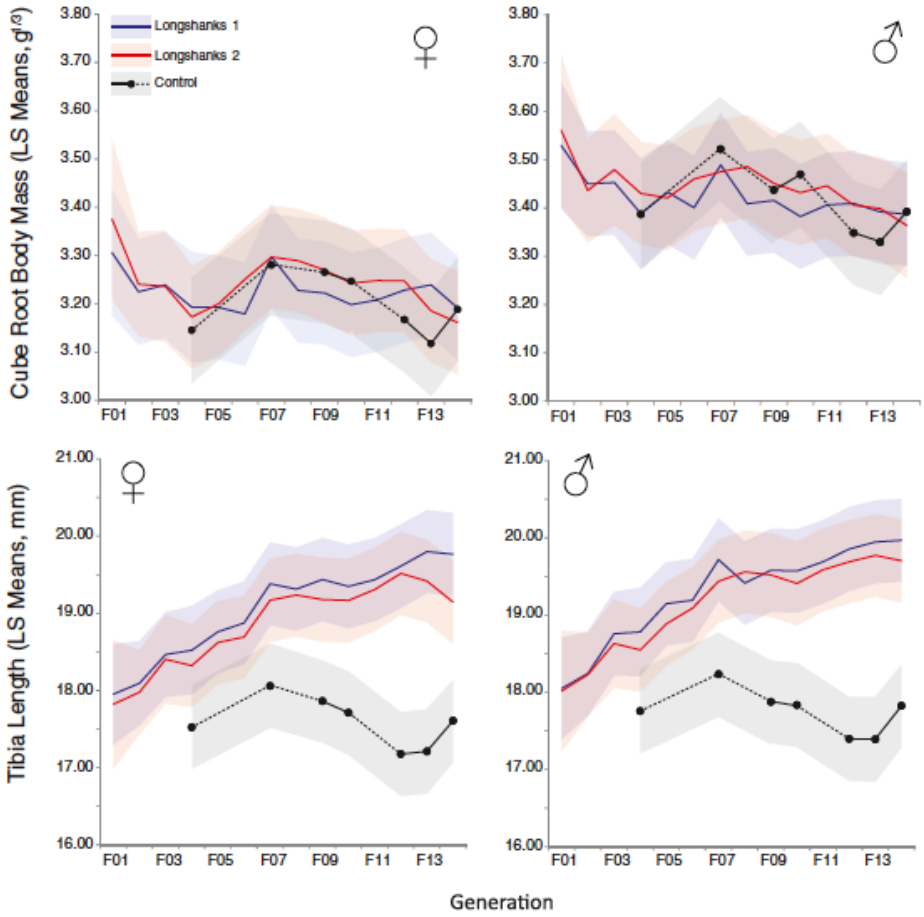
How best to distinguish selection on discrete loci from the infinitesimal model?

Nick Barton

Longshanks

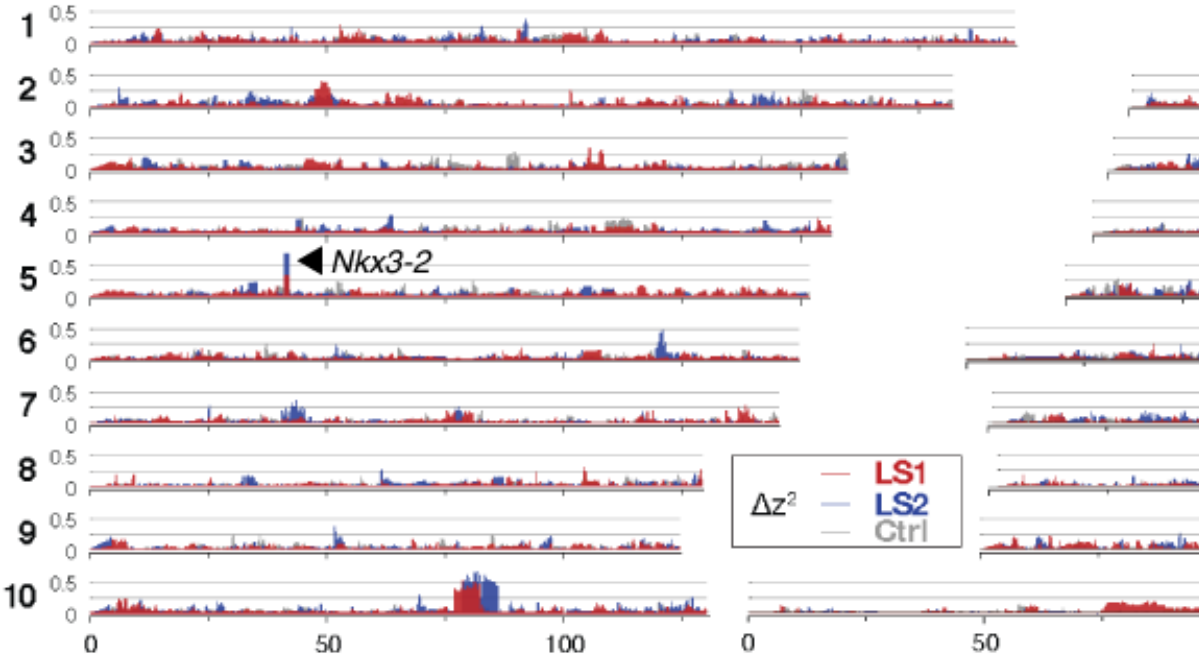
Frank Chan, Layla Hiramitsu (Tübingen); Campbell Rolian (Calgary); Stefanie Belohlavy (IST); bioRxiv

Two replicates of ~30 mice: within-family selection for the longest tibia



Use a composite trait $\log[TM^{-0.57}]$

Some 10kb windows show strong allele frequency change:
z=2 arcsin$\left(\sqrt{p}\right)$; Δz^2 in 10kb windows



Motivation

There is a rapid, consistent response to selection

We know the selection, the pedigree, the sequence …

Can we find the causal alleles ?

A small experiment - but it represents larger populations, selected for a longer time.

Outline

The infinitesimal as the null model

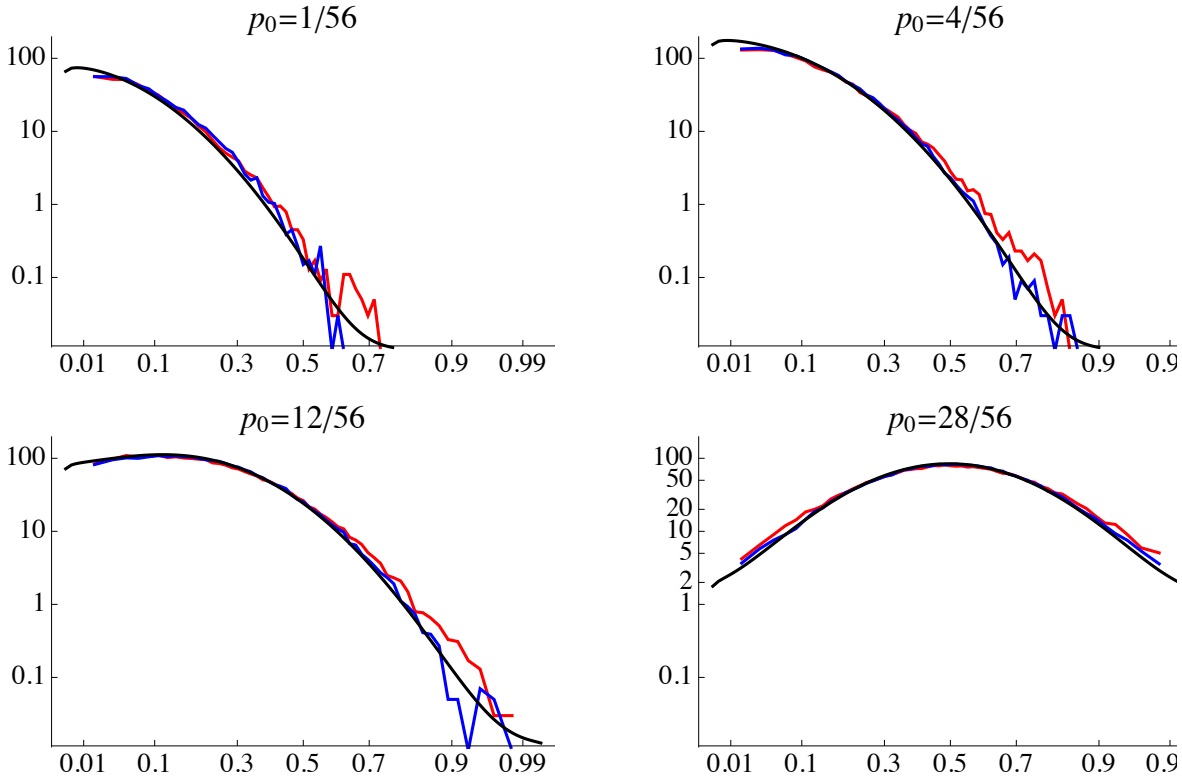
Variation in SNP and haplotype frequencies

Estimating effects of candidate loci on fitness and trait

The infinitesimal with linkage

In this experiment, the pedigree is fixed, and so chromosomes evolve *independently*

How much does infinitesimal selection affect allele frequencies?



Even strong selection has little effect

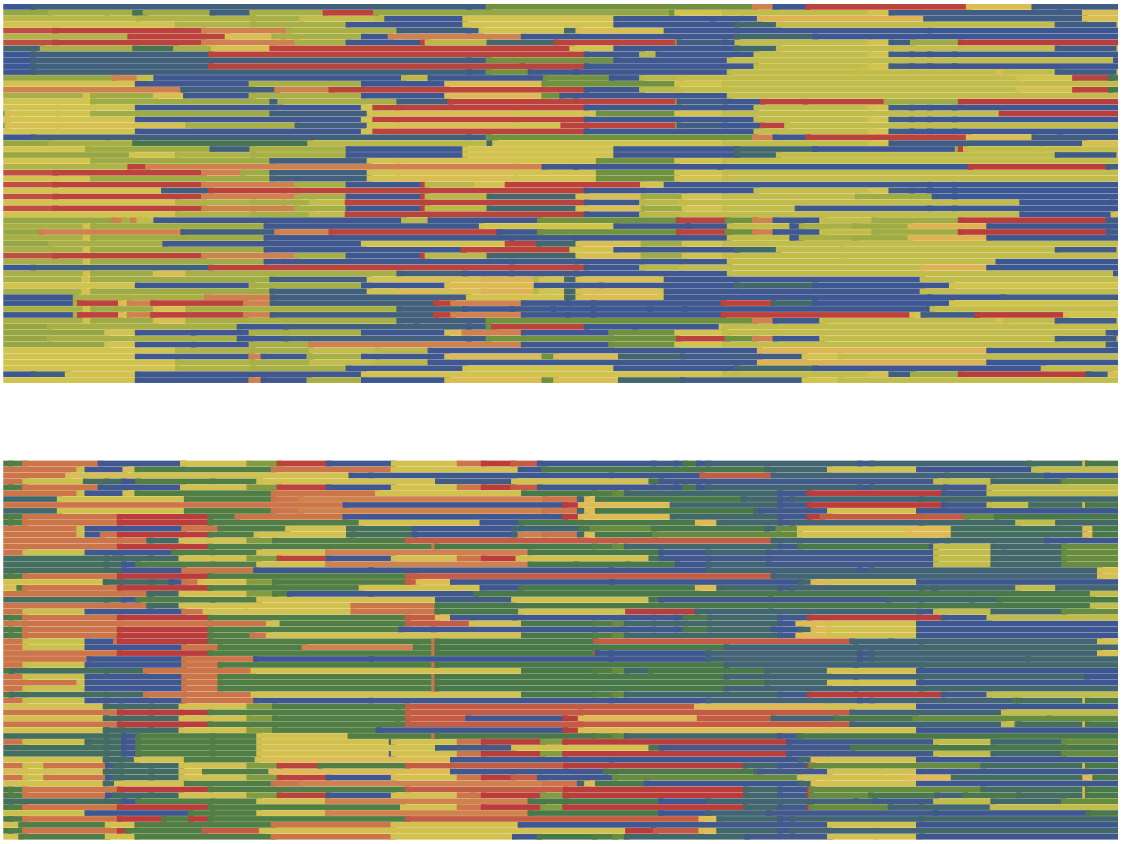
The diffusion approximation works well

Infinitesimal selection produces a slight excess of sweeps

SNP are carried on haplotype blocks

Simulate, conditioning on the pedigree, and the observed heritability (assuming additivity)

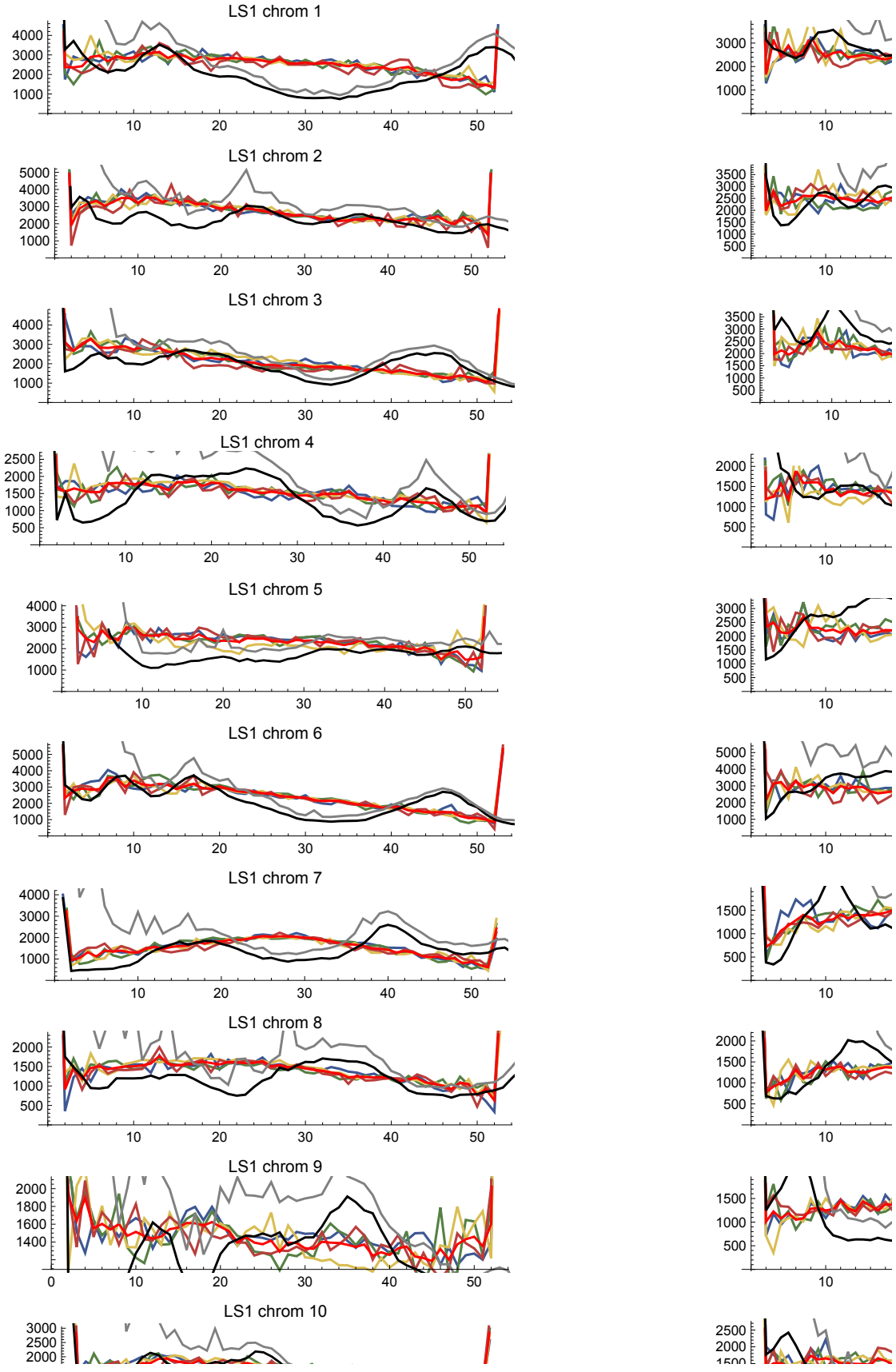
Out[ ]=

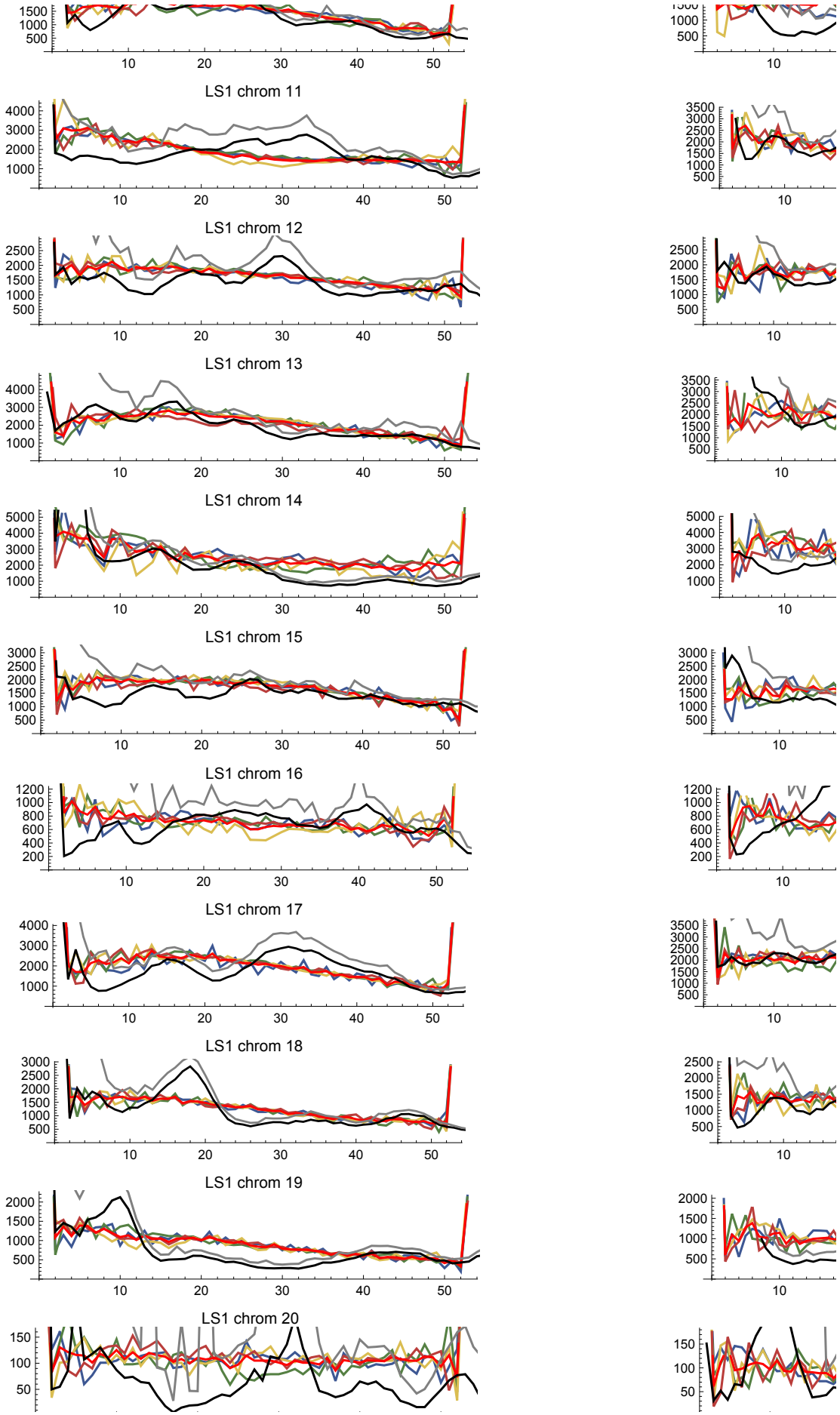


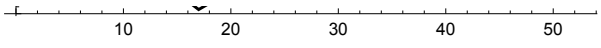
SNP are thrown down onto the haplotype blocks

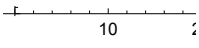
Variance in SNP frequency is inflated

(grey/black: old/new data; colours: replicate simulations)




LS1 chrom 11
LS1 chrom 12
LS1 chrom 13
LS1 chrom 14
LS1 chrom 15
LS1 chrom 16
LS1 chrom 17
LS1 chrom 18
LS1 chrom 19
LS1 chrom 20


10
20
30
40
50


10

Variation in SNP frequency is inflated by LD in the base population

In any window, we have k_i of the i'th haplotype: Variation in SNP frequencies reflect k_i

$$< \Delta p^2 > = \frac{1}{n_S} \sum_{i=1}^{n_s} \Delta p_i^2 \tag{1}$$

$$\mathbb{E}\left[< \Delta p^2 > \right] = \frac{j}{n_0^2} \, \text{var}[k] \left(1 - \frac{j_i - 1}{n_0 - 1} \right) \tag{2}$$

where $\frac{j}{n_0}$ is the initial SNP frequency

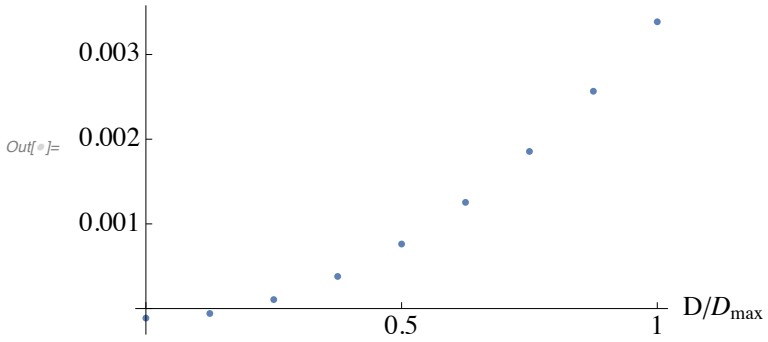
$$\text{var}\left[< \Delta p^2 > \right] = \frac{1}{n_S^2} \left(\sum_{i=1}^{n_s} \text{var}\left[\Delta p_i^2 \right] + \sum_{\substack{i,j=1 \\ i \neq j}}^{n_s} \text{cov}\left[\Delta p_i^2, \Delta p_j^2 \right] \right) \tag{3}$$

var$\left[< \Delta p^2 > \right]$ depends on the moments of k_i and is inflated by LD in the base population.

With large # of SNP, var$\left[< \Delta p^2 > \right]$ ~ cov$\left[\Delta p_i^2, \Delta p_j^2\right]$, which increases with D^2.

e.g. $n_0 = 32$, $k_i = \{10, 4, 2, 1, 1, 1, 1, 0, 0, \ldots\}$, $p_0 = 0.5$:

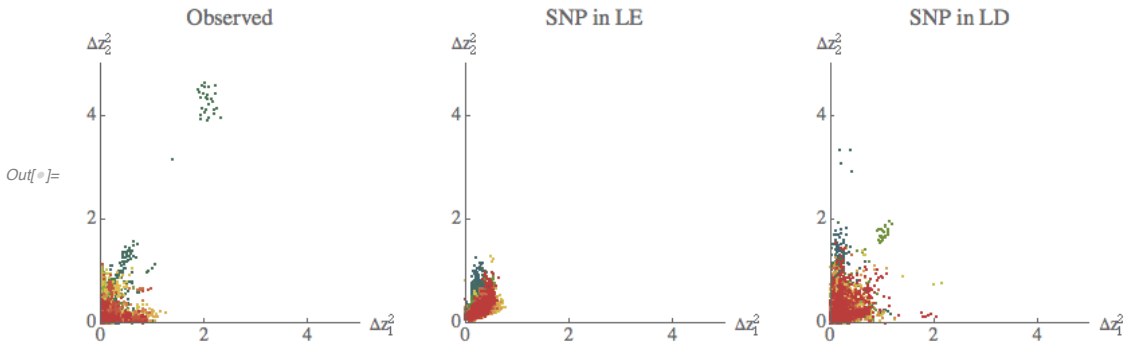
Out[•]=



Is the candidate on chrom. 5 significant?

Is the candidate on chrom. 5 significant?

Out[ ]=



Is the candidate on chrom. 5 significant?

Pairs of simulations, starting from the same founder genomes, give outlier Δz^2 that overlap the signal from LS1 (red) but not LS2 (orange)
Based on SNP frequencies, the signal is marginally significant

Three sources of variation in SNP frequencies

- effects of founders
- evolution of replicates
- random SNP on haplotypes

Variation due to LD amongst SNP can be strong:

- coalescent simulations of a well-mixed population
- Kelly & Hughes: *D. simulans*

This source of error can be *eliminated* by working with haplotypes

- haplotypes can be reconstructed from SNP frequencies (Kessner et al., 2013, Franssen et al., 2016)

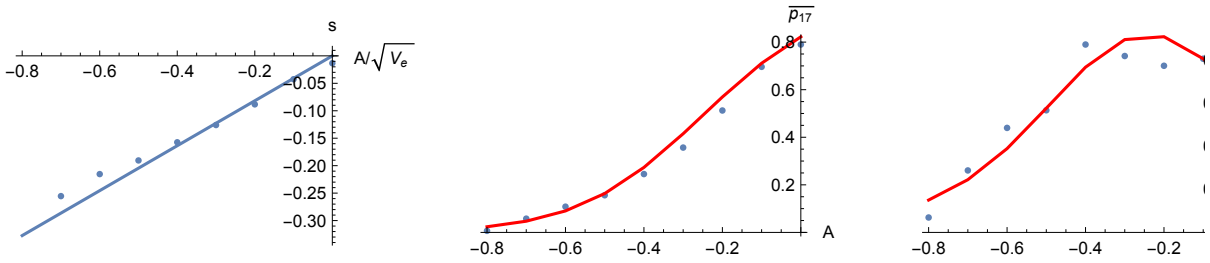
How strong is selection ?

Alleles in the candidate region on chrom. 5 sweep from p= 0.178 → 0.833, 0.981 in LS1, LS2 ⇒ s ~ $\frac{1}{t}\log\left[\frac{p_{17}}{q_{17}}\frac{q_0}{p_0}\right]$ ~ 0.25 (cf. Taus et al., 2017)

How large an effect on the trait? Simulate an additive allele, effect A; 40 replicates; $s = 0.41\,A/\sqrt{V_e}$ (left)

The mean and sd from infinitesimal simulations (dots) fit with a single-locus WF model, N_e ~ 44(red)

Out[•]=



The locus on chromosome 5 has effect $\hat{A} = 0.59\sqrt{V_e}$ ($0.32\sqrt{V_e}$ to $-0.87\sqrt{V_e}$). This single locus is responsible for ~ 9.4% (3.6% - 15.5%) of the response .

Summary

- The infinitesimal should be used as the null model
- In Longshanks, even strong infinitesimal selection has little effect
  (but: selection was within families; the map is long)
- Substantial variation is generated by random assignment of SNP to haplotypes
  - especially with LD in the base population
- Even an obvious signal is marginally significant in any one line
- How many loci contribute to the selection response ?