The effect of missing information on gene mapping

Benjamin Yakir
Dept. of Statistics
The Hebrew University
and Stat. & App. Prob., NUS

Workshop on Genomics, IMS, Singapore
November, 2005

Topics

- Admixture mapping.
- The statistic and its non-centrality parameter.
- The effect of reconstruction.
- Other effects.

Admixture mapping

- **Population based** (characteristics of experimental genetics): Co-segregation of

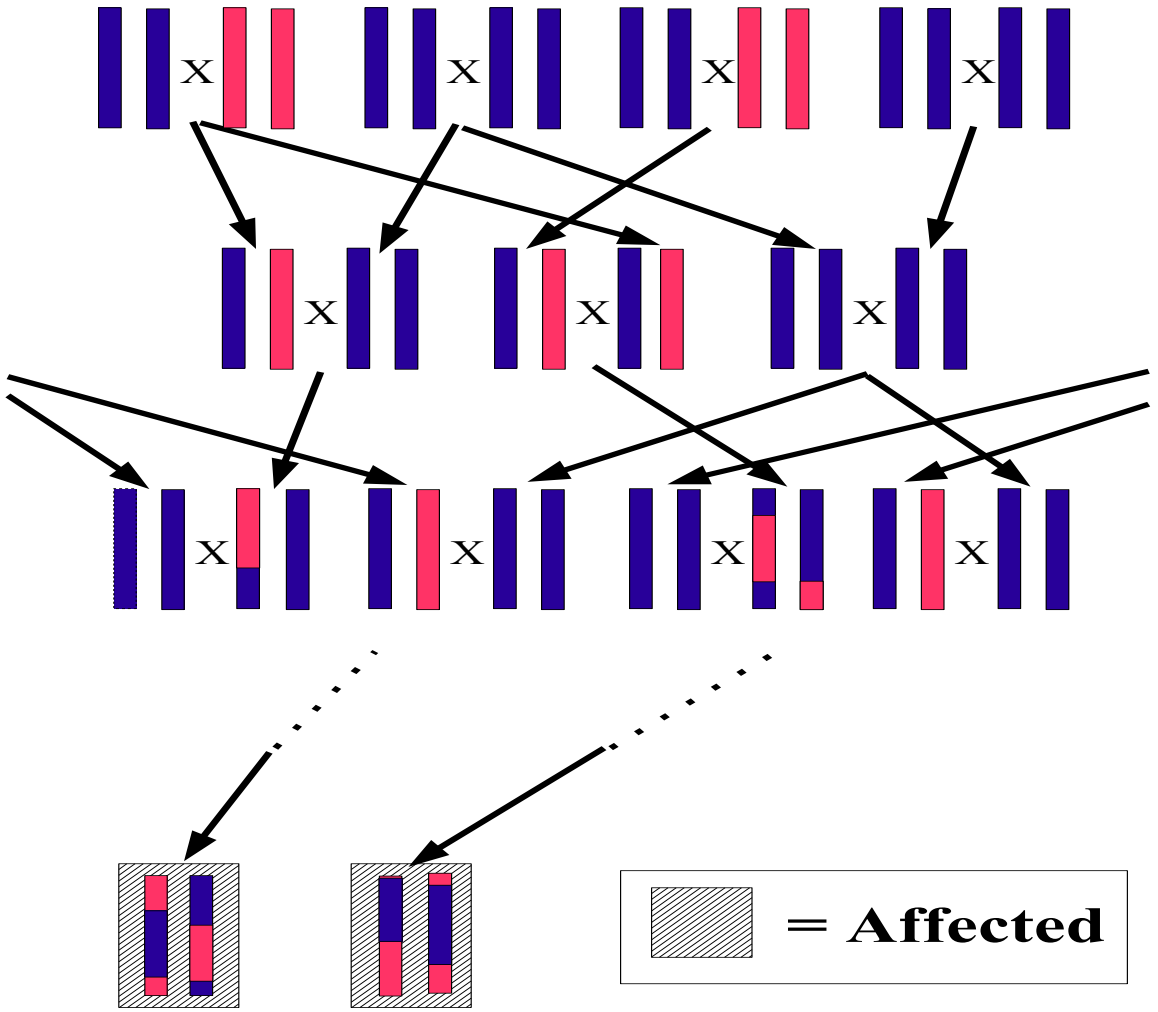
  Phenotypes = disease status, and

  Founder origin = detected by molecular markers.

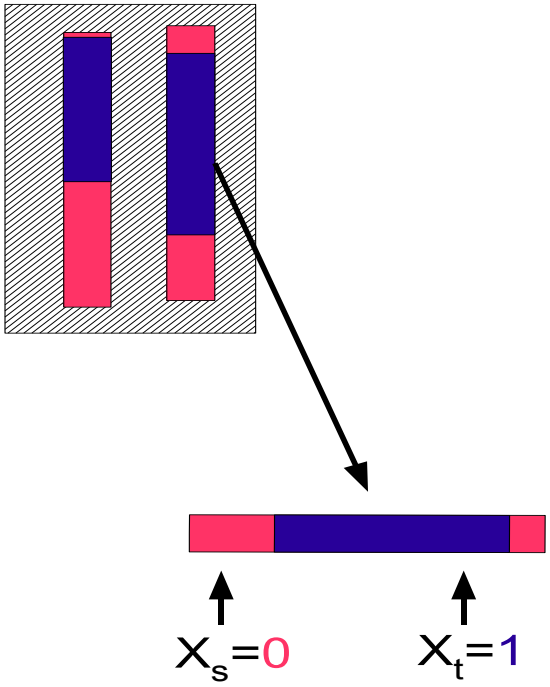
- **Affected only**: Scanning for discrepancies from expected background levels.

- **A case-random design**: Scanning for discrepancies between cases and random controls.

Admixed Population



A random gamete



$P(X_t=1) = p$

$P(X_t=0) = 1-p$

A random gamete

Denote:

$$\begin{aligned} X_t &= \text{The population source of locus } t \text{ (0 or 1).} \\ p &= \mathbb{P}(X_t = 1), \text{ for a random gamete.} \\ D,\, d &= \text{Two alleles of a gene linked to } t. \end{aligned}$$

Then

$$p = \overbrace{\frac{p\,\mathbb{P}(D \,|\, X_t = 1)}{p\,\mathbb{P}(D \,|\, X_t = 1) + (1-p)\mathbb{P}(D \,|\, X_t = 0)}}^{(1)} \times \mathbb{P}(D)$$

$$+ \overbrace{\frac{p\,\mathbb{P}(d \,|\, X_t = 1)}{p\,\mathbb{P}(d \,|\, X_t = 1) + (1-p)\mathbb{P}(d \,|\, X_t = 0)}}^{(2)} \times \mathbb{P}(d)$$

A susceptibility gene

- Terms (1) and (2) are determined by:

  1. $p \Rightarrow$ History of admixture, and

  2. $\mathbb{P}(D \,|\, X_t = i) \Rightarrow$ Characteristics of the founders.

- $\mathbb{P}(D) = 1 - \mathbb{P}(d) \Rightarrow$ The penetrance associated with the gene and the selected sampling.

- Hardy-Weinberg + multiplicative GRR $\Rightarrow$ Binomial distribution of D alleles among cases and controls.

Distribution of X_t at a QTL

Among cases:

- $X_t \sim B(p_\theta, 2)$, where
- $\theta = \log\Big([p_\theta(1-p)]/[(1-p_\theta)p]\Big)$,
- and:

$$\frac{p_\theta - p}{p(1-p)} = \frac{\mathbb{P}_\theta(D) - \mathbb{P}(D)}{\mathbb{P}(D)(1-\mathbb{P}(D))} \times \Big[\mathbb{P}(D \,|\, 1) - \mathbb{P}(D \,|\, 0)\Big]$$

Topics

- Admixture mapping.
- The statistic and its non-centrality parameter.
- The effect of reconstruction.
- Other effects.

The test statistic

- For a sample of n affected: $S_t = \sum_{i=1}^{n} X_{it} \sim B(2n, p_\theta)$.

- Reject $\theta = 0$ for a given locus t if $|Z_t|$ is large.
  (Z_t = standardized version of S_t.)

- Consider all t over the entire length of the genome.

- Significance is discounted by multiple testing.

The non-centrality parameter

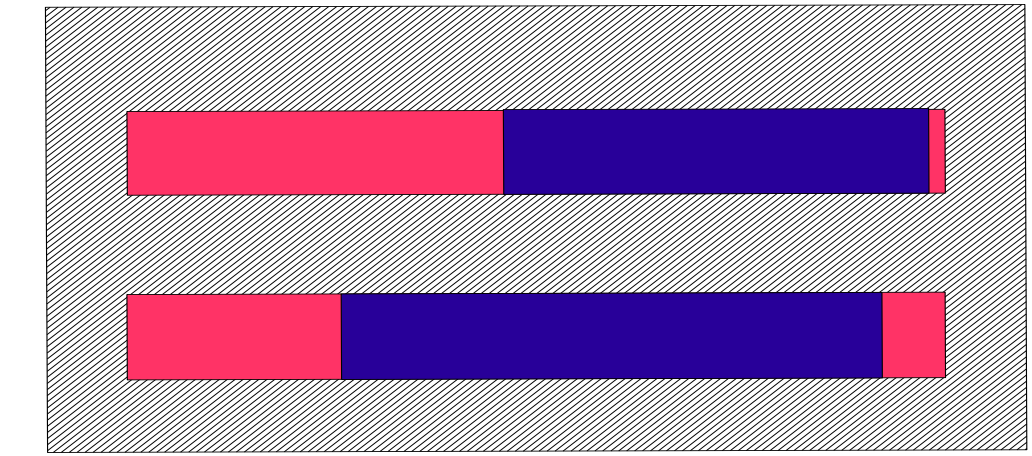
- Assume a QTL at t.
- Let $\xi = \mathbb{E}_\theta(Z_t)$ be the non-centrality parameter.
- Then:

$$\begin{aligned} \xi &= \mathbb{E}\Big(Z_t e^{\theta S_t - 2n\psi(\theta)}\Big) \\ &\approx \theta\,\mathbb{E}\Big(Z_t(S_t - 2n\dot{\psi}(0))\Big) \\ &= \theta\left\{2n(p(1-p)\right\}^{1/2}. \end{aligned}$$

Genotypes and the reconstruction of X_t

- Unfortunately, X_t cannot be observed directly.
- Instead, one observes molecular markers.
- The distribution of markers may depends on the state of X_t.
- Consequently, one may reconstruct the state from the genotypic information.

Genotypes



A/A G/C G/T

$P(A/A) = p_0^2$

$P(G/C) = p_1(1-p_0)+p_0(1-p_1)$

$P(G/T) = 2p_1(1-p_1)$

A hidden markov model

- $X = \{X_t\}$ = Population origin within an individual.
- Assumed to be a stationary, reversible and continuous markov process. ($Q = (q_{ij})$ = transition rates.)
- Hardy-Weinberg $\Rightarrow X_t = X_t^F + X_t^M$, independent.
- G = The genotypic information for the individual.
- If the components of G are conditionally independent given $X \Rightarrow (X, G)$ = HMM.

Topics

- Admixture mapping.
- The statistic and its non-centrality parameter.
- The effect of reconstruction.
- Other effects.

The reconstructed scanning process

- Assume Q_i and the conditional distributions of G_i are known.

- $\hat{X}_{it} = \mathbb{E}(X_{it} \,|\, G_i) =$ the reconstructed process.

- $\mathbb{E}(\hat{X}_{it}) = \mathbb{E}(\hat{X}_{it}) = 2p$ and $\sigma_i^2 = \mathbb{V}ar(\hat{X}_{it})$.

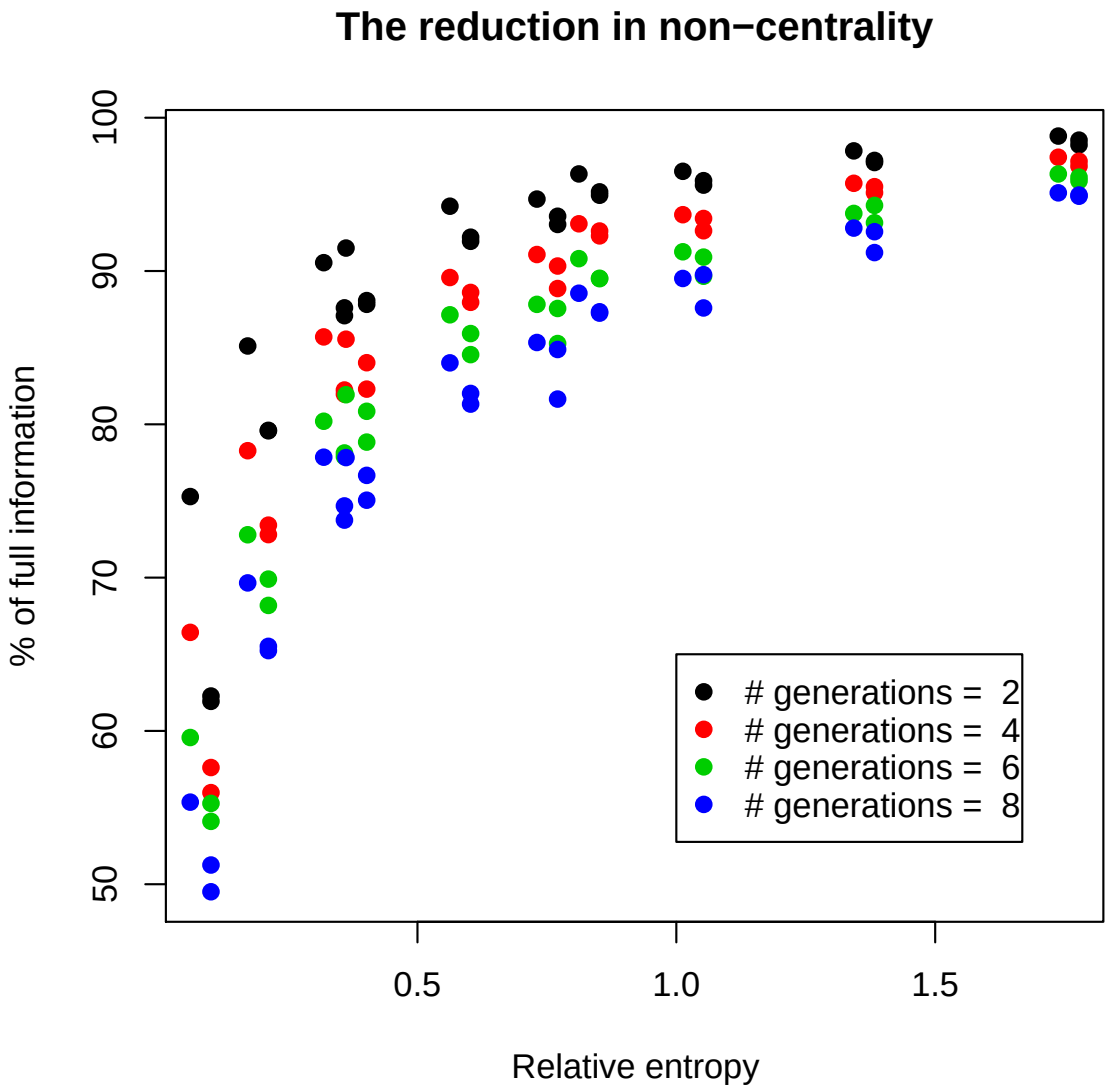
- $\hat{Z}_t =$ Scanning statistic $= \frac{\sum_{i=1}^{n}(\hat{X}_{it} - 2p)}{\sqrt{\sum_{i=1}^{n}\sigma_i^2}}$.

The non-centrality of the reconstructed statistic

- For a QTL at t:

$$\begin{aligned}
\mathbb{E}(\hat{Z}_t) &= \mathbb{E}\Big(\hat{Z}_t e^{\theta S_t - 2n\psi(\theta)}\Big) \\
&\approx \theta\,\mathbb{E}\Big(\hat{Z}_t(S_t - 2n\dot{\psi}(0))\Big) \\
&= \xi \times \frac{\mathbb{C}ov(\hat{Z}_t, S_t)}{\{2n(p(1-p)\}^{1/2}} \\
&= \xi \times \Big\{\frac{1}{n}\sum_{i=1}^{n} \frac{\sigma_i^2}{2p(1-p)}\Big\}^{1/2}.
\end{aligned}$$

- Note that $\sigma_i^2 < 2p(1-p)$.


The reduction in non−centrality
% of full information
100
90
80
70
60
50
0.5
1.0
1.5
Relative entropy
generations = 2
generations = 4
generations = 6
generations = 8

A basic equation

- $\mathbb{P}(X_t{=}i|G) = \frac{\mathbb{P}(X_t{=}i,G)}{\mathbb{P}(G)}$: Incomplete likelihood ratio.

- From the likelihood ratio identity:

$$\begin{aligned}\mathbb{E}\Big[\mathbb{P}(X_t{=}j|G)\cdot\mathbb{P}(X_t{=}i|G)\Big] &= \mathbb{E}\Big[\mathbb{P}(X_t{=}j|G); X_t = i\Big] \\ &= \pi_i\mathbb{E}\Big[\mathbb{P}(X_t{=}j|G)\,|\,X_t = i\Big].\end{aligned}$$

- $\pi_i = \mathbb{P}(X_t = i)$: The stationary probability.

An asymptotic approximation of σ^2

- Let $j \neq i$ and consider $\widehat{\pi}_j = \mathbb{P}(X_t{=}j|G)$.
- Assume transition rates are low: $q_{ij} \to 0$.
- G is relatively informative in $[t-r, t+r]$.
- $\widehat{\pi}_j$ small, but non-negligible, only when
  1. $\{X_{t-r} = i, X_t = i, X_{t+r} = j\}$ or
  2. $\{X_{t-r} = j, X_t = i, X_{t+r} = i\}$.

An asymptotic approximation of σ^2 (cont.)

It follows that:

$$\mathbb{E}(\widehat{\pi}_j \widehat{\pi}_i) \approx 2\pi_i q_{ij} \mathbb{E}\left[\frac{R_r^- R_r^+}{R_r^- + R_r^+}\right],$$

where

$$\begin{aligned} R_r^- &= \int_0^r e^{\ell(i,j,-s)-\ell(i,j,0)} ds, \\ R_r^+ &= \int_0^r e^{\ell(i,j,s)-\ell(i,j,0)} ds \text{ and} \\ \ell(i,j,u) &= \text{conditional log-likelihoods of } G. \end{aligned}$$

Analytical expressions

- Assume $\ell(i,j,u) - \ell(i,j,0) \approx$ a Brownian motion.

- Then $1/R_\infty \sim \text{Gamma}\Big(-2\mu/\sigma^2, \sigma^2/2\Big)$.

- If $-\mu = \sigma^2/2$ then

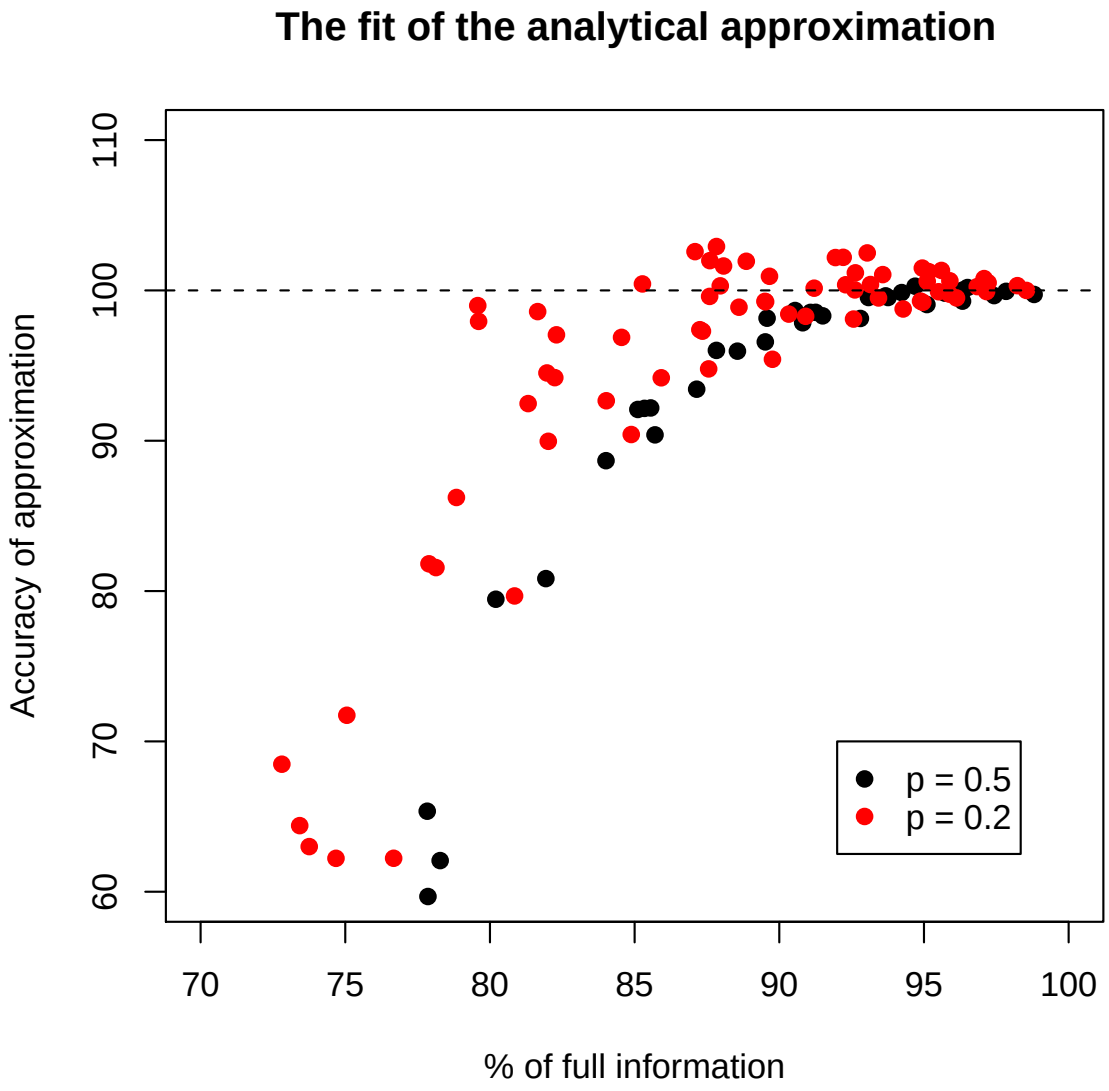
$$\mathbb{E}\left[\frac{R_r^- R_r^+}{R_r^- + R_r^+}\right] \approx H(\mu_{ij}, \mu_{ji})$$

$$= \begin{cases} -\frac{\mu_{ij}\mu_{ji}}{\mu_{ij}-\mu_{ij}} \log(\mu_{ij}/\mu_{ji}), & \text{if } \mu_{ij} \neq \mu_{ji}, \\ -\mu_{ij}, & \text{if } \mu_{ij} = \mu_{ji}. \end{cases}$$

Analytical expressions (cont.)

For admixture mapping

- Δ = Distance between markers.
- $\mu_{ij} = \mathbb{E}\Big[\log \Big\{ \frac{\mathbb{P}(G_t \,|\, X_t{=}j}{\mathbb{P}(G_t \,|\, X_t{=}i)} \Big\} \,\Big|\, X_t = i \Big] / \Delta.$
- $H_{ij} = H(\mu_{ij}, \mu_{ji}).$
- $\widehat{\sigma}^2 \approx 2p(1-p) - 2\Big\{ (1-p)^2 q_{01} H_{01} + p^2 q_{21} H_{21} \Big\}.$


The fit of the analytical approximation
Accuracy of approximation
110
100
90
80
70
60
70
75
80
85
90
95
100
% of full information
p = 0.5
p = 0.2

Topics

- Admixture mapping.
- The statistic and its non-centrality parameter.
- The effect of reconstruction.
- Other effects.

Other effects:

- Covariance structure and significance level.
- Estimation of unknown parameters — both global or local.
- Robustness to modeling assumptions — Markov process, Brownian process.
- Statistic which involves sums of dependent components.

Thank you!