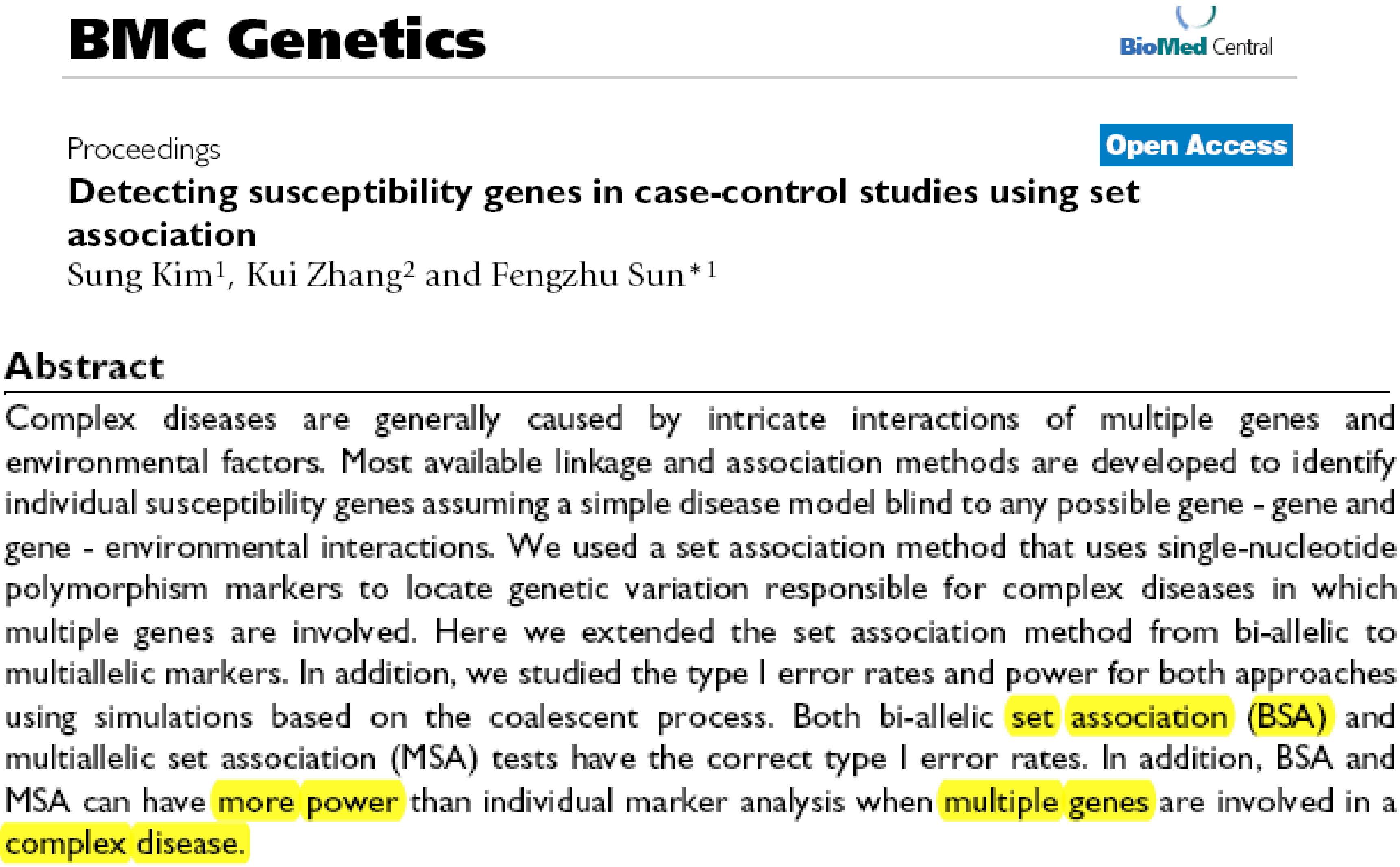


Application: Restenosis Data

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 9 SNPs (7 are the same as selected with nested bootstrap).
- Significance level associated with smallest p is 0.04.

Power



BMC Genetics

BioMed Central

Proceedings

Open Access

Detecting susceptibility genes in case-control studies using set association

Sung Kim[1], Kui Zhang[2] and Fengzhu Sun*[1]

Abstract

Complex diseases are generally caused by intricate interactions of multiple genes and environmental factors. Most available linkage and association methods are developed to identify individual susceptibility genes assuming a simple disease model blind to any possible gene - gene and gene - environmental interactions. We used a set association method that uses single-nucleotide polymorphism markers to locate genetic variation responsible for complex diseases in which multiple genes are involved. Here we extended the set association method from bi-allelic to multiallelic markers. In addition, we studied the type I error rates and power for both approaches using simulations based on the coalescent process. Both bi-allelic set association (BSA) and multiallelic set association (MSA) tests have the correct type I error rates. In addition, BSA and MSA can have more power than individual marker analysis when multiple genes are involved in a complex disease.

Association Rules

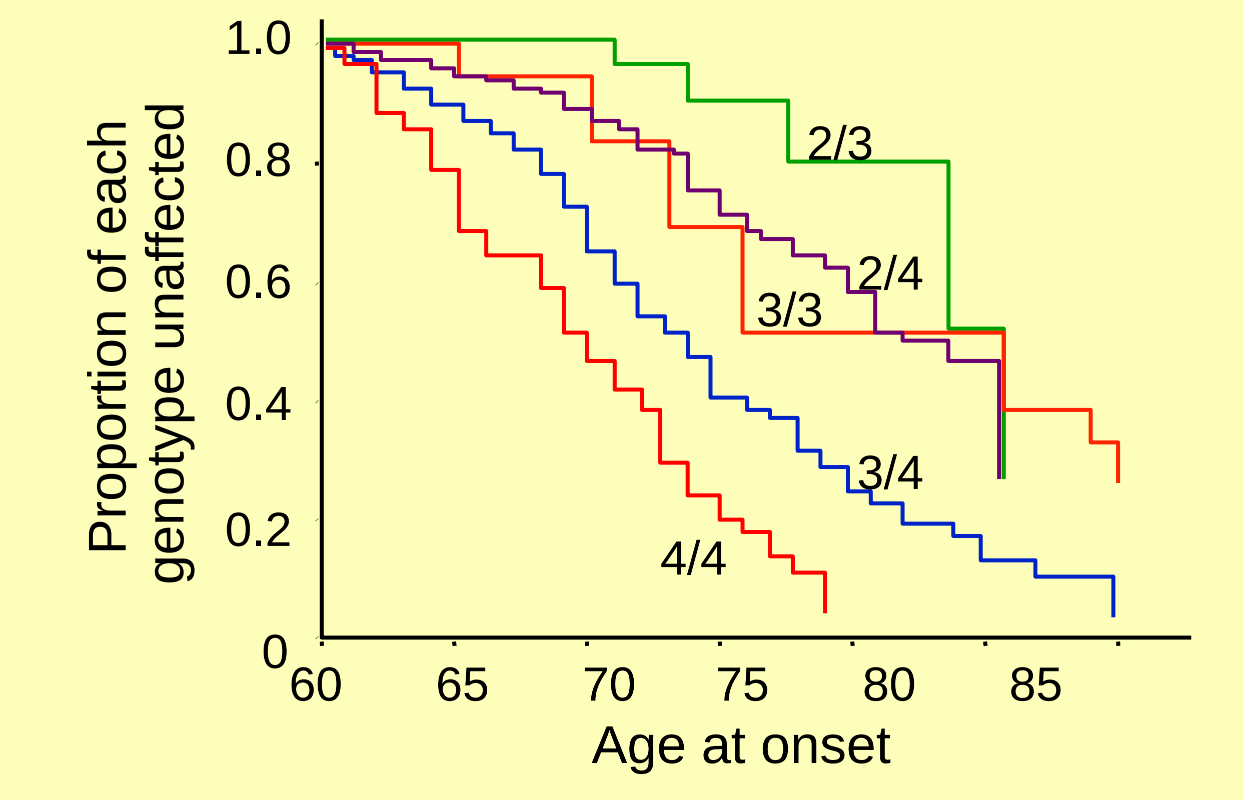
http://fuzzy.cs.uni-magdeburg.de/~borgelt/software.html

- 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.
- Very fast. Few applications so far to genetic data (Toivonen et al [2000] *Am J Hum Genet* **67**, 133).

Apriori Algorithm

- Association rule R = expression X ⇒ Y, with X and Y being sets of items, itemsets. Example: R = “Wine and Bread ⇒ Cheese”.
- For itemset S, support (S) = #S/n × 100%
- For R = “A and B ⇒ C”, confidence (R) = support (“A and B ⇒ C”)/support (A and B) × 100%
- *Apriori* algorithm finds association rules with minimum support and confidence.
- Genetics: Target item may be disease status, other items = genotypes.

Strong Effects of Single Genes



Alzheimer disease: Strong effect of APOE genotype on risk to disease.

“Fish Consumption and the Risk of Alzheimer Disease“

Friedland (2003) *Arch Neurol* **60**, 923 (editorial)

- Subjects who eat fish at least once a week have a 60% lower risk for developing AD than those who consume fish less frequently (Morris et al. [2003] *Arch Neurol* **60**, 194).
- Previous studies point in same direction.

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.
- Real-life examples of epistatic traits?

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 $p = 3.6 \times 10^{-9}$.
- More manageable 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).