Distinguishing causes of reduced diversity

with Alison Etheridge, Oxford

Introduction

- **Can we detect adaptive substitutions by finding regions of low diversity??**

- **Diversity can be reduced in several ways**

  - low mutation rate
  - recent common ancestry
    - chance
    - bottlenecks
    - population structure
    - selection

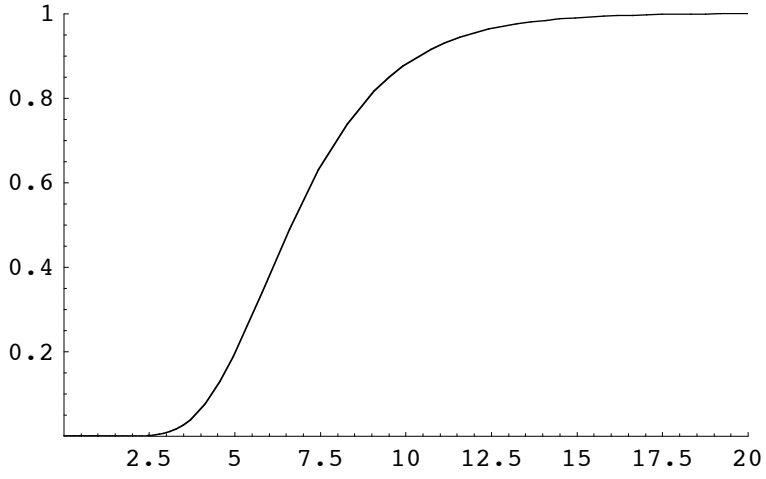
■ Severely reduced diversity is rare under the standard neutral model (SNM)

- **We expect chance variation under the standard neutral model (SNM)**

  The depth of the genealogy (t_{MRCA}) is dominated by the time taken for the last two lineages to coalesce

- **Nevertheless, extremely shallow genealogies are unlikely**

  This is the cumulative density of tree length for a sample of 20:

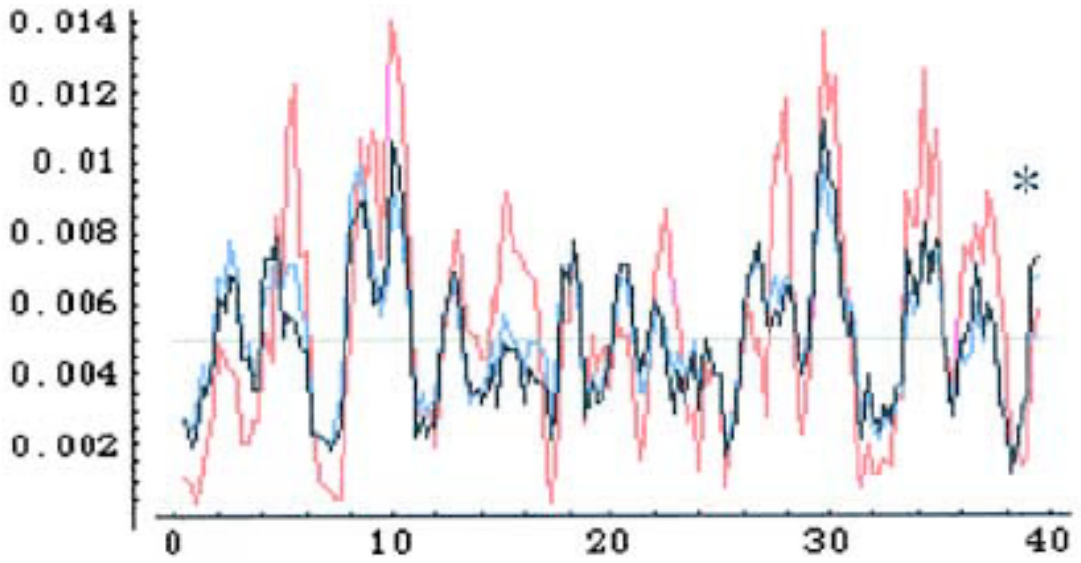


$F_n[L] = (1 - e^{-L/2})^{n-1}$

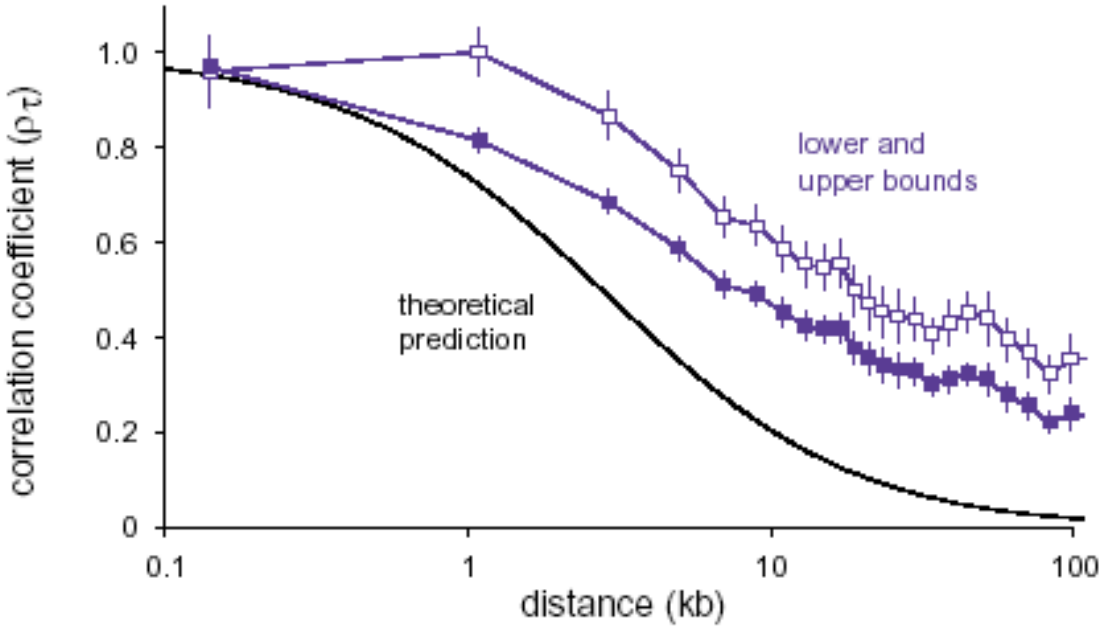
For n=20, we expect $L = 2\sum_{j=1}^{n-1} \frac{1}{j} = 7.1$ (scaling to T = t/2N)

The P= 0.001 tail is L=2.38

- **Kim & Stephan (2002) show that regions of low diversity are short and not very deep**



- **Reich et al. (2002) show that correlations in history extends over longer regions than expected under the SNM**



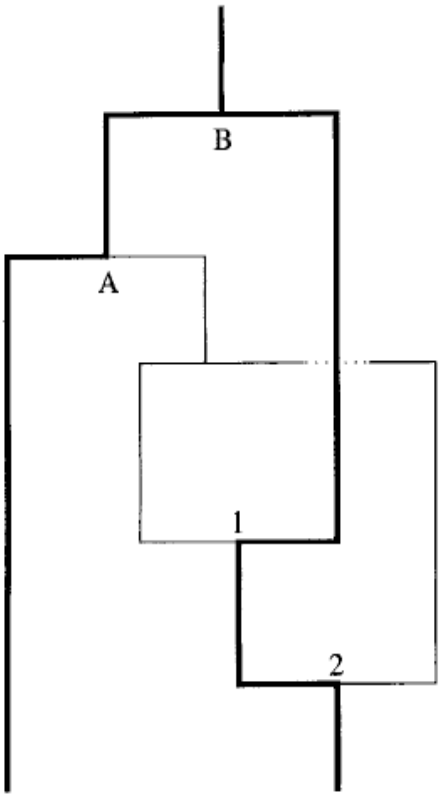
- **The SNM is not an appropriate null model:**
  - in general, diversity is far lower than expected from population size (Lewontin, 1974)

The effects of extreme events on genealogies

The coalescent with recombination is hard to analyse

Throw down a genealogy, of length L
The next recombination event along the genome is exponentially distributed, with rate L
This recombinant lineage coalesces somewhere further back....

A complication: the next genealogy along the genome depends on more than just the local genealogy (Wiuf & Hein, TPB 1999)



The genealogies for regions (0,1) and (2,3) share MRCA at A, but the intervening region has MRCA at B.
Recombinant lineages may coalesce back to lineages that are not part of the adjacent genealogy (thin lines above)

All preceding lineages must be stored, and new recombinant lineages can coalesce with any of them.
Coalescence events recur many times in the ancestral graph

■ Extreme events are easier to analyse

- Coalescence will be with distantly related lineages
- All information about the event is contained in the distribution of *family sizes*
- Concentrate on *recent* events: their causes most easily distinguished

A recent bottleneck

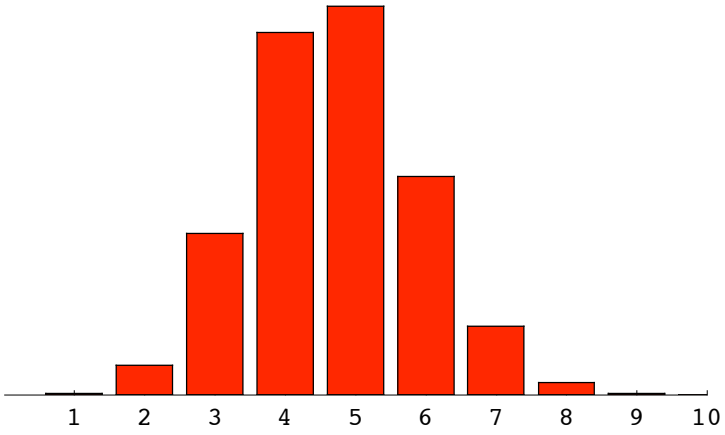
■ A mild bottleneck may by chance cause complete coalescence

The chance that k lineages coalesce down to j in T_b is:

$$p_{k,j} = \sum_{i=j}^{k} e^{-\lambda_i T_b} (-1)^j \frac{(2i-1)(j)_{(i-1)} (-k)_{(i)}}{j!\,(i-j)!\,(k)_{(i)}} \qquad (1)$$

where $a_{(i)} \equiv a(a+1)\ldots(a+i-1)$

A bottleneck of T_b =0.36, and $n = 20$ sampled genomes: P=0.001 that *all* lineages will coalesce at this bottleneck.



Little average effect:

Mean pairwise diversity ~0.7 rather than 1
Mean # of segregating sites is ~ 3.91 rather than 7.1
Mean # of ancestral lineages is ~ 4.6 rather than 20

■ Population structure has similar effects

Mean coalescence time between two nearby genes is $2N_{tot}$ regardless of structure

- **Island model**

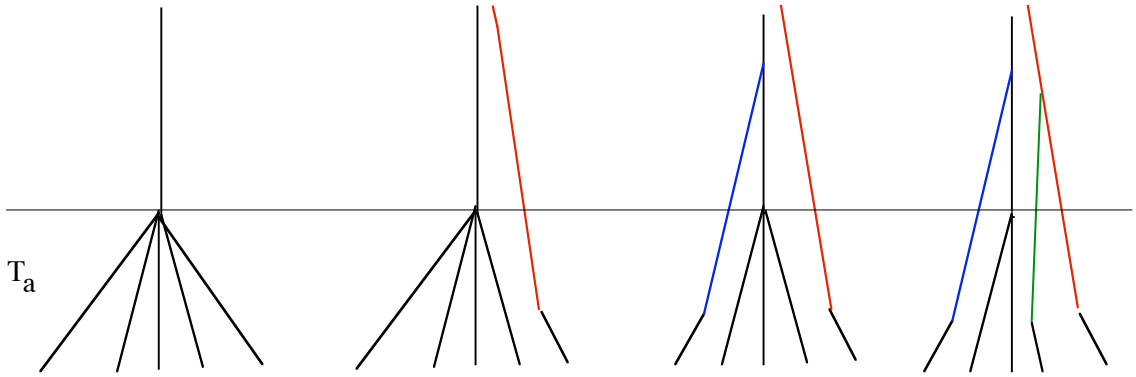
  - chance $\frac{1}{1+4\,N_e\,m}$ of coalescence within a deme
  - otherwise, E[T] ~ $2\,N_e\,d + \frac{d}{2\,m}$ with $d >> 1$ demes

- **Two dimensions**

  - density ρ, area L^2
  - chance ~ $\frac{1}{4\,\pi\rho\sigma^2}$ of coalescence within the local area
  - otherwise, E[T] ~ $2\rho\,L^2 + \frac{L^2}{2\,\pi\sigma^2}\,\log[K\,\frac{L}{\sigma}]$

Moving along the genome…

Lineages recombine away and are added one by one (more or less):

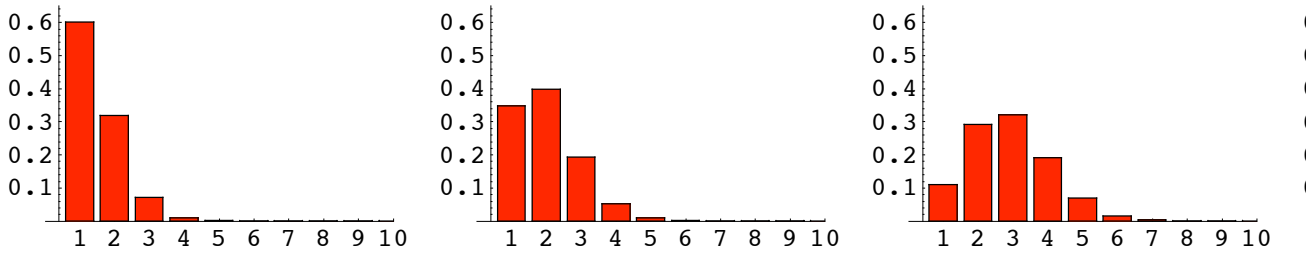


The rate of recombination is ~ $nT_a << 1$
A map distance r ~ $\frac{1}{nT_a}$ away, there will be ~k recombinant lineages, j of which pass back through the bottleneck
The probability that k→j, given that n-k→1 is:

$$2^{-(n-1)}\,(n-k)!\left(\sum_{y=j+1}^{n}\frac{e^{-\lambda_y\,T_a}}{\prod_{y\neq z}(\lambda_z-\lambda_y)}\right)\left(2^j\,\frac{k!}{j!}\,\frac{(2n-k-1)!}{(j+1)!}\,\frac{(k-j+1)_{n-m}}{(j+2)_{n-1}\,(n+1)_{n-1}}\right) \qquad (2)$$

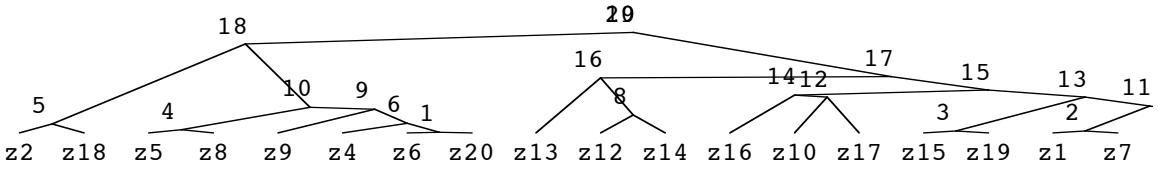
where $\lambda_y = y\,(y-1)/2$, $a_n = a\,(a+1)\ldots(a+n-1)$

The distribution of # of ancestral lineages for $T_b = 0.36$ $rT_a = 0.25,\ 0.5,\ 1,\ 2$

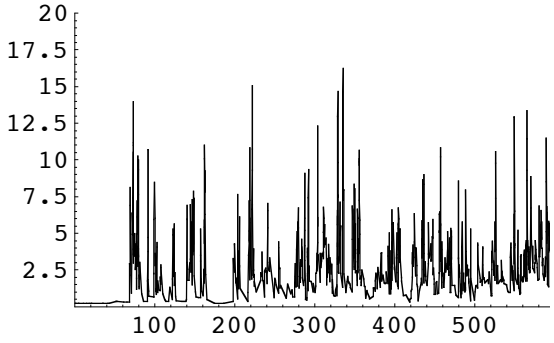


Immediately after a bottleneck ($T_a << 1$), diversity is reduced over a long region ($R \sim 1/T_a$)
Genealogies *prior* to the bottleneck are therefore independent (because $R >> 1$)

- **Genealogies along the chromosome**

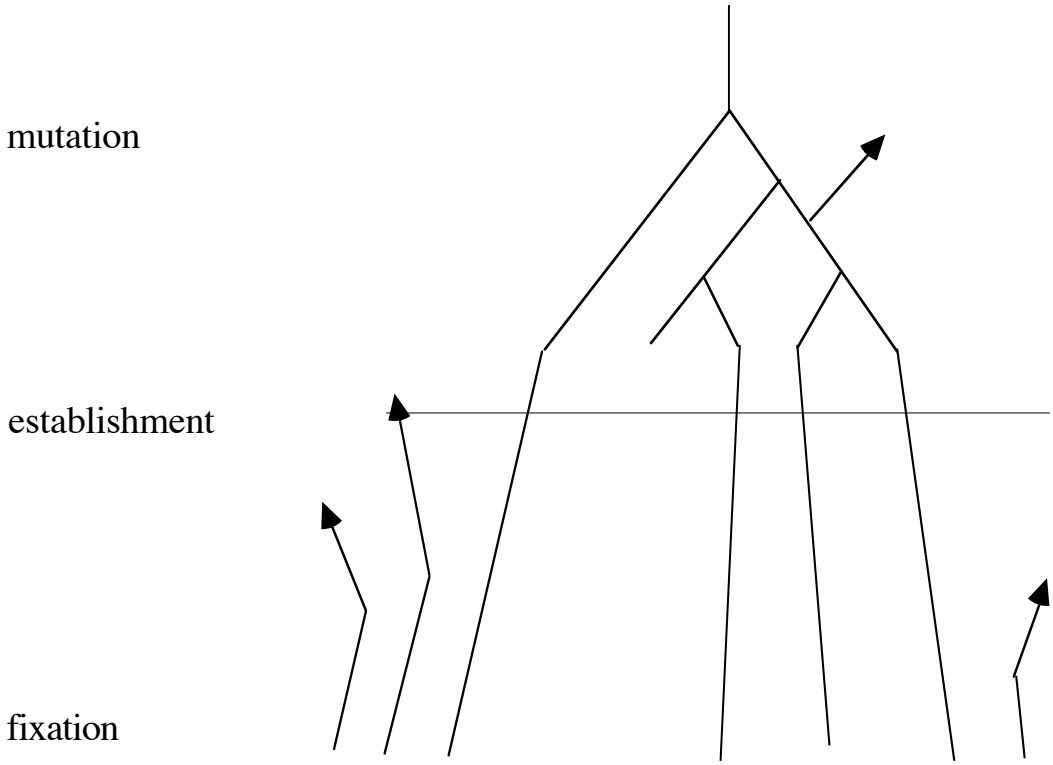


This shows the tree length along the genome, allowing for the bottleneck:



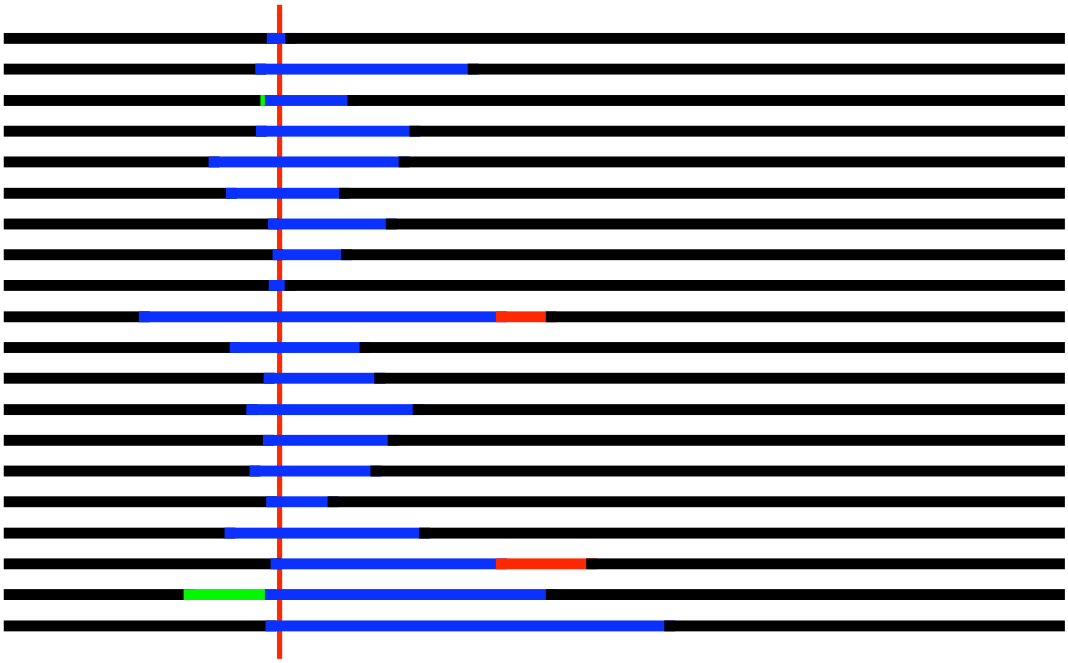
Selective sweeps

At the selected locus, the entire population traces back to one favourable mutation
The pattern of coalescence along the genome is determined by the *branching process* that establishes this mutation



The chance that a lineage traces back to one of k^* copies is $(\frac{k^*}{2N})^{r/s} \frac{\Gamma[1+r/s]}{\Gamma[1-r/s]}$
Coalescence occurs within the branching process, and is likely to lead back to the one ancestral mutation

- **Along the genome…**



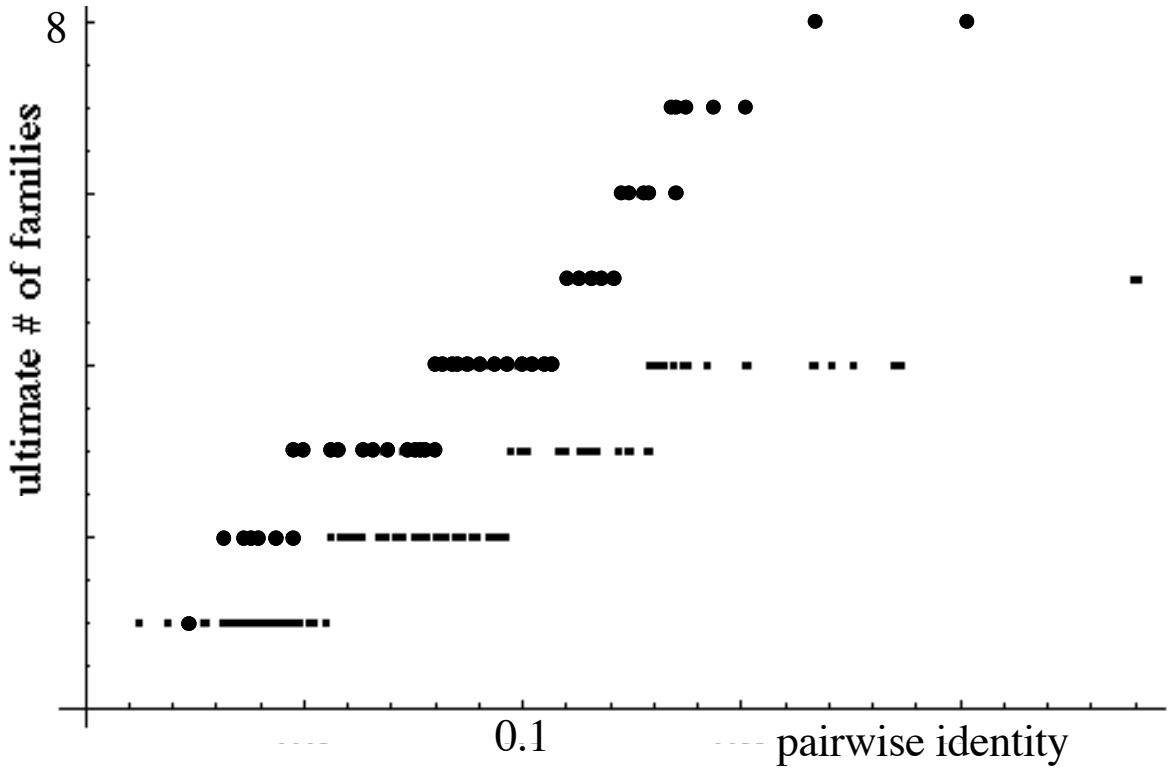
To a good approximation, coalescence is into one family
Breakpoints are independently distributed with rate ~ s/log(2N)

More precisely, we need the distribution of times T_k for which the branching process has k copies;
This leads to the distribution of # of lineages extant when the first recombination event occurs

As for a bottleneck, unrelated lineages are added one-by-one as we move along the genome

- **Ultimate # of families vs pairwise identity.**

circles: $r = 0.01$, $s = 0.1$, $N = 10^6$; dots: r=0.00453, s=0.1, $N = 10^{12}$.



How can we distinguish bottlenecks from selective sweeps?

- **Genealogical patterns are indistinguishable in practice**

- **Bottlenecks are expected to have the same effect across the whole genome**

  BUT - bottlenecks have a high variance in effect
  - if selective sweeps are common, their effects will not be distinguishable

- **A bottleneck will reduce diversity at the same time at all loci (Galtier et al. 2000)**

  BUT - there may be multiple bottlenecks
  - reduced diversity may be due to population structure

- ***A priori* differences between different parts of the genome**
  Lower diversity in regions of reduced recombination in *Drosophila melanogaster*
  Candidate loci: self-incompatibility loci, G6PD…