Gene Mapping using Coalescence Theory

Ola Hössjer

July, 2009

Joint work with Linda Hartman, Keith Humphreys and Fredrik Olsson

Problem Statement

- Goal:
  - Test if given chromosomal regions harbors a disease causing mutation.
- Data from a nr. of seemingly unrelated individuals):
  - Affection status (case or control).
  - DNA marker data from two copies of the chromosomal region[1]
- Strategy:
  - Check if cases' chromosome regions tend to have similar DNA.
- Rationale:
  - Common mutated ancestor has passed on mutation and surrounding DNA material.

[1] Inherited from the mother and father respectively.

Hypotheses and Data

Let

$$\begin{aligned} (0,1) &= \text{chromosomal region} \\ m &= \text{nr. of individuals} \\ K &= \text{nr. of markers} \end{aligned}$$

and test

$$\begin{aligned} H_0 &: \text{ Disease mutation unlinked to } (0,1), \\ H_1 &: \text{ Disease mutation within } (0,1), \end{aligned}$$

using test statistic Z based on data

$$\begin{aligned} \mathbf{Y} &= 1 \times m \text{ phenotype vector} \\ \mathbf{g} &= m \times K \text{ SNP marker genotype matrix} \end{aligned}$$

and permutation test

$$\text{p-value} = \frac{1}{Q} \sum_{q=1}^{Q} 1_{\{Z_q \geq Z\}}$$

where Z_q is test statistic based on $\mathbf{g}$ and q:th random permutation of $\mathbf{Y}$.

Test statistics

- Maximal χ^2 test statistic

$$Z = \max_{1 \leq k \leq K} Z(x_k)$$

where

  - $0 \leq x_k \leq 1$ is k:th marker position,
  - $Z(x_k)$ is χ^2 test statistic of independence between phenotypes and k:th marker.

- Maximal lod score

$$Z = \max_{0 \leq x \leq 1} \log_{10} \frac{L(x)}{L(\infty)},$$

where

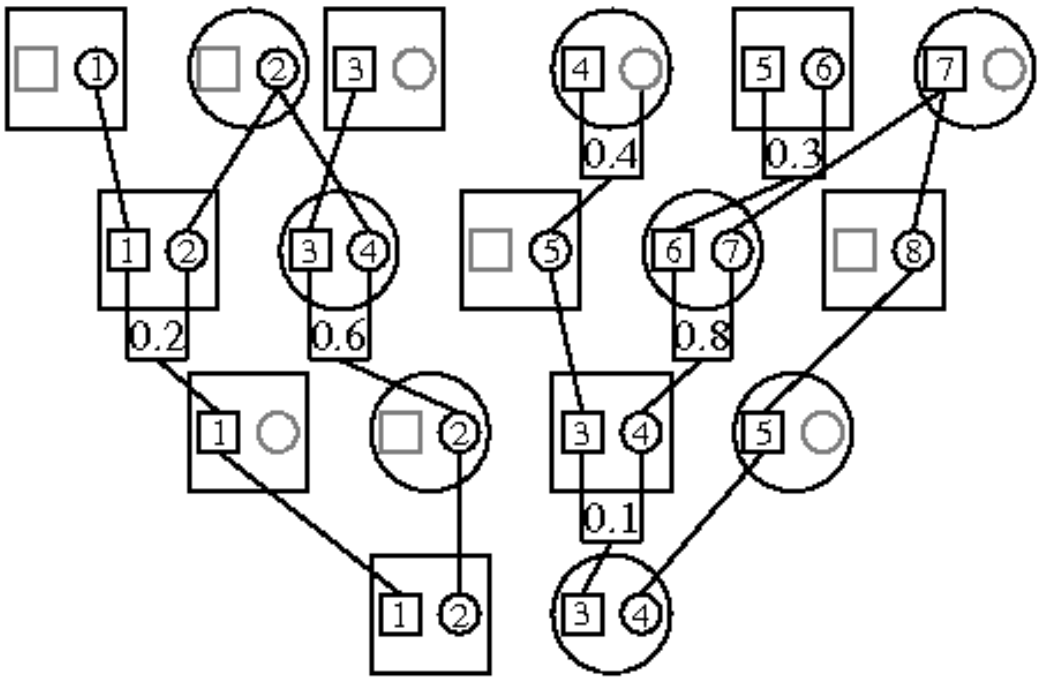
  - $L(x) = P_x(\mathbf{g}|\mathbf{Y})$ is likelihood assuming mutation at x.

Expanded Likelihood

$$P_x(\mathbf{g}|\mathbf{Y}) = \sum_{\mathcal{A},\mathcal{M}} P(\mathbf{g}|\mathcal{A})P_x(\mathcal{A}|\mathcal{M})P(\mathcal{M}|\mathbf{Y})$$

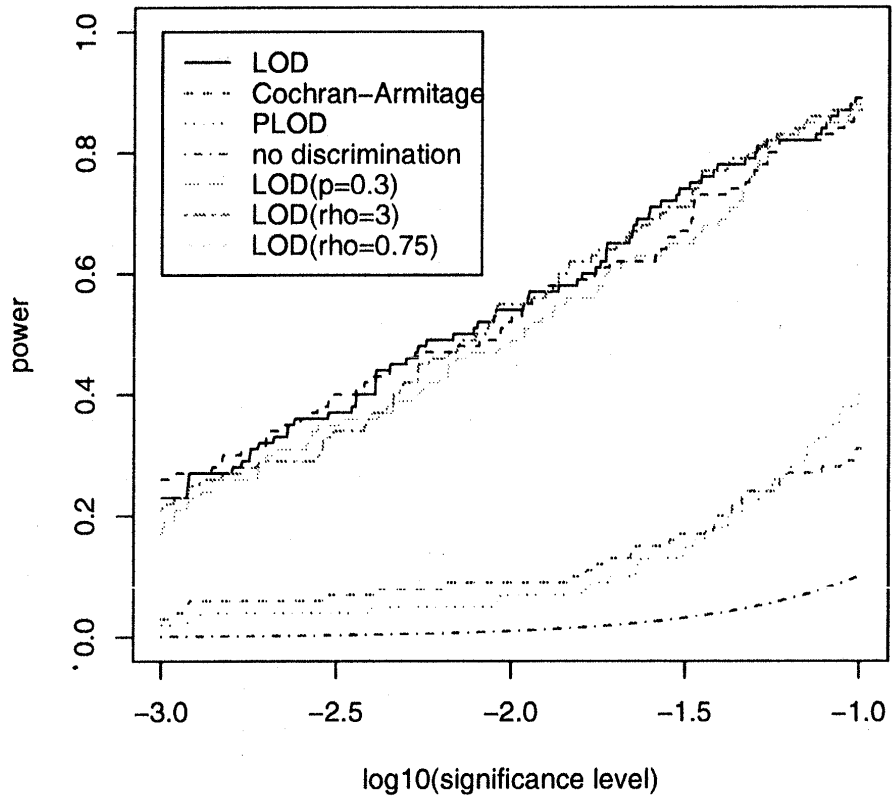
- $\mathcal{A}$ is joint ancestry of sample along (0,1)
- $\mathcal{M}$ contains mutation status of all $2m$ chromosomes
- Likelihood depends on
  - Penetrance of mutated variant
  - Allele frequency of mutated variant
  - Population genetic model for $\mathcal{A}$
- Fast HMM algorithm to compute likelihood ratio.

Ancestral Recombination Graph $\mathcal{A}$



- **Founder population** $G(=3)$ generations back.
- **Coalescence**: Common 'parent' of two chromosomes (merge).
  - Colescence rates λ_M and λ_U for mutated and unmutated chr.
- **Recombination**: Two 'parents' of one chromosome (split).
  - Recombination rate ρ for all chr.

ROC Curves, With/Without Parameter Misspecification

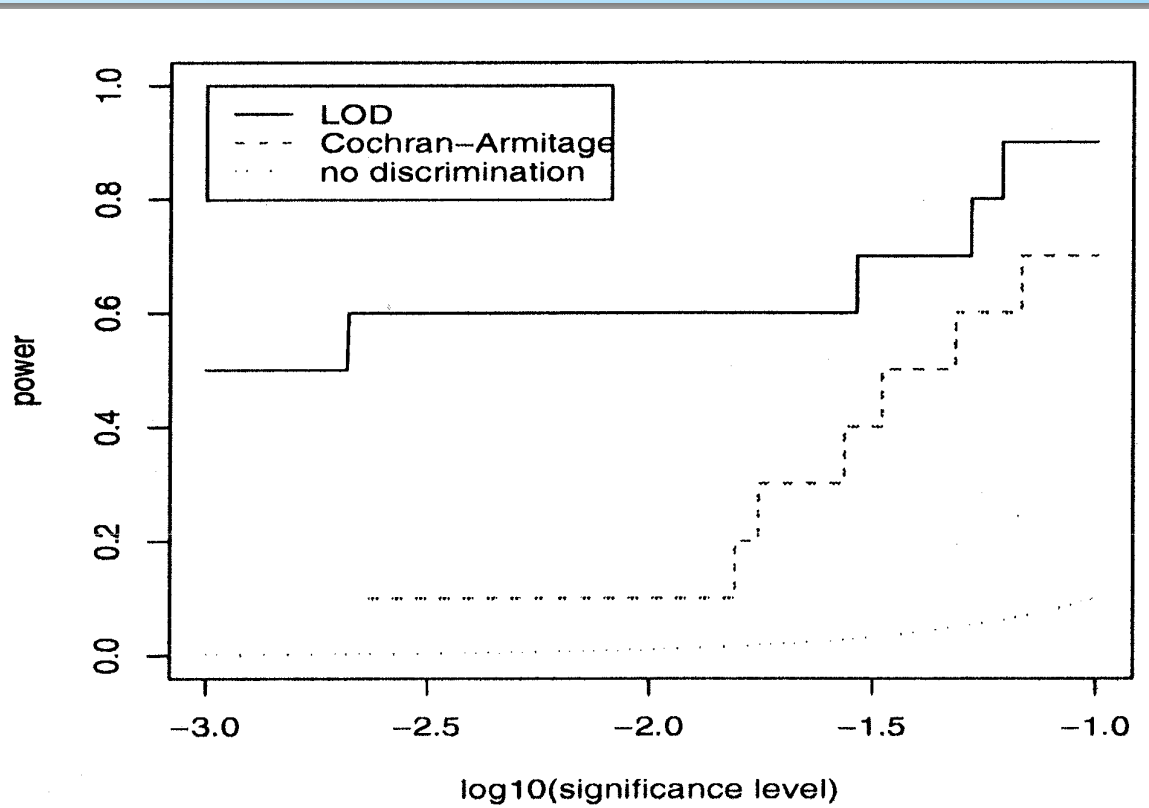


ROC curve: $\beta(\alpha) = \sum_{i=1}^{N} 1_{\{p_i \leq \alpha\}}/N,\ (N = 100)$

p_i = i^{th} simulated p-value (10 000 permutations),

m	=	400 (200 cases, 200 controls),
K	=	10 markers,
P(mutated variant)	=	0.1 (misspecified as 0.3 in one curve),
increased risk per mutated variant	=	2
ρ	=	1.5 (misspecified as 0.75 and 3 in two curves).

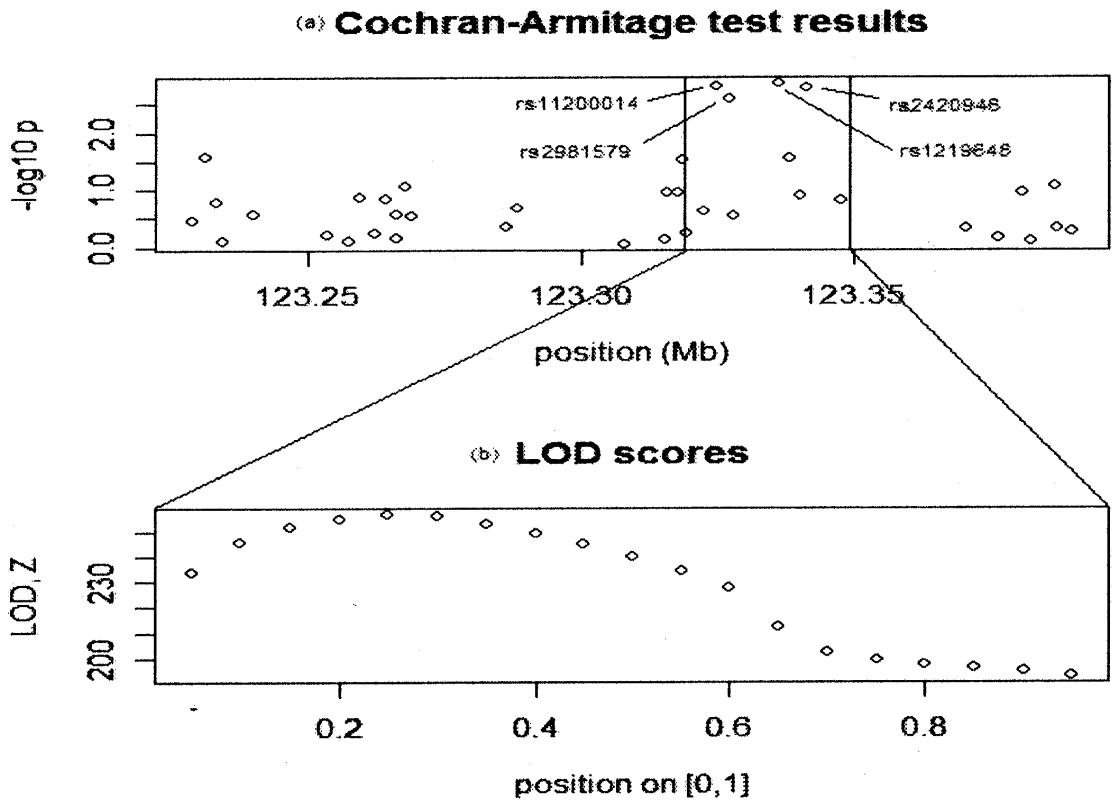
ROC Curve, Misspecified Genealogy



$\mathcal{A}$ = Neutral Wright-Fisher conditioned on ascertainment (misspecified as star topology Markov model),

N	=	10 simulations,
Q	=	10000 permutations,
m	=	1000 (500 cases, 500 controls),
K	=	10 markers,
P(mutated variant)	=	0.2 (0.05 in analysis),
increased risk per mutated variant	$\approx$	1.4 (1.8 in analysis)
ρ	=	1.5.

Swedish Breast Cancer Data Set



400 cases, 400 controls

Upper: 160 kb region around the gene FGFR2
38 markers
Pointwise p-values (Cochran Armitage test $\approx$ 1 df χ^2 test)

Lower: Lod score along chromosomal subregion
Subregionwide p-values: 0.0102 (CA) and 0.0029 (Lod score)

References

Einarsdóttir, K., Humphreys K., Bonnard C. et al. (2006) Linkage disequilibrium mapping of CHEK2: common variation and breast cancer risk. *PLoS Med*, 3:e168.

Griffiths, R. and Marjoram, P. (1997). An ancestral recombination graph. In Donnelly, P. and Tavaré, S., editors, *Progress in Population Genetics and Human Evolution*, pages 257–270. Springer, New York.

Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theor Popul Biol*, 23(2):183–201.

Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338.

Hössjer, O., Hartman, L. and Humphreys, K. (2009). Ancestral recombination graphs under nonrandom ascertainment, with applications to gene mapping. *Statistical Applications of Genetics and Molecular Biology*, **8**(1), Article 35.

Morris, A.P., Whittaker, J.C. and Balding, D.J. (2002). Fine-scale mapping of disease loci via shattered coalescent modeling. *Am J Hum Genet*, 70:686-707.

Wang, Y. and Rannala, B. (2005). In silico analysis of disease-association mapping strategies using the coalescent process and incorporating ascertainment and selection. *Am J Hum Genet*, 76(6):1066–1073.

Wiuf, C. and Hein, J. (1999). The ancestry of a sample of sequences subject to recombination. *Genetics*, 151(3):1217–1228.

Zöllner, S. and Pritchard, J.K. (2005). Coalescent-based association mapping and fine mapping of complex trait loci. *Genetics*, 169(2), 1071-1092.