A Probabilistic Framework for the Statistics of Selective Breeding

Benny Yakir
Department of Statistics
The Hebrew University

M.S.R.I.
February 10, 2004

Topics

- Selective breeding and genotyping.
- The basic Markov processes.
- Large deviations for a function of basic processes.
- Analytic approximations and simulation results.

Selective breeding and genotyping

- Selective breeding may enhance the relative part of the genetic component of the phenotypic variability.

- It may reduce the genetic heterogeneity by dissecting a complex trait into simpler components.

- Selective genotyping may reduce the cost of genotyping.

The Basic Markov processes

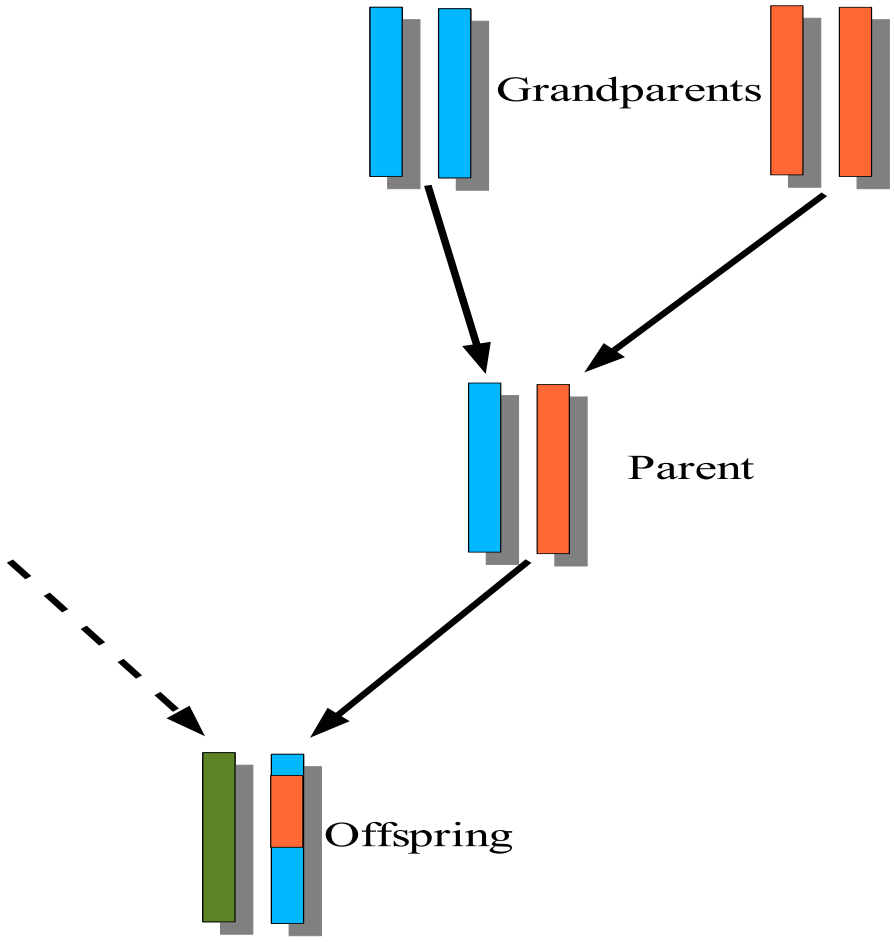
- In selective breeding one is attempting to detect disruptions in the pattern of segregation of alleles.

- We consider scanning statistics which can be represented as a function of basic processes.

- The basic process: Identity of the parental source of each locus for a given chromosome of an offspring.

- For a chromosome in which the selection force is irrelevant the basic process will be a two-states Markov process.

Segregation of Chromosomes



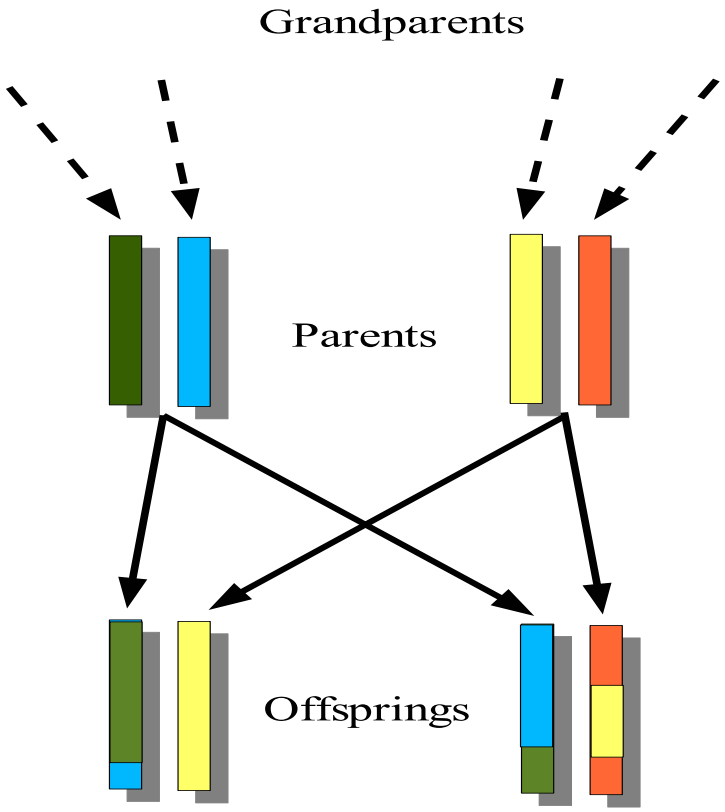
Example: Identity By Decent mapping in Affected Sib-Pairs

- A large collection of affected pairs of siblings is recruited.

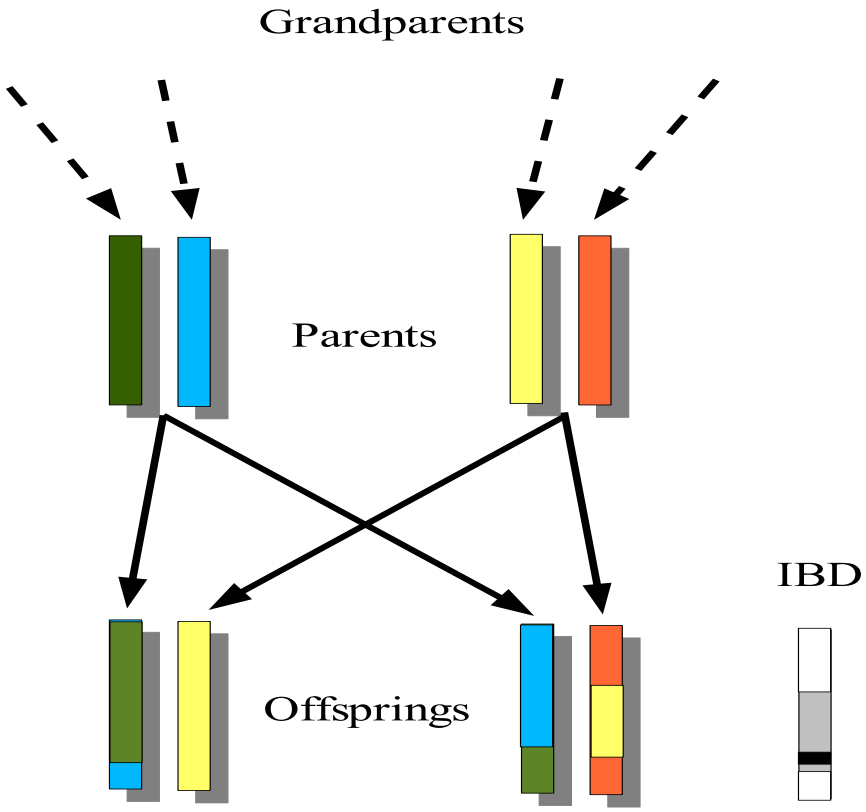
- In each pair the parental source of each locus is examined. Identities of parental source within siblings are recorded.

- Scanning is based on the detection of loci with an access level of such identities.

IBD mapping



IBD mapping



The scanning statistic for IBD mapping:

- Let $\{X_{tji}\}$ be the genotypic status at locus t of the 4 chromosomes for n pedigrees.

- One may observe $\{X_{ti} := I(X_{t1i} = X_{t3i}) + I(X_{t2i} = X_{t4i})\}$.

- The scanning statistic is $\max_t X_t$, for $X_t = \sum_{i=1}^{n} X_{ti} \sim B(2n, 1/2)$.

- In the limit, X_t is the Ornstein-Ulenbeck process. The distribution of the scanning statistic may be obtained from the distribution of the maximum of this process.

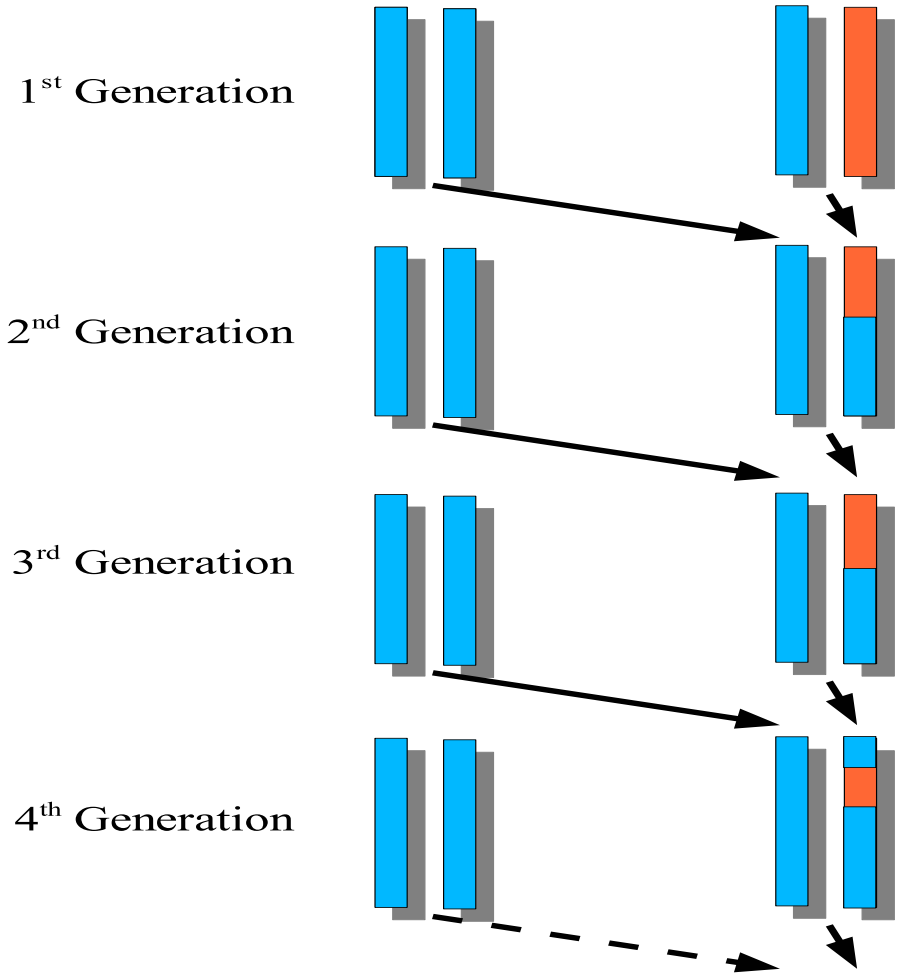
Example: Congenic lines

- Animals that express the trait are repeatedly back-crossed with a neutral inbred strain.

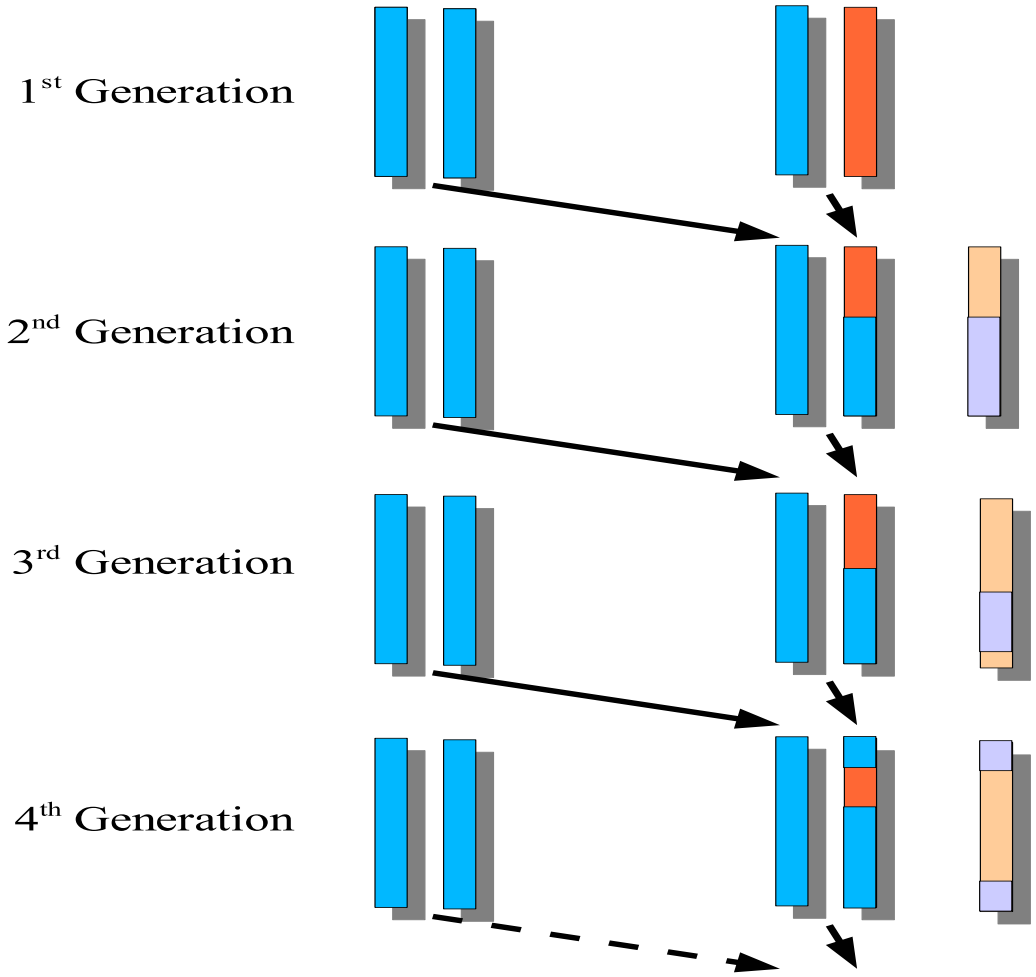
- In each generation the parental source of each locus is examined. Heterozygous loci are recorded.

- Scanning is based on the detection of loci which remain heterozygous for many generations.

A congenic line



A congenic line



The scanning statistic for congenic strains:

- Let $\{X_{ti}\}$ be the genotypic status at locus t for the chromosome segregated from the donating parent.

- The Likelihood-Ratio test statistic for a given congenic line at locus t is $X_t = \min\{i : X_{ti} = 0\} \sim G(1/2)$.

- The problem: $\mathbb{P}(\max_t X_t = b) =?$

Large deviations for integer-valued processes:

- Let $\{X_t; 0 \leq t \leq L\}$ be an integer valued random process.

- The marginal distribution of X_t is independent of t.

- A basic question:

$$\mathbb{P}\Big(\max_{0\leq t\leq L} X_t = b\Big) = ?$$

A basic identity:

- The Log-Moment Generating Function: $\psi(\theta) = \log \mathbb{E}[\exp\{\theta X_t\}]$.

- Define: $T_j = \Big\{t : X_t = \max_{0\le s\le L} X_s - j\Big\}$. Then:

$$\mathbb{P}\Big(\max_{0\le t\le L} X_t = b\Big) = e^{\psi(\theta)-\theta b}\frac{1}{L}\int_0^L \mathbb{E}_t\left[\frac{L}{\sum_{j=0}^{J}|T_j|e^{-\theta j}}; \max_{0\le s\le L} X_s = b\right]d$$

where $\mathcal{L}_t(X_t) = \mathbb{P}_\theta$ and $\mathcal{L}_t(\{X_s\}|X_t) = \mathbb{P}$.

The basic identity for a congenic line:

- Let: $C \sim \text{Poisson}(bL/100)$, $\nu \sim \text{Unif}(1, 2, ..., C+1)$, and $X(\nu) \sim G\Big(1 - e^{\theta}/2\Big)$.

- Define: $S = \sum_{h=1}^{\nu} |\Delta_{(h)}| \cdot e^{\theta(X_{(h)}-b)} + \sum_{h=\nu}^{C+1} |\Delta_{(h+}$ $e^{\theta(X_{(h)}-b)}$.

- Then:

$$\mathbb{P}\Big(\max_{0\le t\le L} X_t = b\Big) =$$

$$e^{\psi(\theta)-\theta b} \cdot \mathbb{E}\left(\mathbb{E}\left[\frac{L}{S}; \max_{1\le h\le C+1} X_{(h)} = b \middle| C, \nu, X_{(\nu)}\right]\right)$$

Analytic approximation (first order):

$$\approx e^{\psi(\theta)-\theta b}\cdot\mathbb{P}_\theta(X_{(\nu)}=b)\cdot\mathbb{E}\left(\mathbb{E}\left[\frac{L}{|\Delta_{(\nu)}|+|\Delta_{(\nu+1)}|}\middle|C\right]\right).$$

However,

$$e^{\psi(\theta)-\theta b}\cdot\mathbb{P}_\theta(X_{(\nu)}=b)=\mathbb{P}(X_{(\nu)}=b)=2^{-b},$$

and

$$\mathbb{E}\left[\frac{L}{|\Delta_{(\nu)}|+|\Delta_{(\nu+1)}|}\middle|C\right]=C+1.$$

Thus

$$\mathbb{P}\Big(\max_{0\le t\le L}X_t=b\Big)\approx\left[\frac{Lb}{100}+1\right]\cdot 2^{-b}.$$

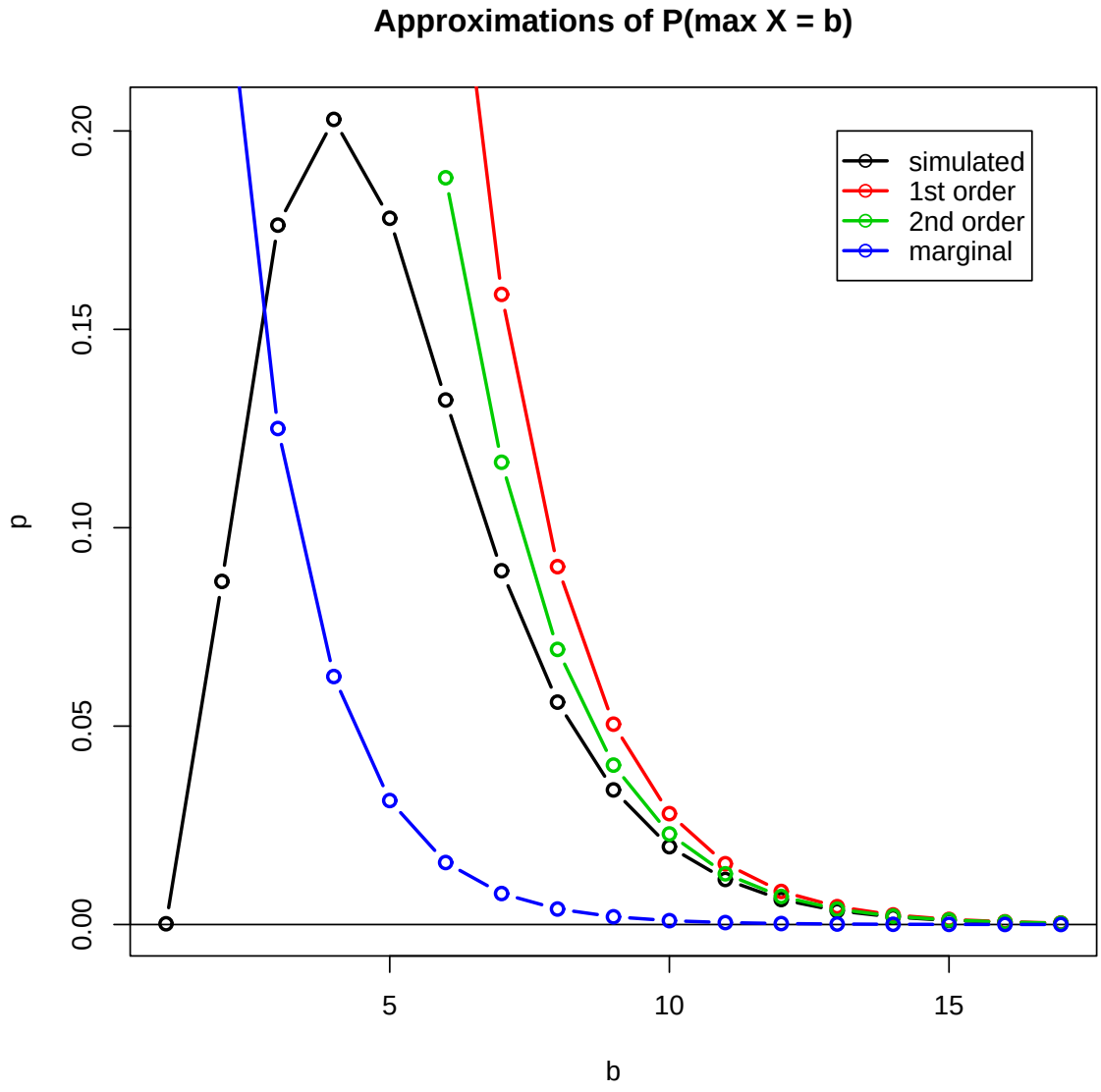
Analytic approximation (second order):

- Analyze the paths of multidimensional embedded Markov chain in the vicinity of ν and for $X_{(\nu)} \in \{1, 2, \ldots, b\}$.

- Collect all terms up to the order of $1/b$.

- The result is the second order approximation:

$$\mathbb{P}\Big(\max_{0 \le t \le L} X_t = b\Big) \approx \left[\frac{Lb}{100} + 1\right]\left[1 - \frac{g(\theta)}{b}\right] \cdot 2^{-b}.$$


Approximations of P(max X = b)
simulated
1st order
2nd order
marginal
p
0.00
0.05
0.10
0.15
0.20
5
10
15
b

Future directions

- Deal with the power and confidence sets.

- Extend the analysis to more realistic designs.

- Help in the actual mapping of complex traits in yeast.