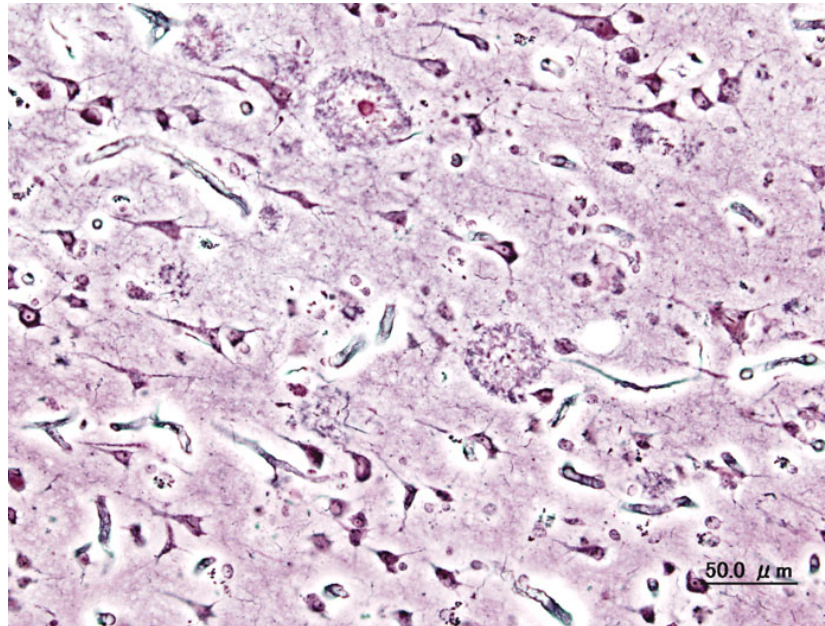


# A Strategy for Detecting Multiple Trait Loci in Disease Association Studies

Caroline Uhler in collaboration with Anna-Sapfo Malaspinas



# Overview

- **Player 1: Biology**
  - Disease association studies
- **Player 2: Statistics**
  - Interaction and how this can be detected
- **Player 3: Algebra**
  - Toric ideals

# 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

# $\chi^2$ -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,

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

## 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  $\varepsilon \in \{\pm 1\}$ ,

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

# Computing Markov bases

Given a  $d \times n$  matrix  $\mathcal{A} := (\mathbf{a}_1 \dots \mathbf{a}_n)$  with  $\mathbf{a}_1, \dots, \mathbf{a}_n \in \mathbb{Z}^d$ , define the homomorphism

$$\pi : \mathbb{C}[x_1, \dots, x_n] \rightarrow \mathbb{C}[t_1, \dots, t_d, t_1^{-1}, \dots, t_d^{-1}], \quad x_i \mapsto \mathbf{t}^{\mathbf{a}_i}.$$

The **toric ideal** of  $\mathcal{A}$  is defined as

$$I_{\mathcal{A}} := \ker(\pi).$$

**Theorem:** A finite family of functions

$(f_1, \dots, f_L : I_1 \times \dots \times I_n \rightarrow \mathbb{Z})$  is a Markov basis (to a given matrix of the form  $\mathcal{A}$ ) if and only if

$$\langle \mathbf{x}^{f_1^+} - \mathbf{x}^{f_1^-}, \dots, \mathbf{x}^{f_L^+} - \mathbf{x}^{f_L^-} \rangle = I_{\mathcal{A}}.$$

# Markov basis for testing for no 3-way interactions in $3 \times 3 \times 2$ tables

$$\begin{aligned}
 I_{\mathcal{A}} = & \langle x_3x_6x_{16}x_{17} - x_4x_5x_{15}x_{18}, \quad x_9x_{12}x_{16}x_{17} - x_{10}x_{11}x_{15}x_{18}, \\
 & x_1x_6x_{14}x_{17} - x_2x_5x_{13}x_{18}, \quad x_7x_{12}x_{14}x_{17} - x_8x_{11}x_{13}x_{18}, \\
 & x_3x_6x_{10}x_{11} - x_4x_5x_9x_{12}, \quad x_1x_6x_8x_{11} - x_2x_5x_7x_{12}, \\
 & x_1x_4x_8x_9 - x_2x_3x_7x_{10}, \quad x_1x_4x_{14}x_{15} - x_2x_3x_{13}x_{16}, \\
 & x_7x_{10}x_{14}x_{15} - x_8x_9x_{13}x_{16}, \quad x_4x_5x_7x_{12}x_{14}x_{15} - x_3x_6x_8x_{11}x_{13}x_{16}, \\
 & x_1x_6x_{10}x_{11}x_{14}x_{15} - x_2x_5x_9x_{12}x_{13}x_{16}, \\
 & x_2x_3x_7x_{12}x_{16}x_{17} - x_1x_4x_8x_{11}x_{15}x_{18}, \\
 & x_1x_6x_8x_9x_{16}x_{17} - x_2x_5x_7x_{10}x_{15}x_{18}, \\
 & x_1x_4x_9x_{12}x_{14}x_{17} - x_2x_3x_{10}x_{11}x_{13}x_{18}, \\
 & x_3x_6x_7x_{10}x_{14}x_{17} - x_4x_5x_8x_9x_{13}x_{18} \rangle
 \end{aligned}$$

# Extended Fisher's exact test

- i) Choose a model and determine its minimal sufficient statistics.
- ii) Compute the corresponding matrix  $\mathcal{A}$ .
- iii) Compute the corresponding toric ideal  $I_{\mathcal{A}}$  (e.g. by using CoCoA and an algorithm described by Sturmfels (1995)).
- iv) Run MCMC starting with the given data matrix and the Markov basis from iii).
- v) Analyze the posterior distribution of the  $\chi^2$ -statistic: The exact p-value of the given data can be approximated by the quantiles of this distribution.

## Simulations: Hap-Sample

For our simulations we used Hap-Sample (<http://hapsample.org/>):

- 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 for a set of user-defined loci

# Input to Hap-Sample

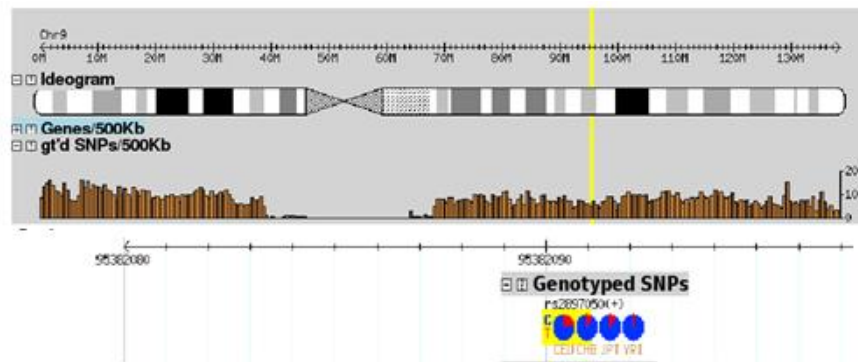
As a first step we restricted the data set to:

- CEU population
- 2 causative loci, unlinked
- Chromosome 9 and chromosome 13
  - 9935 SNPs total
- 2 interaction models: additive effects, epistasis
- 400 cases, 400 controls

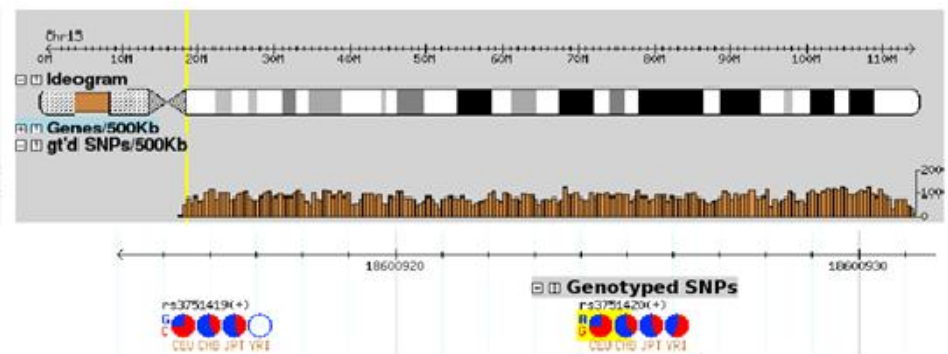
# Data matrix

- We selected the two SNPs with highest  $\chi^2$ -statistics for further analysis.
- The two selected SNPs have to be more than 1000 bp apart.

rs2897050 frequency=0.217 (CEU)

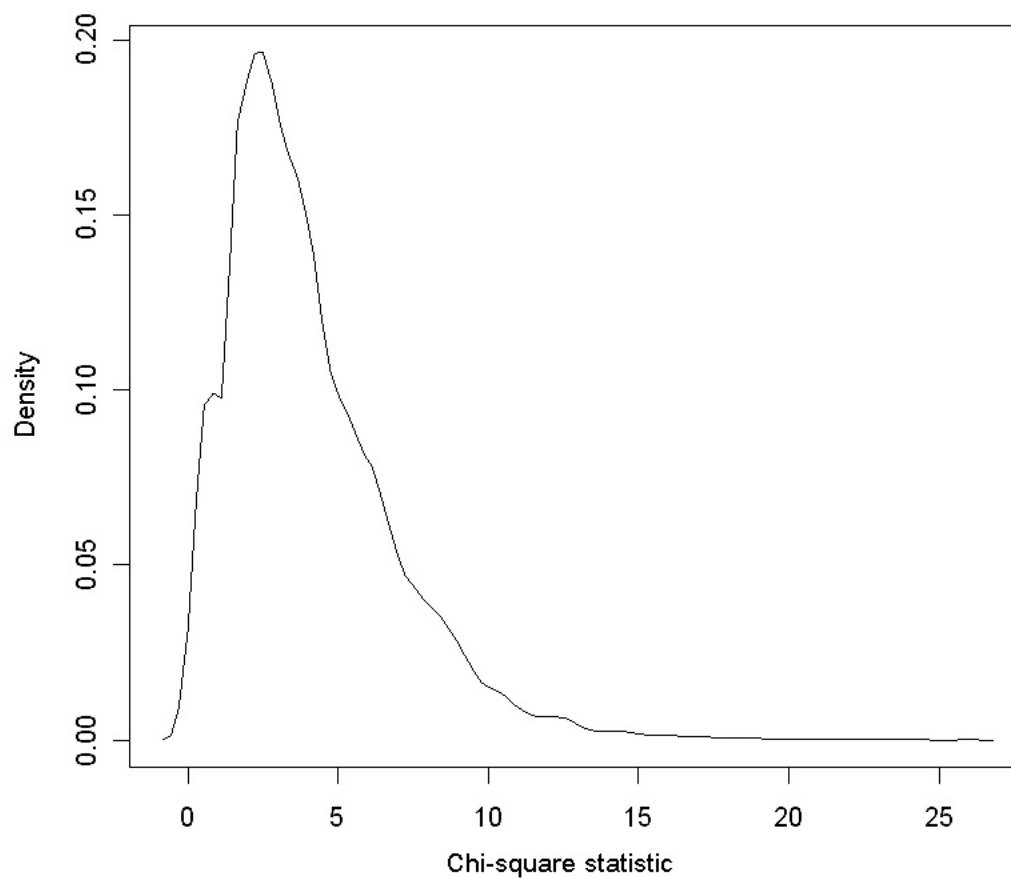


rs3751420, frequency=0.275 (CEU)



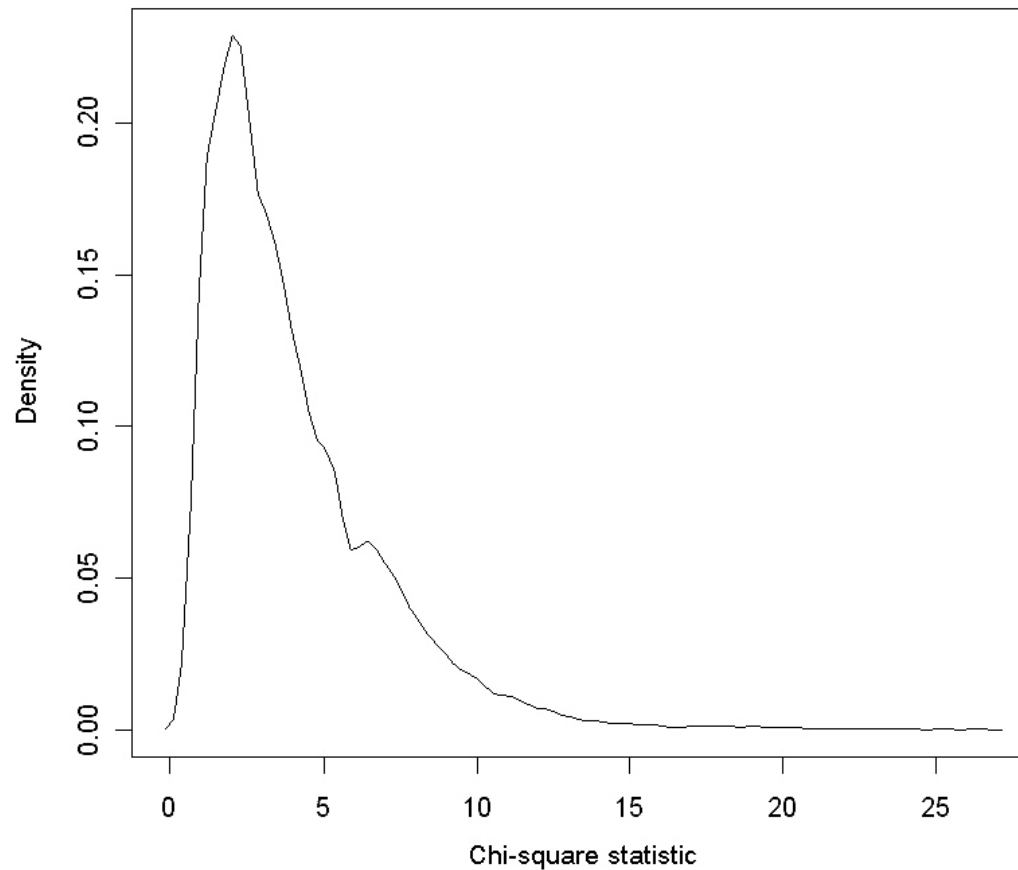
## Results: Additive model

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



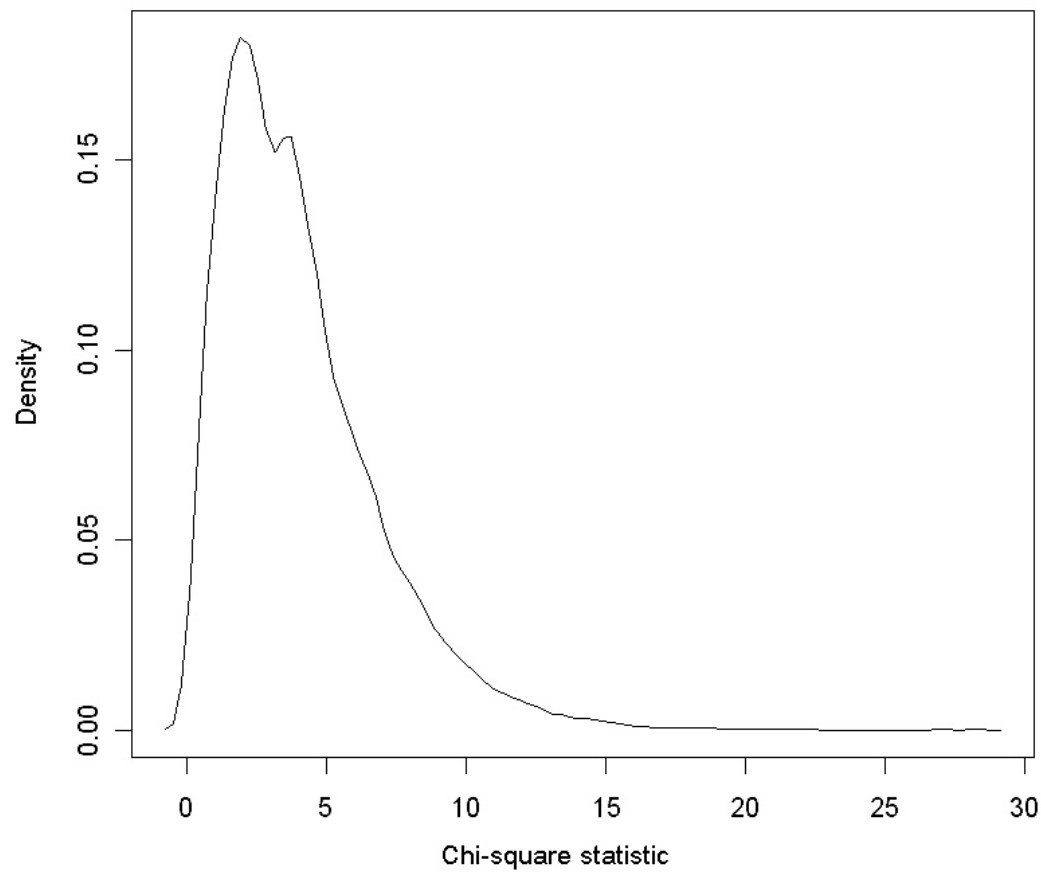
## Results: Model with epistasis

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



## Results: Control

Data matrix:  $\begin{pmatrix} 90 & 72 & 10 \\ 87 & 77 & 14 \\ 25 & 21 & 4 \end{pmatrix}$   $\begin{pmatrix} 36 & 108 & 24 \\ 41 & 119 & 31 \\ 6 & 27 & 8 \end{pmatrix}$   $\chi^2$ -statistic = 0.87



**Thank you!**

**Questions?**