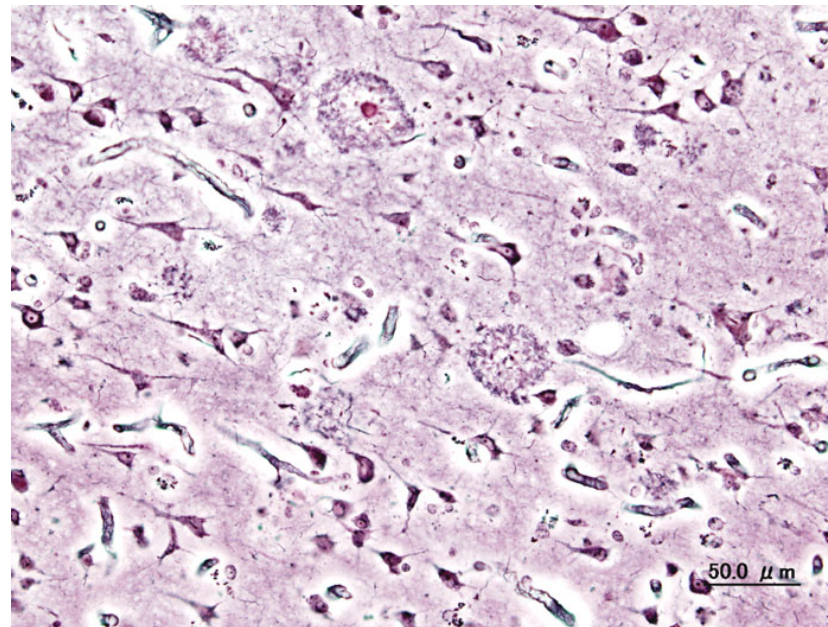


Markov Bases in Statistical Genetics

Caroline Uhler* and Anna-Sapfo Malaspinas**

* Department of Statistics ** Department of Integrative Biology

University of California, Berkeley



Overview

- **Player 1: Biology**
 - Disease association studies
- **Player 2: Statistics**
 - Interaction and how this can be detected
- **Player 3: Algebra**
 - Markov bases and Lawrence liftings

Disease association studies

- Single gene disorders
 - E.g. Cystic fibrosis, haemophilia A and B, Huntington's disease.
- Multifactorial disorders
 - Caused by the interaction of multiple genes and the environment.
 - E.g. Alzheimer's disease, diabetes, cancer, multiple sclerosis.

Data

- SNPs: Genotype data encoded by 0, 1 and 2
- Disease status encoded by 0, 1

For this talk: 2 markers

$\Rightarrow 3 \times 3 \times 2$ contingency table

Interaction between markers

- In a biological context, interaction means:

Epistasis

- In a mathematical context, interaction means:

Correlation

$$\mathbb{P}(\text{marker 1} = i, \text{marker 2} = j) \neq \mathbb{P}(\text{marker 1} = i)\mathbb{P}(\text{marker 2} = j)$$

Correlation can be caused by

- Fitness differences
- Linkage disequilibrium
- Non-random sampling

χ^2 -goodness-of-fit test

The test statistic is given by

$$T_n = \sum_{i,j} \frac{(n_{ij} - E_{ij})^2}{E_{ij}},$$

where E_{ij} represents the expected number of observations in class (i, j) under H_0 . So under independence:

$$E_{ij} = \frac{n_{i.}n_{.j}}{n_{..}}$$

Asymptotically,

$$T \sim \chi^2_{(s-1)}.$$

Fisher's exact test

$$\begin{pmatrix} 5 & 0 \\ 1 & 4 \end{pmatrix} \quad p(5, 0, 1, 4 \mid \text{margins}) = 0.0238$$

$$\begin{pmatrix} 4 & 1 \\ 2 & 3 \end{pmatrix} \quad p(4, 1, 2, 3 \mid \text{margins}) = 0.2381$$

$$\begin{pmatrix} 3 & 2 \\ 3 & 2 \end{pmatrix} \quad p(3, 2, 3, 2 \mid \text{margins}) = 0.4762$$

$$\begin{pmatrix} 2 & 3 \\ 4 & 1 \end{pmatrix} \quad p(2, 3, 4, 1 \mid \text{margins}) = 0.2381$$

$$\begin{pmatrix} 1 & 4 \\ 5 & 0 \end{pmatrix} \quad p(1, 4, 5, 0 \mid \text{margins}) = 0.0238$$

$$\Rightarrow \text{p-value} = 0.0476 < 0.05$$

Interaction models with two SNPs

Model	Minimal sufficient statistics	Expected counts
(X, Y, D)	$(n_{i..}), (n_{.j.}), (n_{..k})$	$\hat{n}_{ijk} = \frac{n_{i..}n_{.j.}n_{..k}}{(n_{...})^2}$
(XY, D)	$(n_{ij.}), (n_{..k})$	$\hat{n}_{ijk} = \frac{n_{ij.}n_{..k}}{(n_{...})}$
(XD, Y)	$(n_{i.k}), (n_{.j.})$	$\hat{n}_{ijk} = \frac{n_{i.k}n_{.j.}}{(n_{...})}$
(X, YD)	$(n_{i..}), (n_{.jk})$	$\hat{n}_{ijk} = \frac{n_{.jk}n_{i..}}{(n_{...})}$
(XY, YD)	$(n_{ij.}), (n_{.jk})$	$\hat{n}_{ijk} = \frac{n_{ij.}n_{.jk}}{(n_{.j.})}$
(XY, XD)	$(n_{ij.}), (n_{i.k})$	$\hat{n}_{ijk} = \frac{n_{ij.}n_{i.k}}{(n_{i..})}$
(XD, YD)	$(n_{i.k}), (n_{.jk})$	$\hat{n}_{ijk} = \frac{n_{i.k}n_{.jk}}{(n_{..k})}$
(XY, XD, YD)	$(n_{ij.}), (n_{i.k}), (n_{.jk})$	IPF

Markov basis

Definition: A Markov basis for \mathcal{F} is a finite family of tables $\mathcal{T} = (T_1, \dots, T_L)$ such that

- (i) T_1, \dots, T_L have margins equal to 0.
- (ii) For any $T, T' \in \mathcal{F}$ there are $(\varepsilon_1, T_{i_1}), \dots, (\varepsilon_A, T_{i_A})$ with $\varepsilon \in \{\pm 1\}$,

$$T' = T + \sum_{j=1}^A \varepsilon_j T_{l_j} \quad \text{and} \quad T + \sum_{j=1}^a \varepsilon_j T_{l_j} \geq 0 \quad \text{for } 1 \leq a \leq A.$$

Algorithm

- i) Select two SNPs $\Rightarrow 3 \times 3 \times 2$ contingency table.
- ii) Choose a model and determine its minimal sufficient statistics.
- iii) Compute the corresponding Markov basis (e.g. by using `4ti2`).
- iv) Run MCMC starting in the given contingency table and using the Markov basis from iii).
- v) Approximate the exact p-value of the given contingency table by the posterior distribution of the χ^2 -statistic.

Simulations: Hap-Sample

- Simulates SNPs for case control studies.
- Uses Hap-Map and assumes recombination and random mating to get enough cases and controls.
- Predicts genotypes given the disease status.

Input to Hap-Sample

As a first step we restricted the data set to:

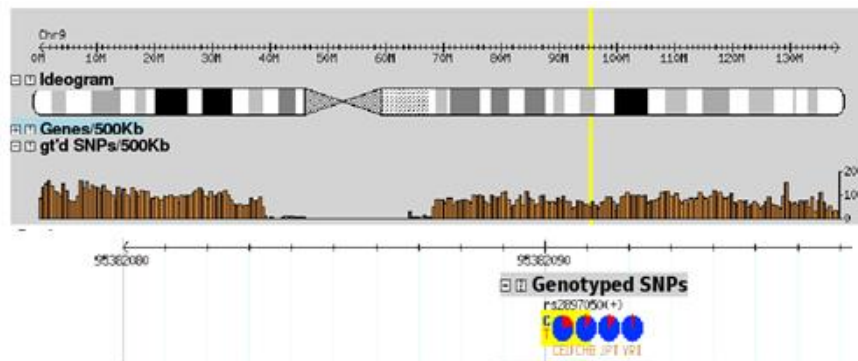
- CEU population
- 2 causative loci, unlinked
- 3 interaction models: control, additive, epistatic
- 400 cases, 400 controls

Choice of SNPs

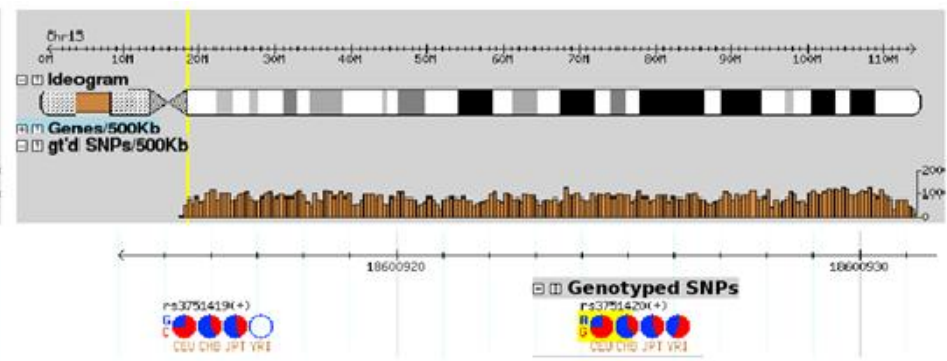
Two causative loci:

- on chromosome 9 and chromosome 13
- 9935 SNPs total

rs2897050 frequency=0.217 (CEU)



rs3751420, frequency=0.275 (CEU)



Models

- Control model:

$\frac{\mathbb{P}(D=1 \text{genotype})}{\mathbb{P}(D=0 \text{genotype})}$	0	1	2
0	ϵ	ϵ	ϵ
1	ϵ	ϵ	ϵ
2	ϵ	ϵ	ϵ

- Additive model:

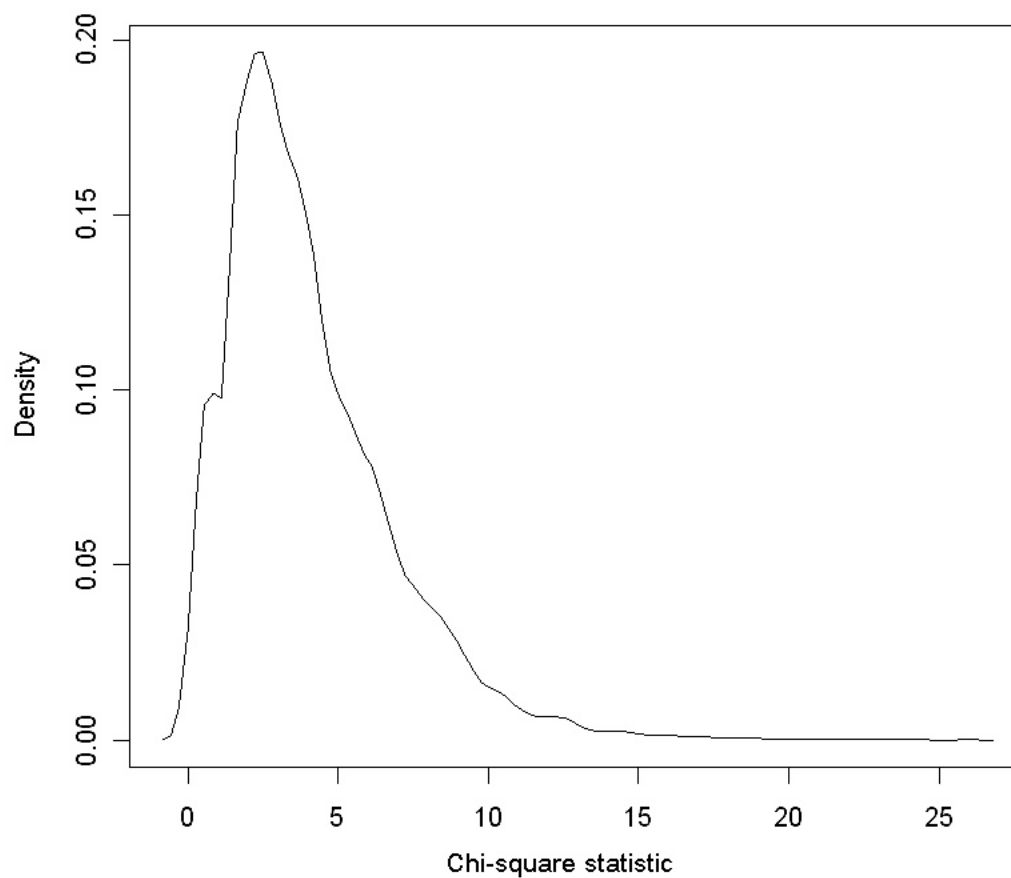
$\frac{\mathbb{P}(D=1 \text{genotype})}{\mathbb{P}(D=0 \text{genotype})}$	0	1	2
0	ϵ	$\epsilon(1 + \theta_2)$	$\epsilon(1 + \theta_2)^2$
1	$\epsilon(1 + \theta_1)$	$\epsilon(1 + \theta_1)(1 + \theta_2)$	$\epsilon(1 + \theta_1)(1 + \theta_2)^2$
2	$\epsilon(1 + \theta_1)^2$	$\epsilon(1 + \theta_1)^2(1 + \theta_2)$	$\epsilon(1 + \theta_1)^2(1 + \theta_2)^2$

- Epistatic model:

$\frac{\mathbb{P}(D=1 \text{genotype})}{\mathbb{P}(D=0 \text{genotype})}$	0	1	2
0	ϵ	$\epsilon(1 + \theta_2)$	$\epsilon(1 + \theta_2)^2$
1	$\epsilon(1 + \theta_1)$	$\epsilon(1 + \theta_1)(1 + \theta_2)(1 + \theta_3)$	$\epsilon(1 + \theta_1)(1 + \theta_2)^2(1 + \theta_3)^2$
2	$\epsilon(1 + \theta_1)^2$	$\epsilon(1 + \theta_1)^2(1 + \theta_2)(1 + \theta_3)^2$	$\epsilon(1 + \theta_1)^2(1 + \theta_2)^2(1 + \theta_3)^4$

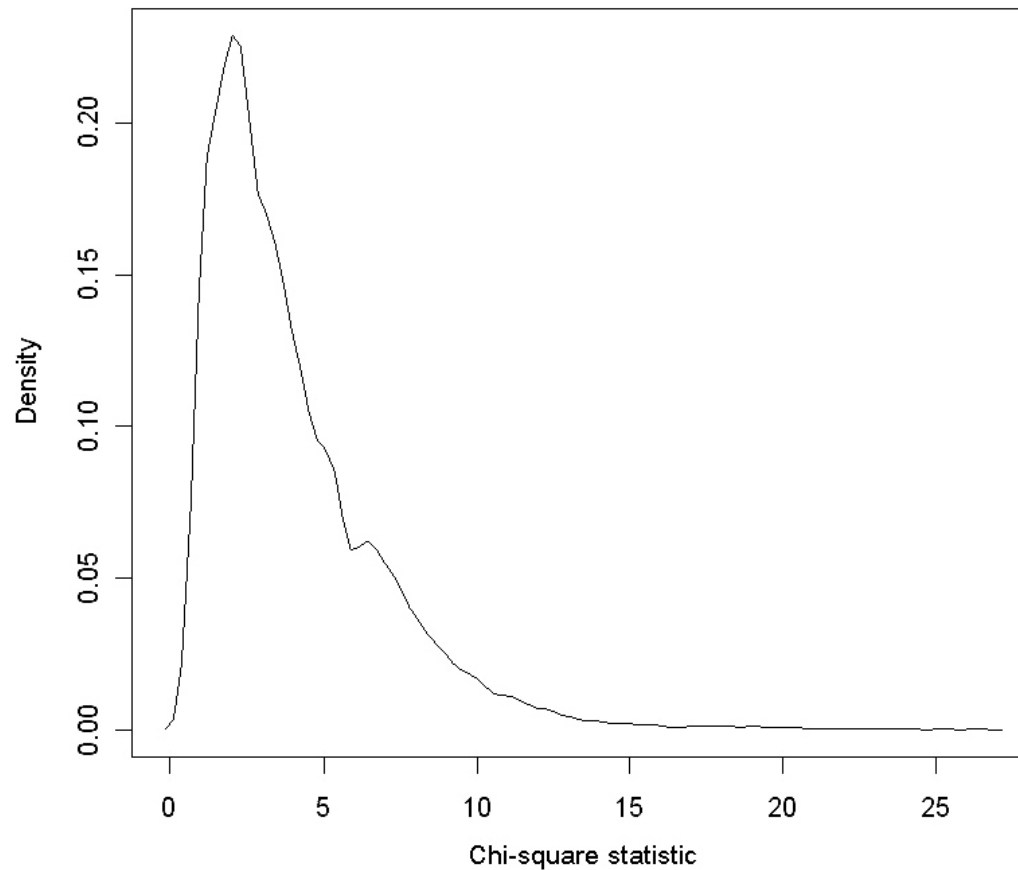
Results: Additive model

Data matrix: $\begin{pmatrix} 127 & 72 & 14 \\ 99 & 47 & 11 \\ 17 & 12 & 1 \end{pmatrix}$ $\begin{pmatrix} 0 & 102 & 19 \\ 132 & 82 & 18 \\ 26 & 18 & 3 \end{pmatrix}$ χ^2 -statistic = 73.06

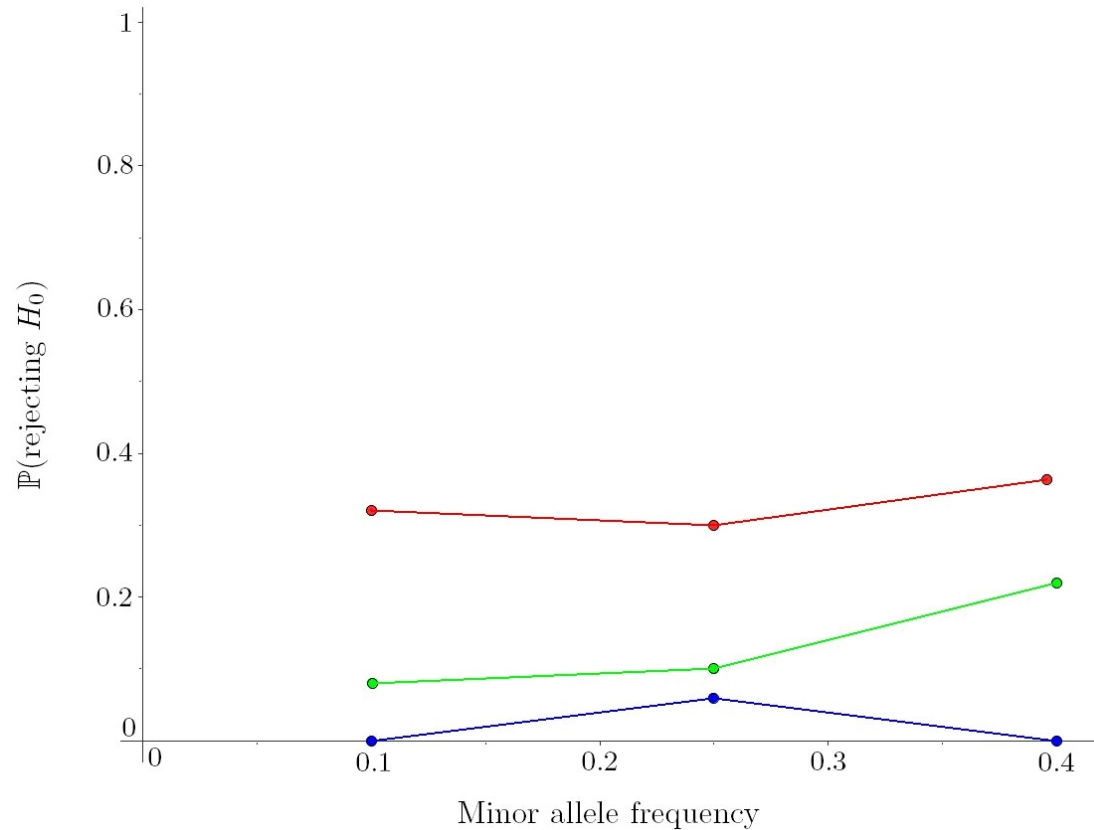


Results: Epistatic model

Data matrix: $\begin{pmatrix} 136 & 56 & 10 \\ 105 & 56 & 9 \\ 18 & 10 & 0 \end{pmatrix} \begin{pmatrix} 0 & 71 & 12 \\ 140 & 101 & 13 \\ 42 & 18 & 3 \end{pmatrix} \quad \chi^2\text{-statistic} = 72.67$



Results: Power analysis



Notation: H_0 : No 3-way interaction
blue: Control with $\epsilon = 1$
green: Additive model with $\epsilon = 0.05, \theta_1 = 1, \theta_2 = 2.5$
red: Epistatic model with $\epsilon = 0.05, \theta_1 = 1, \theta_2 = 2.5, \theta_3 = 2$

Future work

- Incorporate a more sophisticated single-locus search.
- Perform further power analyses for different parameters.
- Compare this method to other existing methods.
- Investigate the effect of non-random sampling.
- Extend method to 3 markers.

Thank you!

Questions?