Gene Mapping using Coalescence Theory

Ola Hössjer

July, 2009

Joint work with Linda Hartman, Keith Humphreys and Fredrik Olsson
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

\[(0, 1) = \text{chromosomal region}\]
\[m = \text{nr. of individuals}\]
\[K = \text{nr. of markers}\]

and test

\[H_0: \text{Disease mutation unlinked to (0, 1)},\]
\[H_1: \text{Disease mutation within (0, 1)},\]

using test statistic \(Z\) based on data

\[Y = 1 \times m\text{ phenotype vector} \]
\[g = m \times K\text{ SNP marker genotype matrix} \]

and permutation test

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

where \(Z_q\) is test statistic based on \(g\) and \(q:\text{th random permutation of } Y\).
Test statistics

- Maximal $\chi^2$ 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 $\chi^2$ 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
- $L(x) = P_x(g|Y)$ is likelihood assuming mutation at $x$. 

Ola Hössjer | Gene Mapping using Coalescence Theory
Expanded Likelihood

\[ P_x(g|Y) = \sum_{A,M} P(g|A)P_x(A|M)P(M|Y) \]

- \( A \) is joint ancestry of sample along \((0,1)\)
- \( M \) contains mutation status of all \(2m\) chromosomes
- Likelihood depends on
  - Penetrance of mutated variant
  - Allele frequency of mutated variant
  - Population genetic model for \( A \)
- 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).
  - Coalescence rates $\lambda_M$ and $\lambda_U$ for mutated and unmutated chr.

- **Recombination**: Two 'parents' of one chromosome (split).
  - Recombination rate $\rho$ for all chr.
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 $= 2$

$\rho = 1.5$ (misspecified as 0.75 and 3 in two curves).
A = Neutral Wright-Fisher conditioned on ascertainment (misspecified as star topology Markov model),

\[ N = 10 \text{ simulations,} \]
\[ Q = 10000 \text{ permutations,} \]
\[ m = 1000 \text{ (500 cases, 500 controls),} \]
\[ K = 10 \text{ markers,} \]
\[ P(\text{mutated variant}) = 0.2 \text{ (0.05 in analysis),} \]
\[ \rho = 1.5. \]
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 $\chi^2$ test)

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


