What can we learn from experimental evolution in sexual populations?

Oregon State University OSU

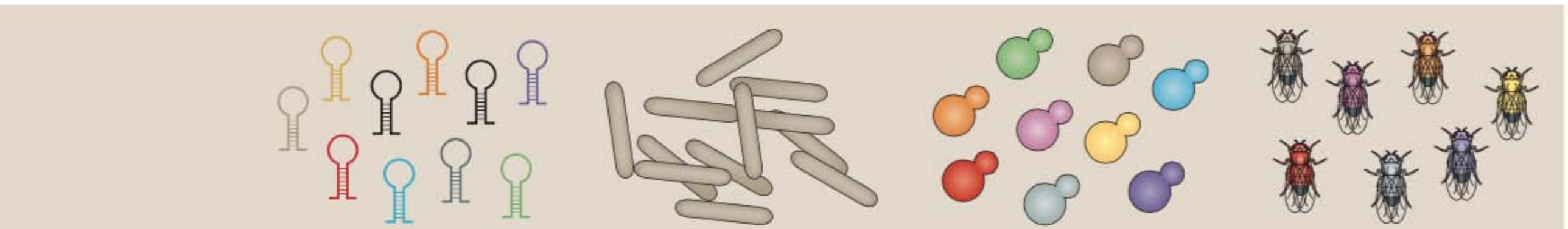
Molly K. Burke
Assistant professor
Dept of Integrative Biology

experimental evolution allows us to test many questions…

- how repeatable is evolution?
- how well can we localize/identify QTL?
- what are the origins and fates of adaptive alleles?

Elucidating the molecular architecture of adaptation via evolve and resequence experiments

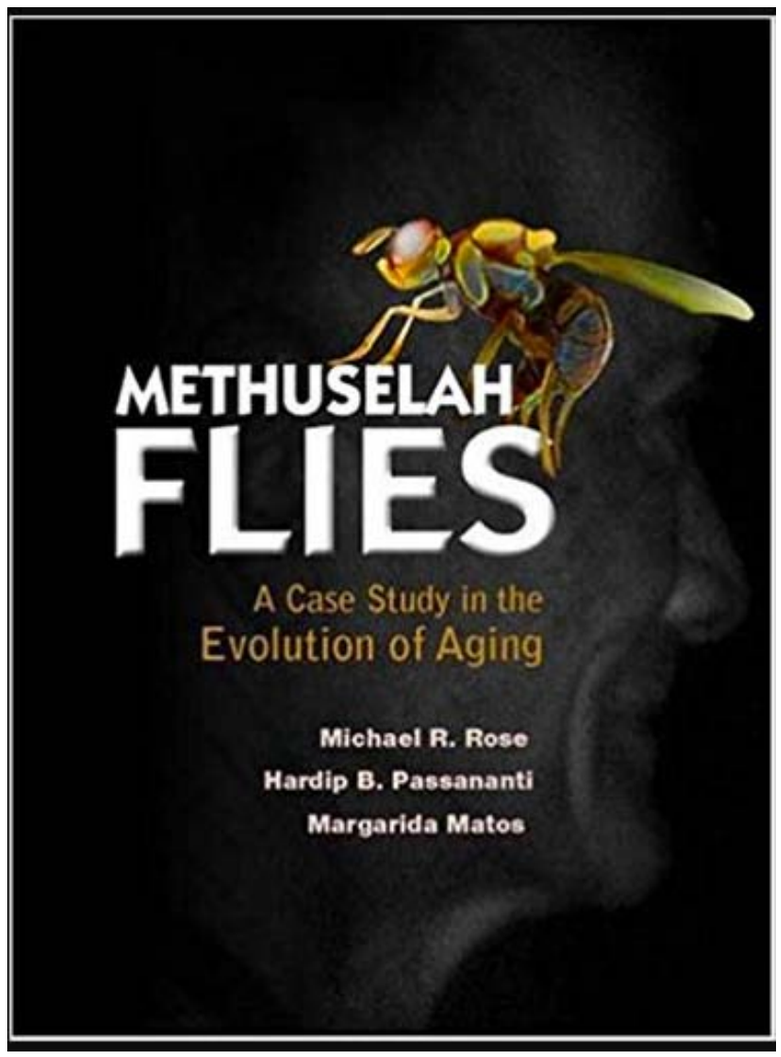
Anthony Long[1], Gianni Liti[2], Andrej Luptak[3] and Olivier Tenaillon[4]



	In vitro	Microbial isogenic	Microbial outbred	Obligate sexual higher eukaryotes
Example models	Synthetic DNA or RNA molecules	Bacteria, haploid yeast or diploid yeast	Diploid yeast	*Drosophila melanogaster*

"The time is right for practitioners in the different systems to learn from one another."

Drosophila melanogaster



selected treatment:

ACO

only 9-day old adults reproduce

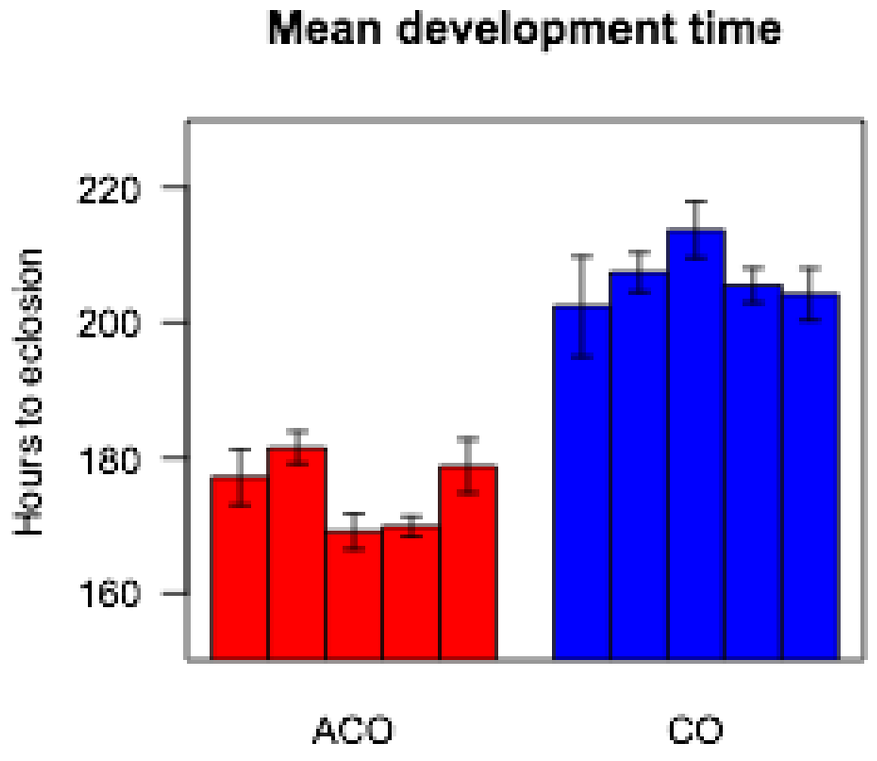
600 generations by 2009

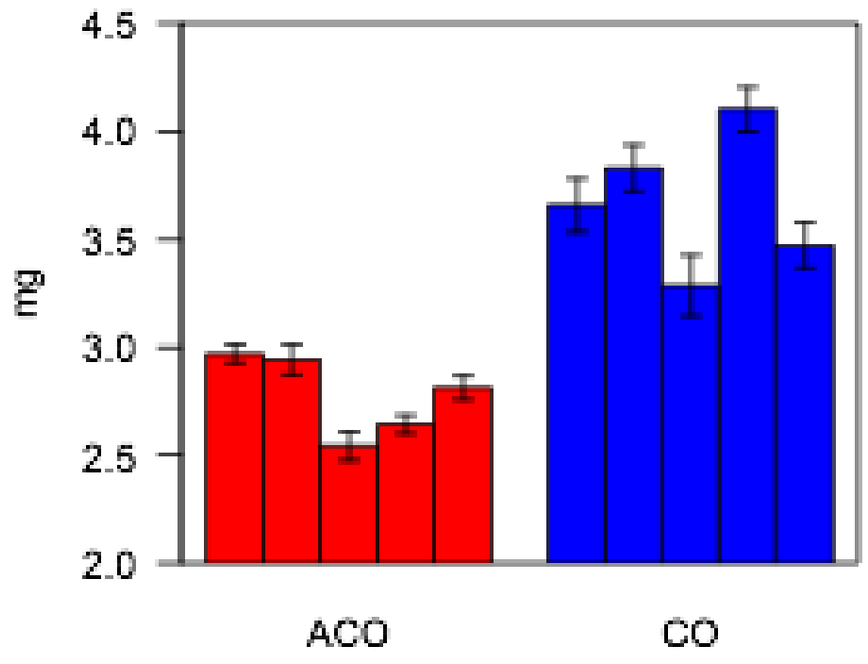
control treatment:

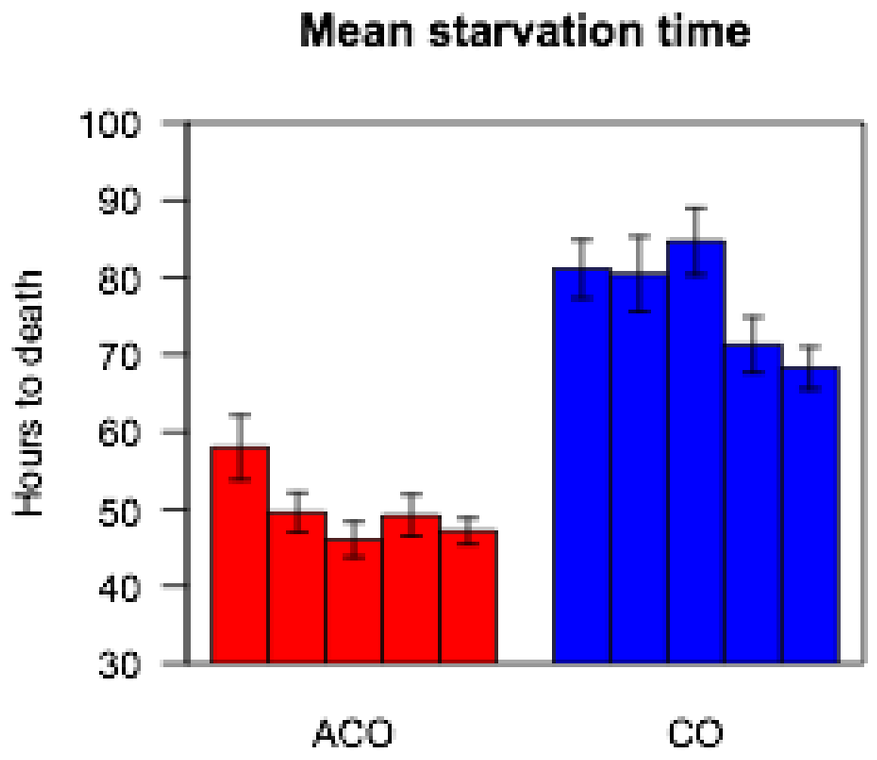
CO

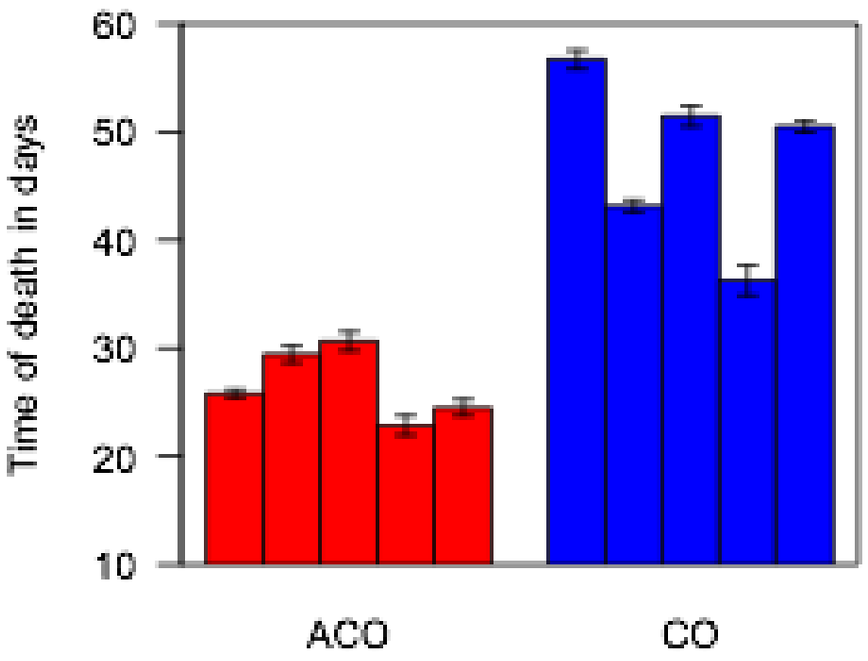
only 28-day old adults reproduce

250 generations by 2009



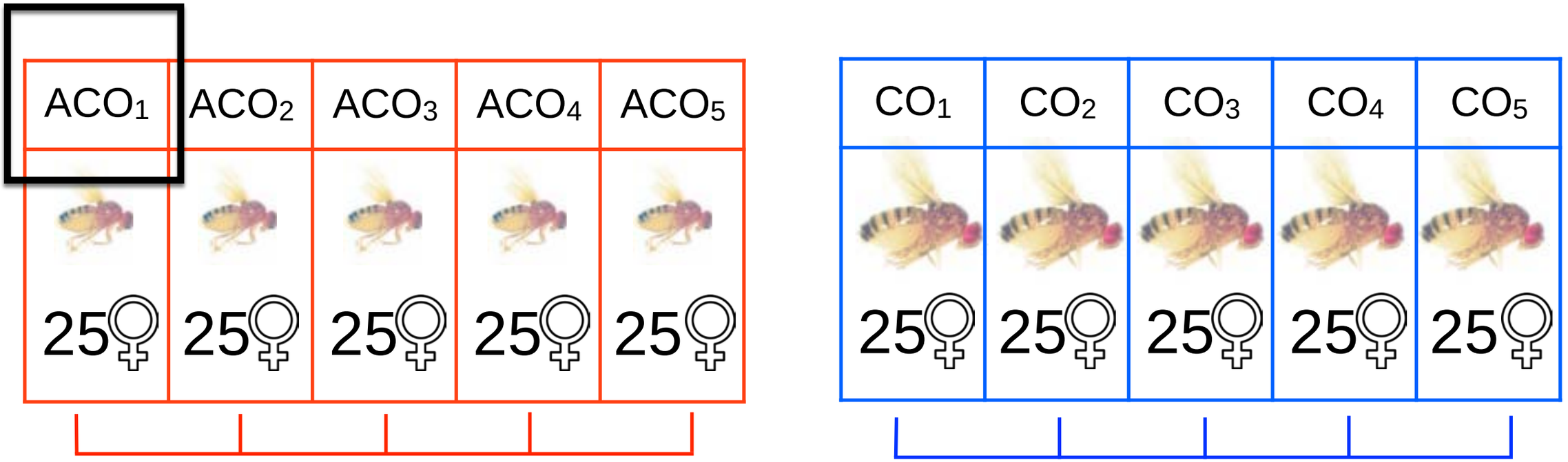






Burke *et al.* 2010 **Nature**

Pooled genome sequencing



ACO library

CO library

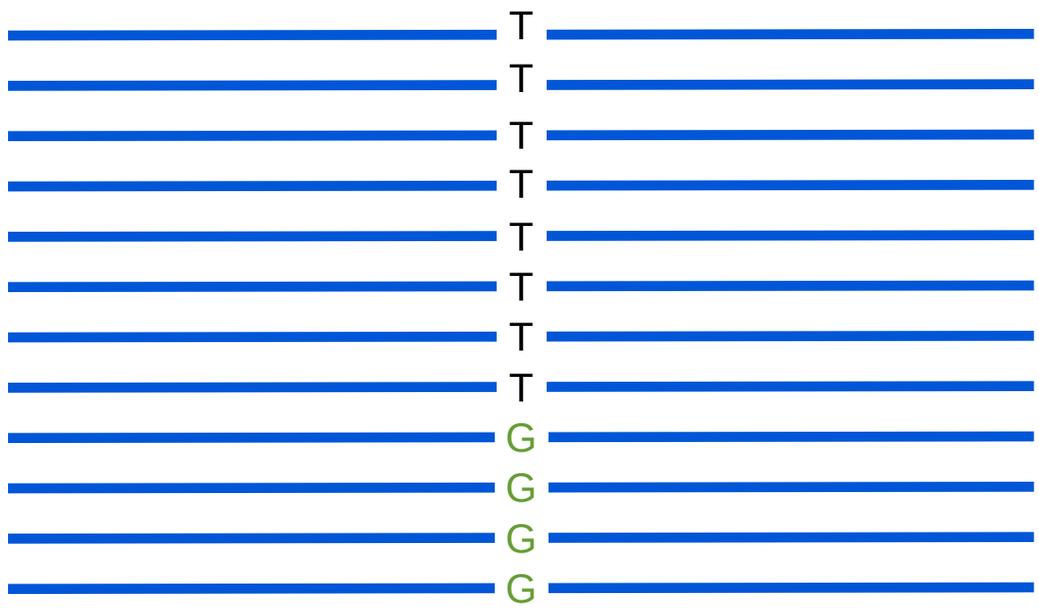
third library sequenced: ACO_1

Pool-SEQ

Drosophila reference genome

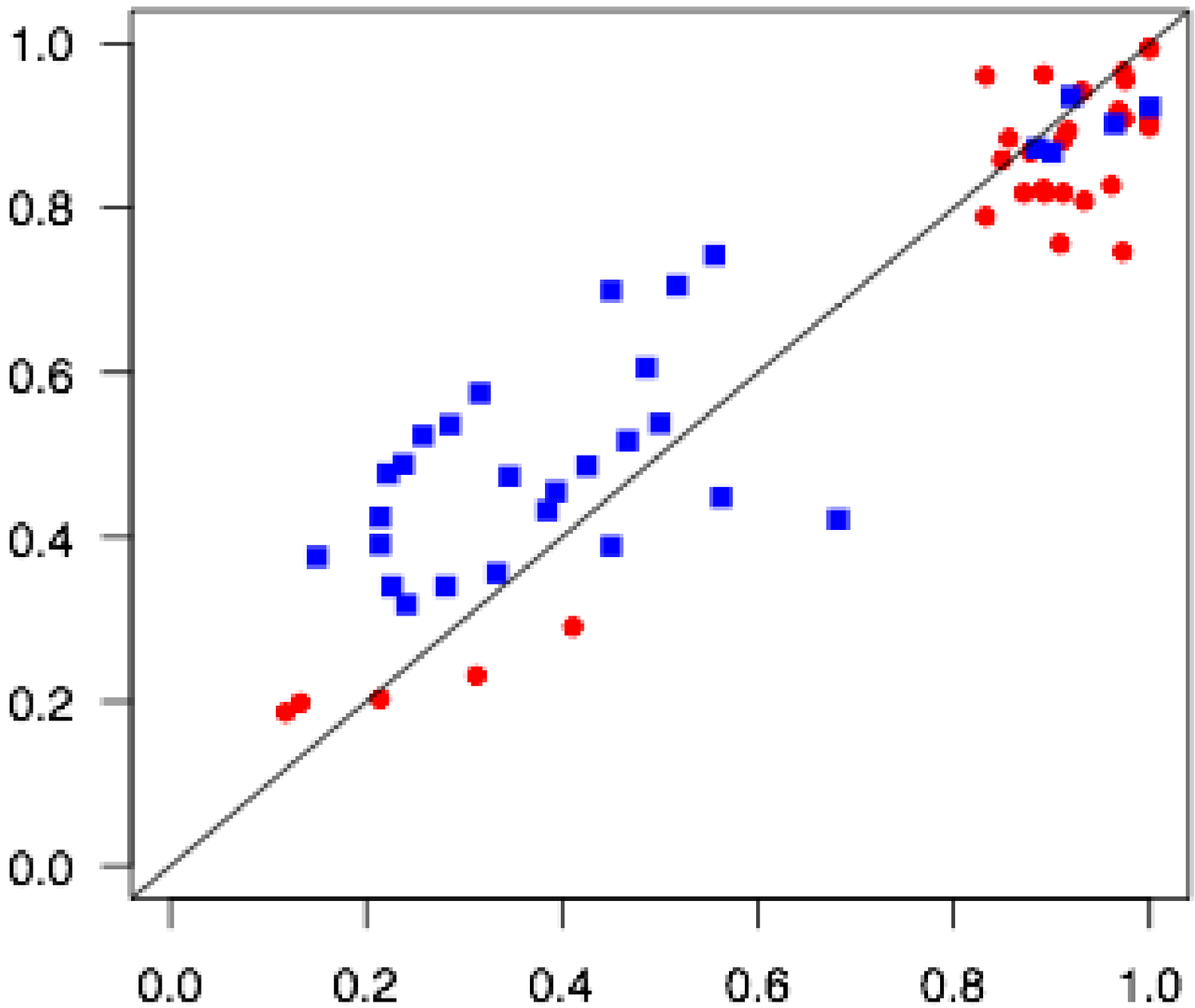
ACO sequences

CO sequences



utility of pooled samples for Pool-SEQ

allele frequency from individual genotypes



APPLICATIONS OF NEXT-GENERATION SEQUENCING

Sequencing pools of individuals — mining genome-wide polymorphism data without big funding

Christian Schlötterer[1], Raymond Tobler[1,2], Robert Kofler[1] and Viola Nolte[1]

Abstract | The analysis of polymorphism data is becoming increasingly important as a complementary tool to classical genetic analyses. Nevertheless, despite plunging sequencing costs, genomic sequencing of individuals at the population scale is still restricted to a few model species. Whole-genome sequencing of pools of individuals (Pool-seq) provides a cost-effective alternative to sequencing individuals separately. With the availability of custom-tailored software tools, Pool-seq is being increasingly used for population genomic research on both model and non-model organisms. In this Review, we not only demonstrate the breadth of questions that are being addressed by Pool-seq but also discuss its limitations and provide guidelines for users.

Next-generation sequencing
(NGS; also known as second generation sequencing) An umbrella term for different sequencing platforms delivering millions of short DNA sequence reads

Reads
DNA sequences that are generated by next-generation sequencing.

About a decade ago, a fully sequenced genome was big news. But now, owing to rapid advances in next-generation sequencing (NGS) technology and computer algorithms for assembling short reads, we are enjoying the availability of an ever increasing number of genomes from a broad spectrum of non-model organisms[1]. In parallel with the growing catalogue of reference genomes, a variety of approaches have emerged that seek to characterize the genome-wide polymorphism patterns. Arguably, the most comprehensive polymorphism data so far have been generated by single-nucleotide polymorphism (SNP) microarrays in humans[2–4]. More recently, the field has begun moving towards the characterization of full genome sequences, with 1000 genome projects completed for humans[4], *Arabidopsis thaliana*[5] and cattle[6]. In *Drosophila melanogaster* too, hundreds of genomes have already been sequenced, and other species, such as pigs and dogs, are catching up. Does this imply that we have now captured all of the relevant variation and that, despite some minor bits and pieces of data remaining to be filled, in essence we are close to what we need to understand variation in these species?

Probably the best demonstration that this is not the case comes again from human genetics. The analysis of human diseases and other complex traits indicated that even the analysis of several thousand individuals frequently turned out to be insufficient to determine the underlying genetic architecture[7,8]. Given this scale, it is clear that many research questions cannot be addressed by whole-genome sequencing of individuals, even though the sequencing costs of a human genome have now decreased below the 'magic line' of US$1,000 (REF. 9).

In this Review, we discuss whole-genome sequencing of pools of individuals (Pool-seq) — an approach that provides genome-wide polymorphism data at considerably lower costs than sequencing of individuals. We explain why Pool-seq is more cost-effective, compare it to other approaches, review dedicated software tools, and discuss limitations and further directions. On the basis of various intraspecific whole-genome Pool-seq studies, we demonstrate its versatility and efficacy in facilitating a broad range of genome-wide analyses. However, we do not cover the metagenomic analysis of pools consisting of multiple species, as this has been reviewed elsewhere[10].

The cost-effectiveness of Pool-seq

Key to population genetic surveys is information about polymorphic positions in the genome and the frequencies of variant alleles in various populations. The power of many genetic analyses increases with the accuracy to which allele frequencies can be determined from population samples. Pool-seq provides more accurate allele frequency estimation at a lower cost than sequencing of individuals[11,12]. To understand the basis of this difference, it is important to remember that allele frequencies are typically estimated from samples drawn from a larger population. Smaller sample sizes

[1]*Institut für Populationsgenetik, Vetmeduni Vienna, Veterinärplatz 1, 1210 Vienna, Austria.*
[2]*Vienna Graduate School of Population Genetics*
Correspondence to C.S.
e-mail: christian.schloetterer@vetmeduni.ac.at
doi:10.1038/nrg3803
Published online
23 September 2014

best practices:

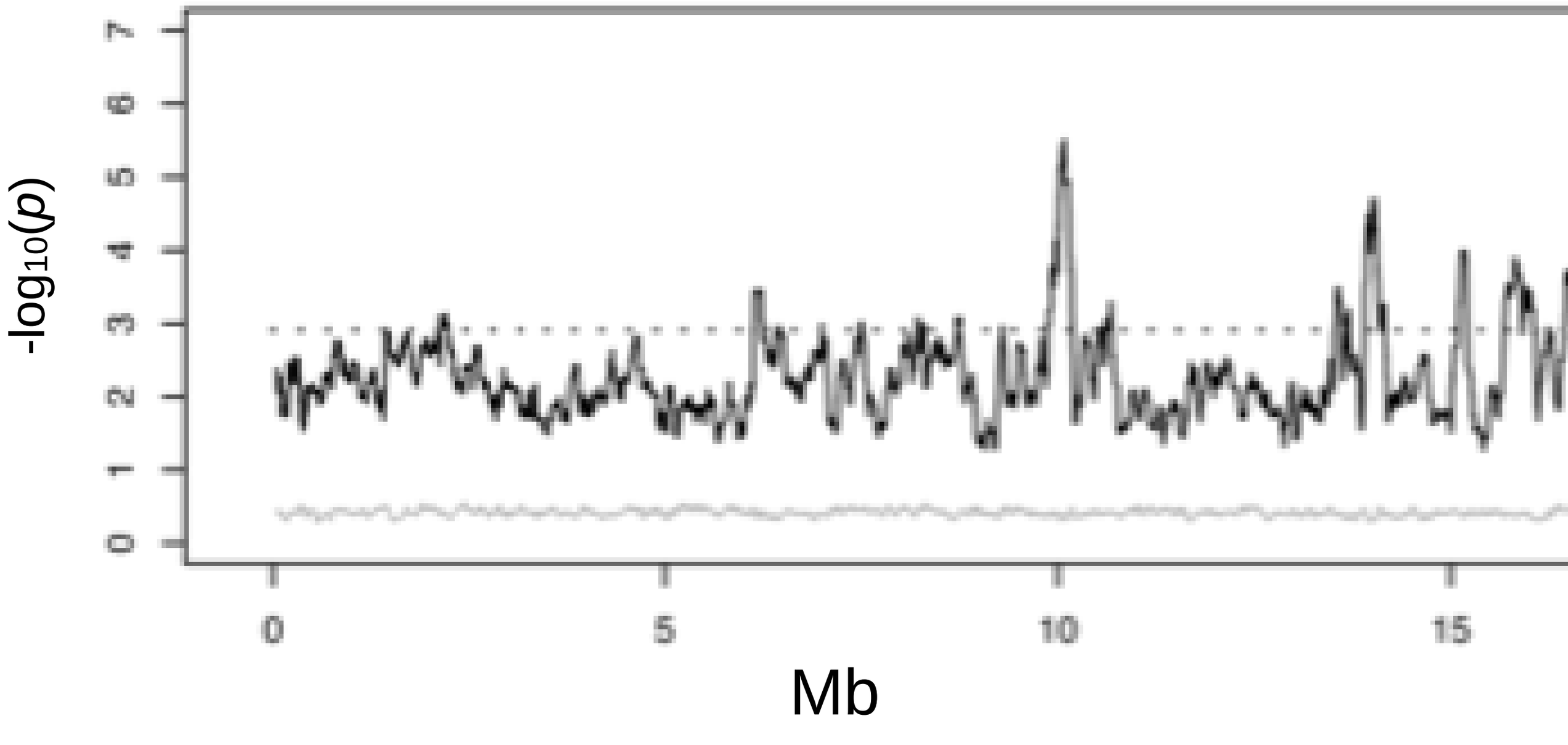
pool > 40 individuals

coverage > 50X

read lengths > 75bp

black line = frequency differences between ACO and CO treatments

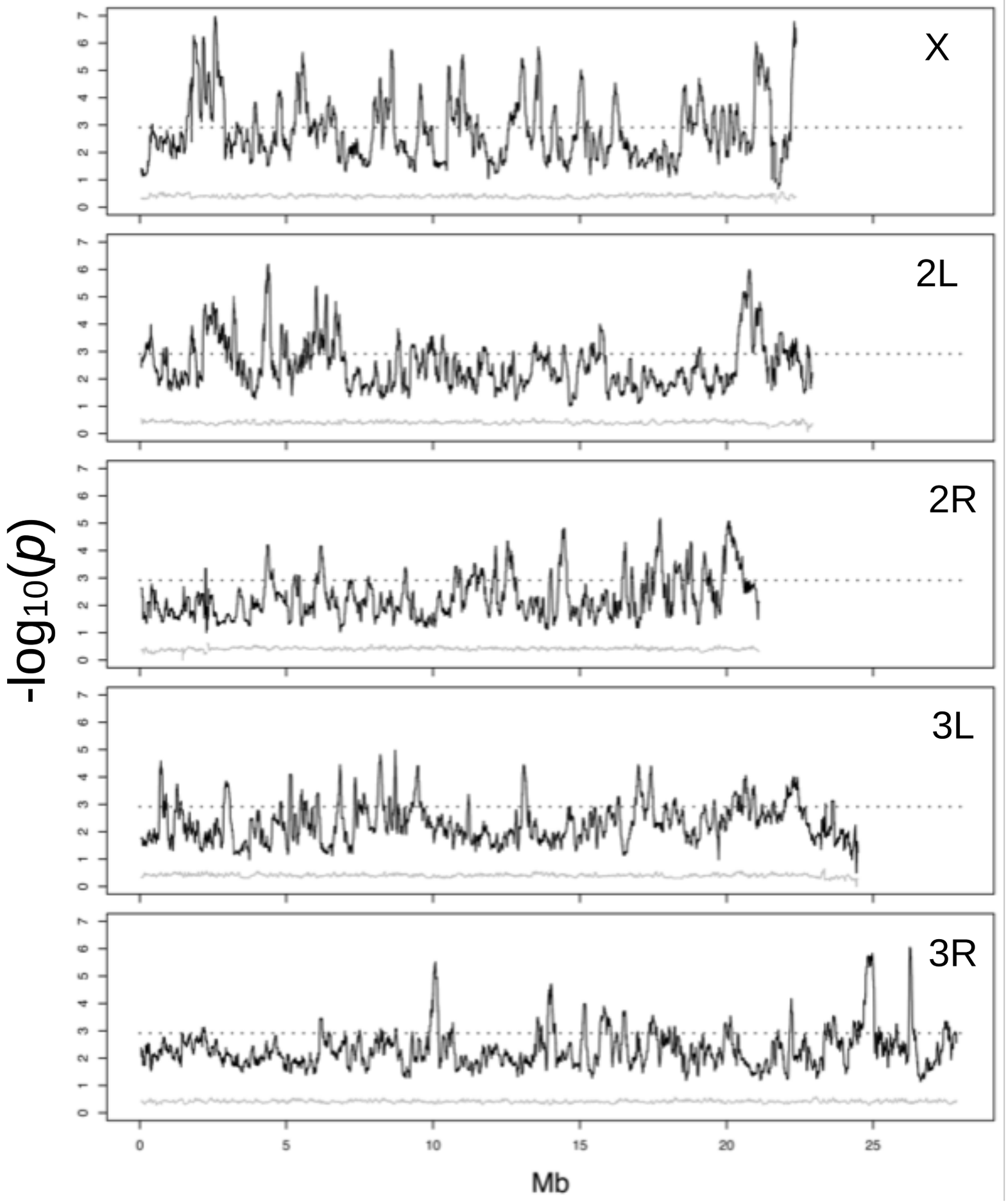
gray line = frequency differences between ACO_1 and the entire ACO pool



lots of differentiation between the ACO and CO pools

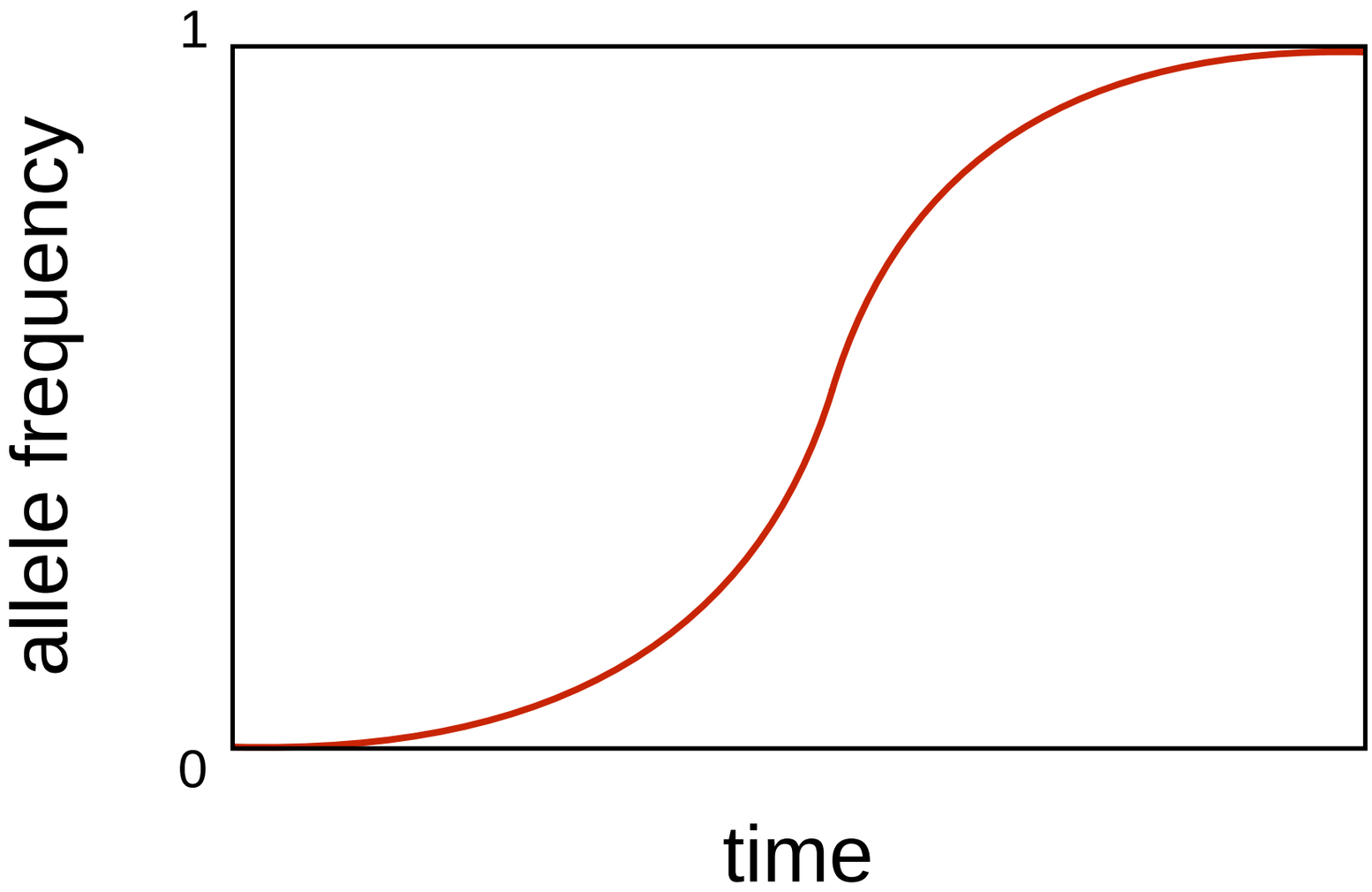
no differentiation between ACO_1 and the ACO pool

~500 genes under peaks



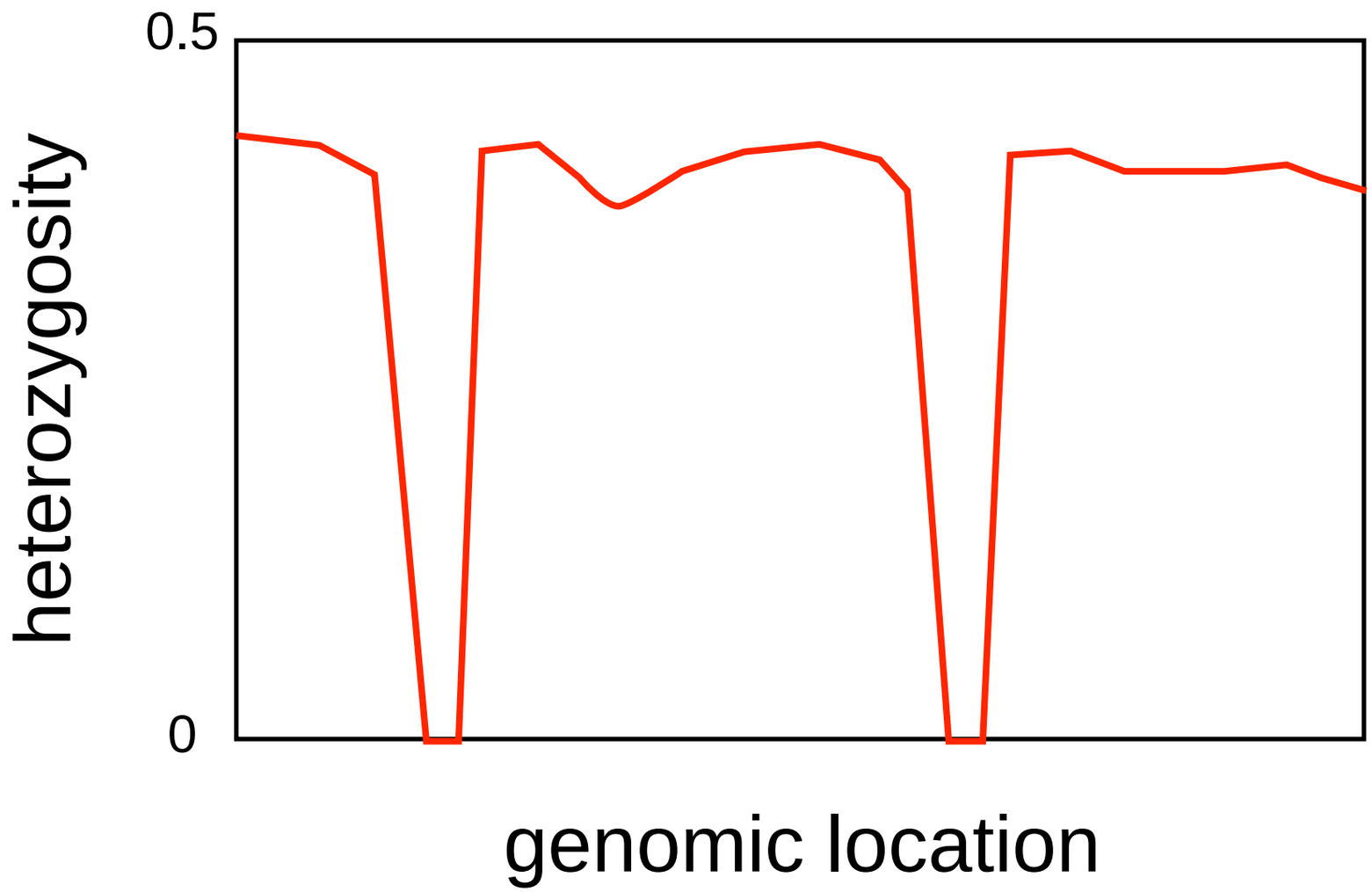
classic selective sweep

a beneficial mutation arises, natural selection increases the frequency of this allele until fixation



classic selective sweep

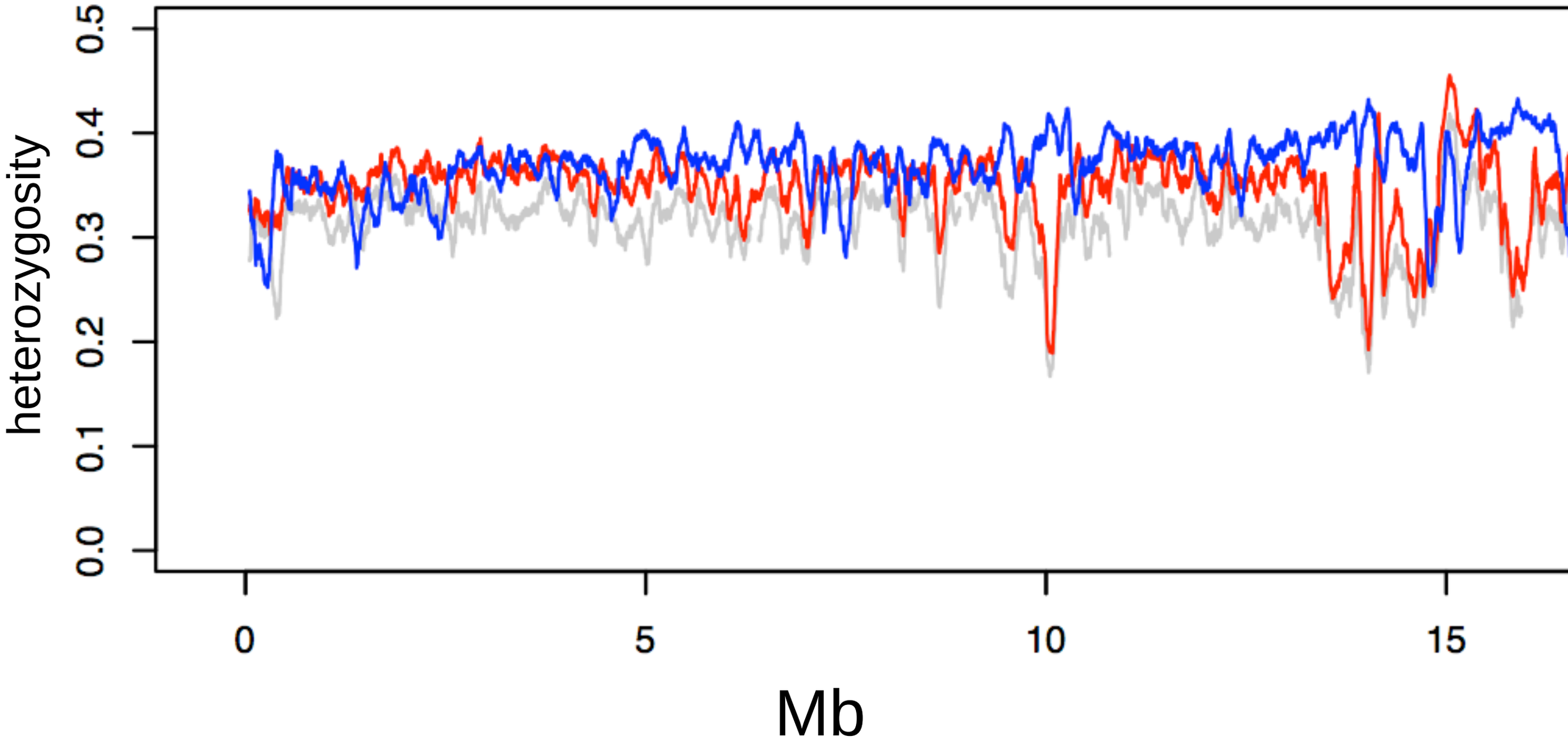
at sweep locations, heterozygosity losses should occur at selected and linked sites



blue = CO pool

red = ACO pool

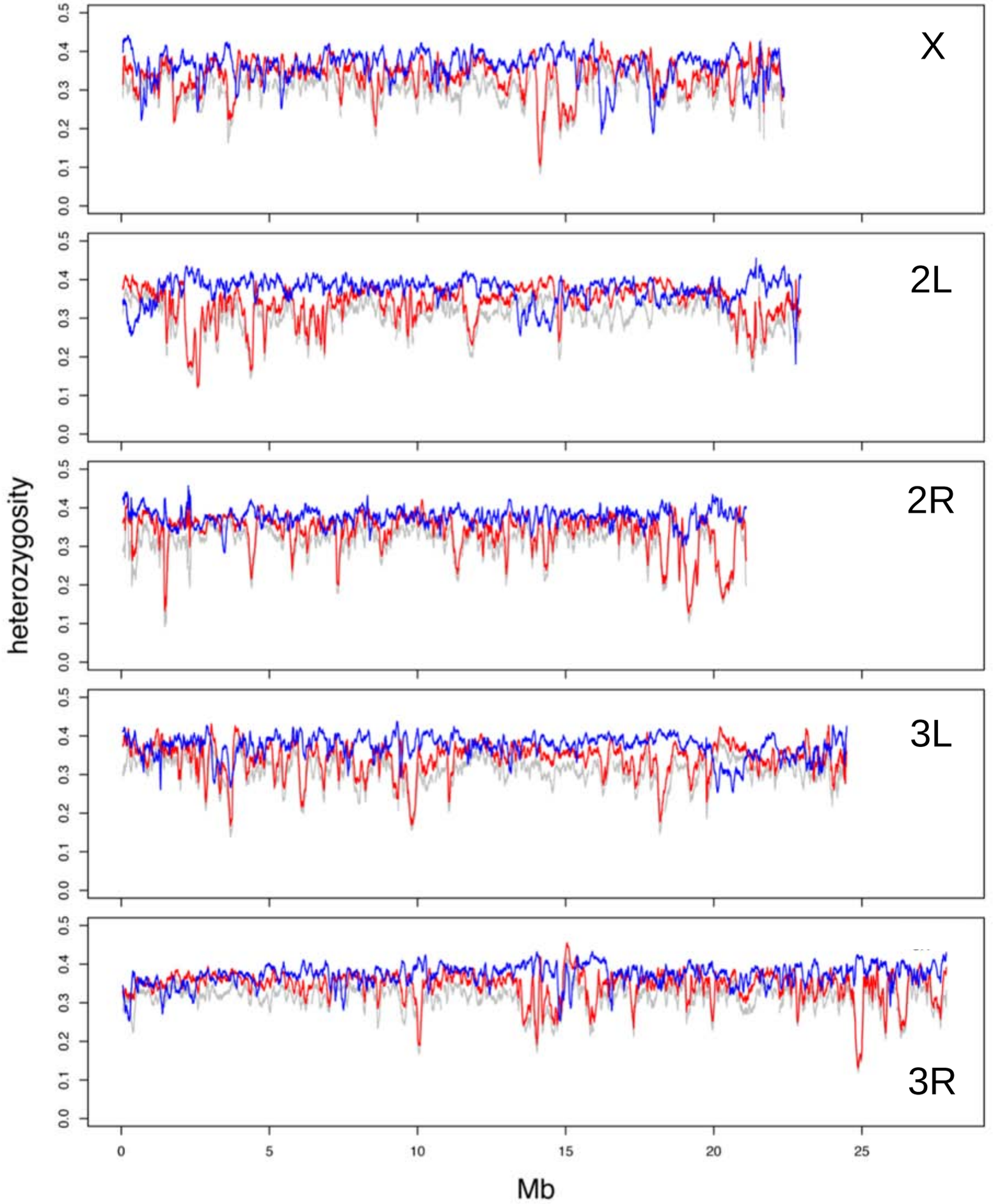
gray = ACO_1



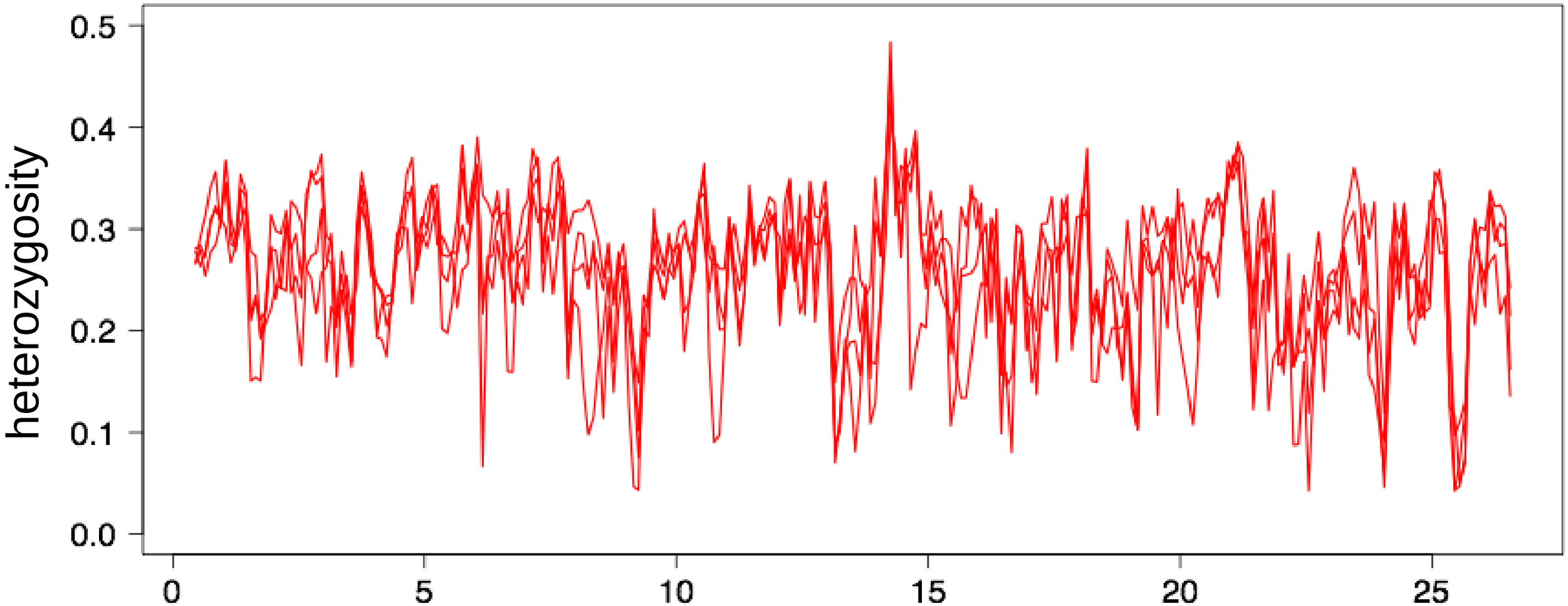
more losses of heterozygosity in ACO

local losses correspond to differentiated regions

ACO_1 heterozygosity resembles ACO pool

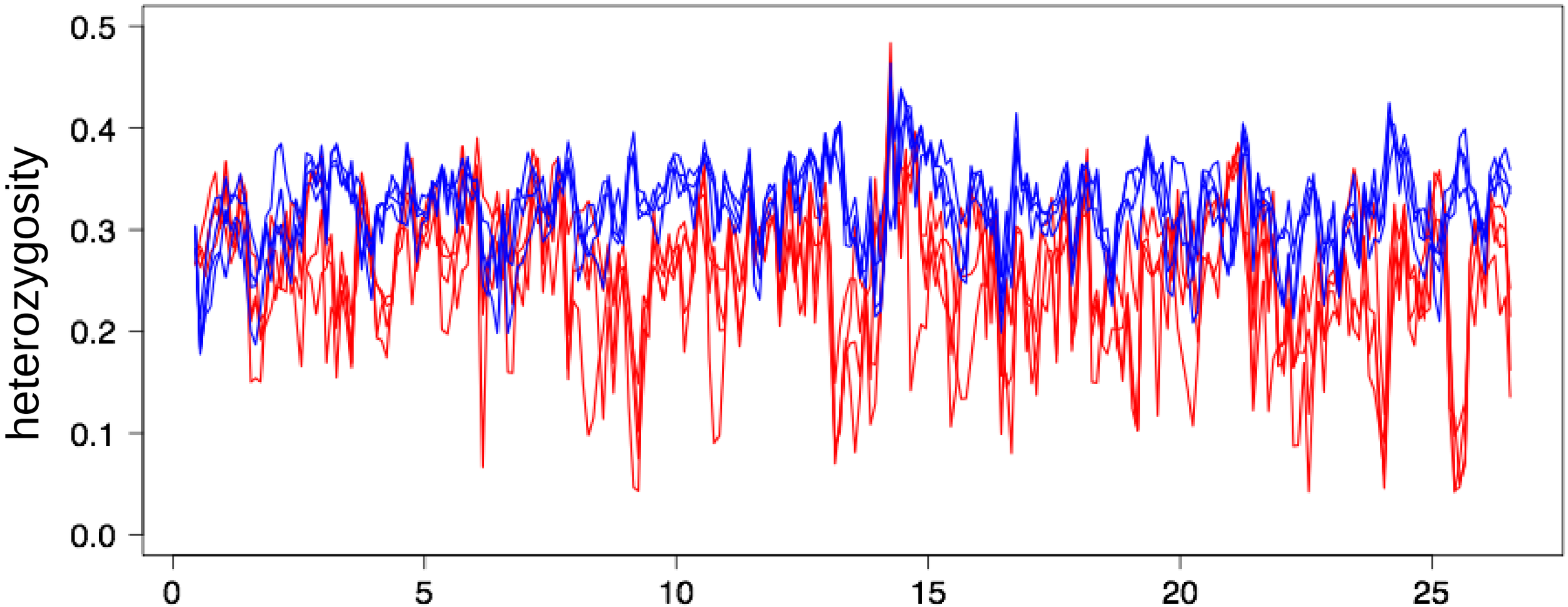


~50X coverage per replicate



data from Graves *et al.* 2016 **MBE**

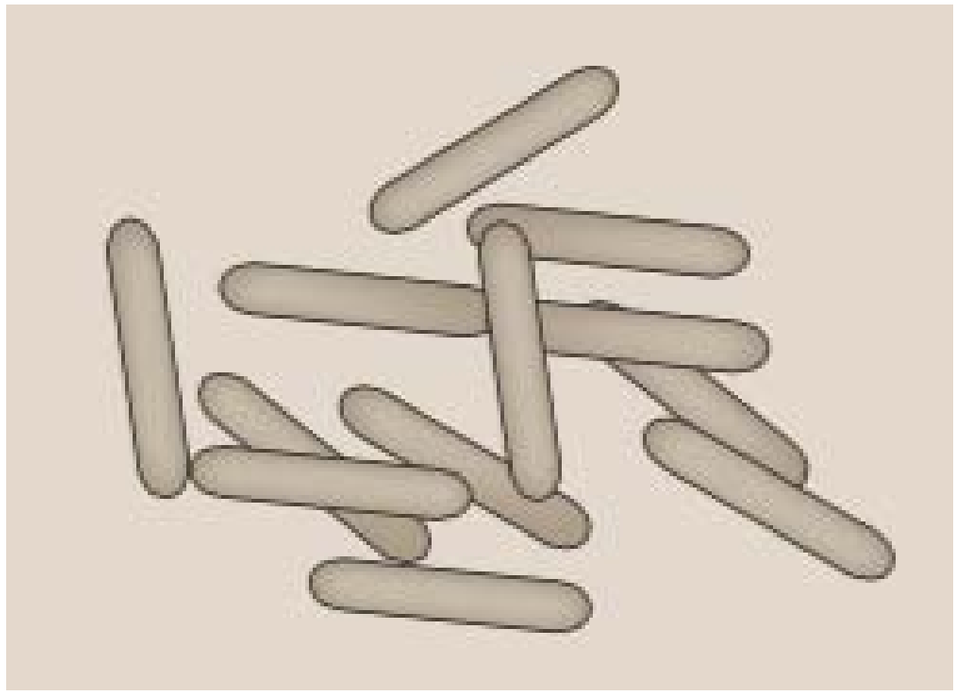
~50X coverage per replicate



data from Graves *et al.* 2016 **MBE**

conclusions

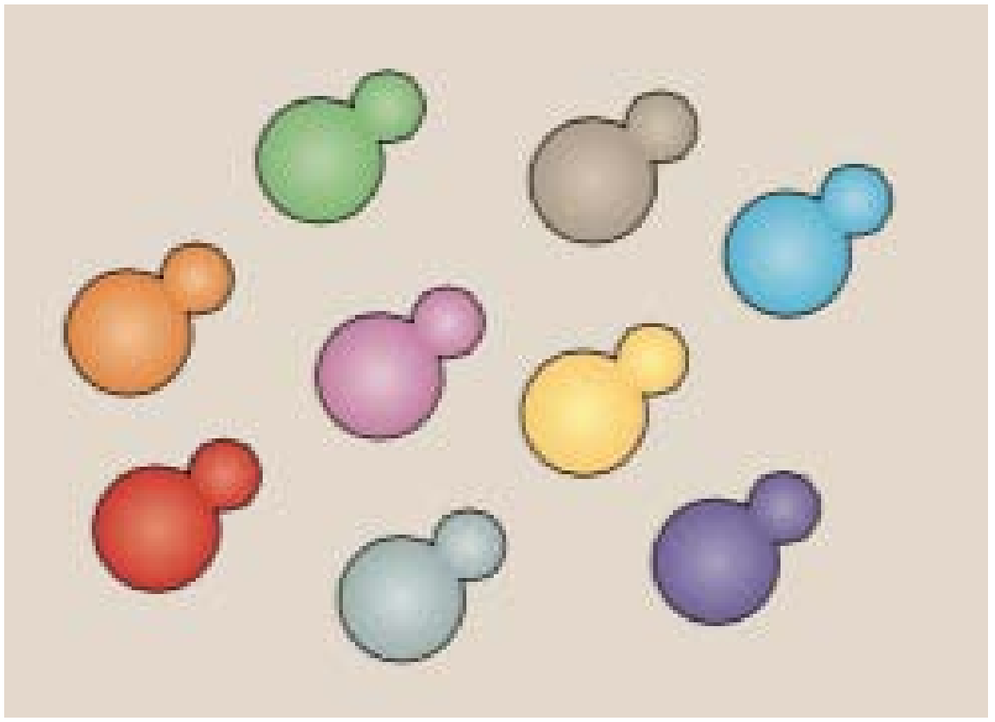
- how repeatable is evolution?
- how well can we localize/identify QTL?
- what are the origins and fates of adaptive alleles?



- eukaryote
- sexual recombination

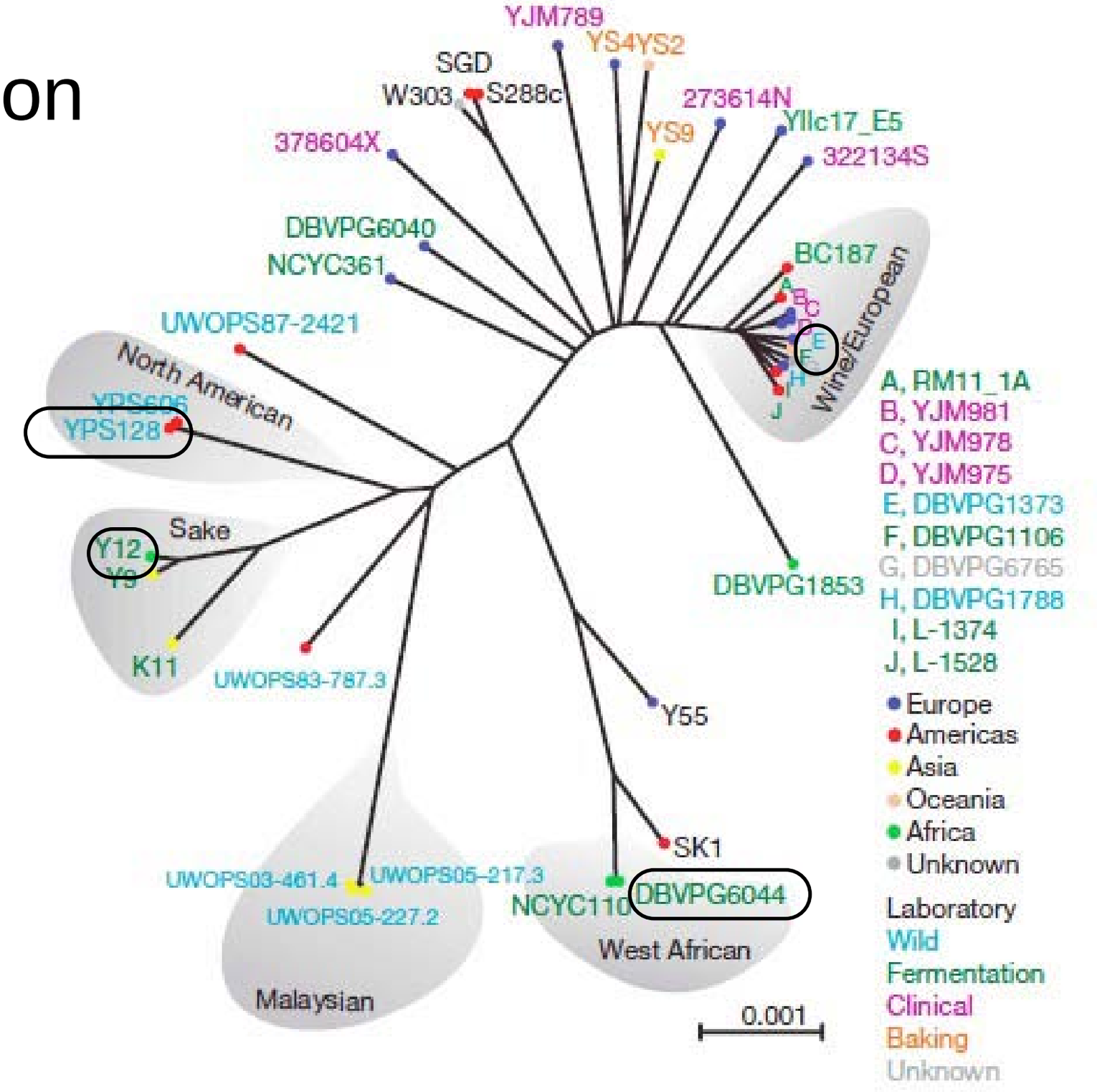
- short generations
- archivable

Saccharomyces cerevisiae

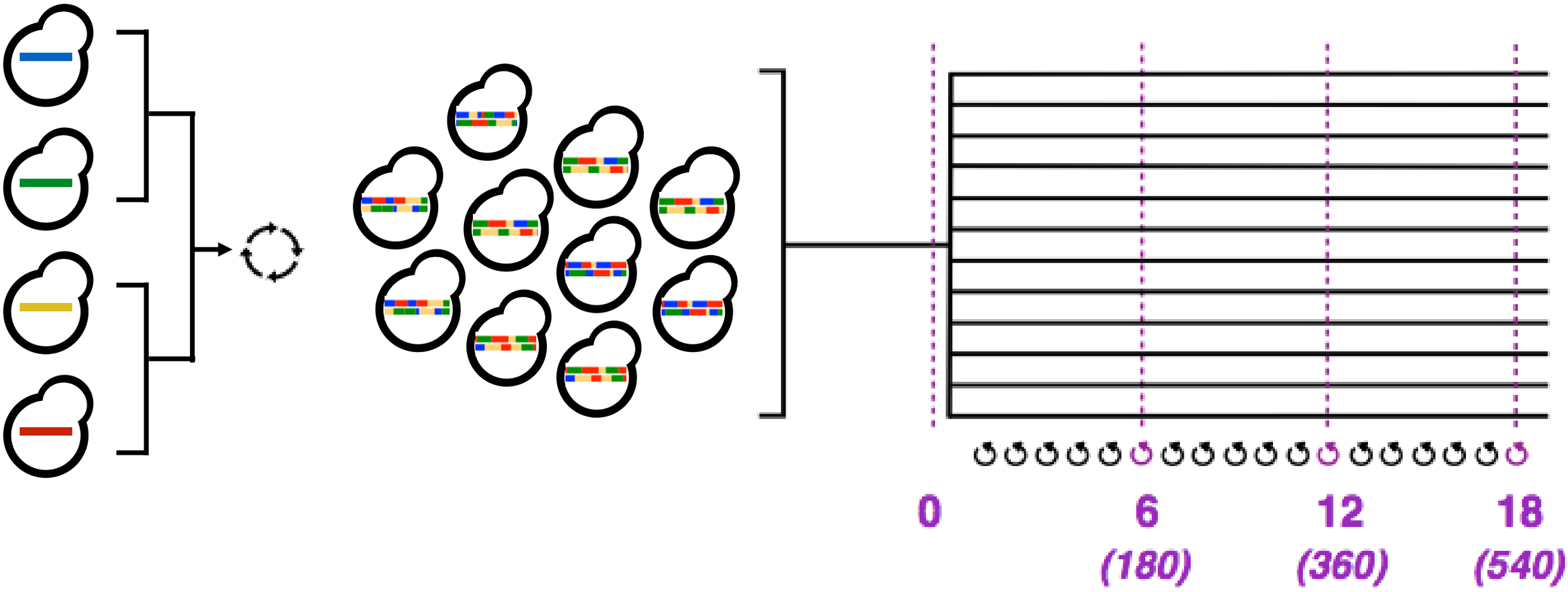


- eukaryote
- sexual recombination
- short generations
- archivable

ecombinant population
with 4 founders



Liti *et al.* 2009 **Nature**
Cubillos *et al.* 2013 **Genetics**

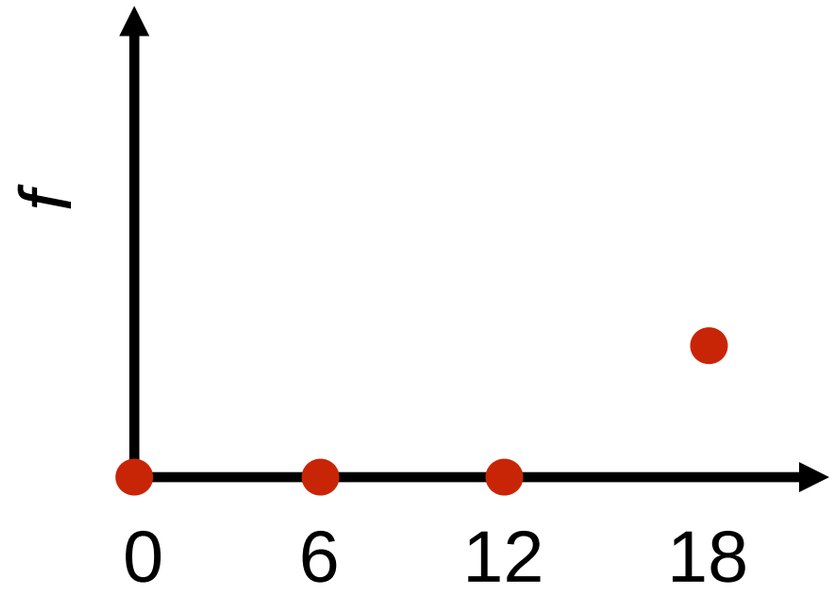


selection treatment = regular outcrossing

Burke *et al.* 2014 **MBE**

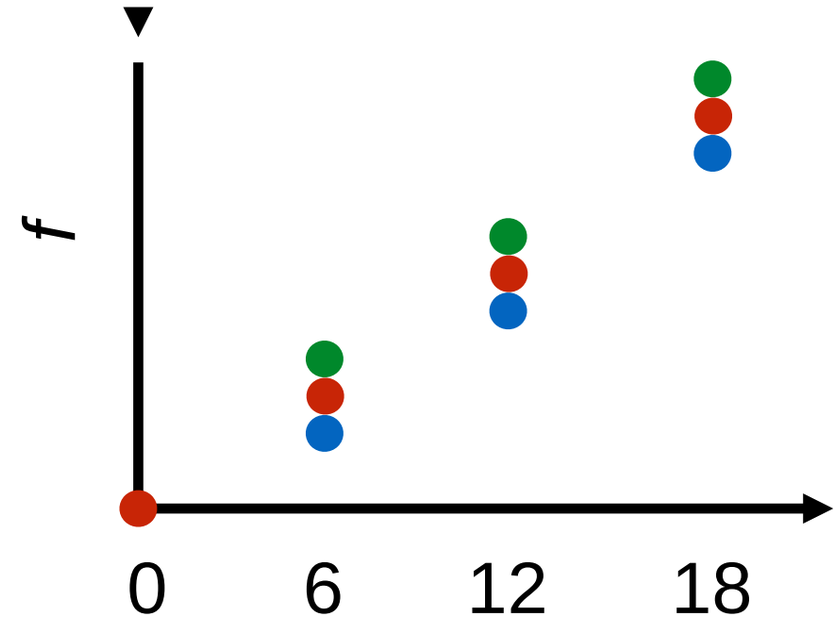
15 candidate de novo mutations

- none > 0.2 in any replicate
- none private to any replicate



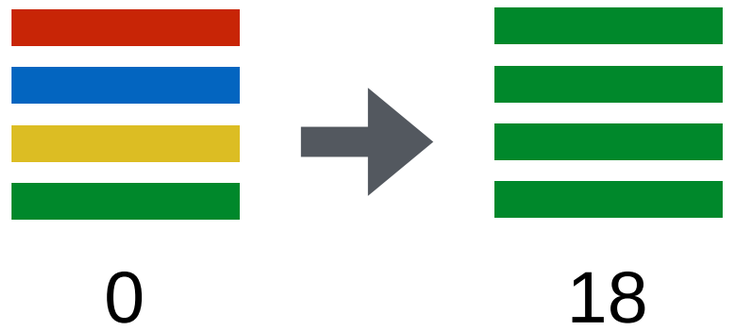
~75K standing variants

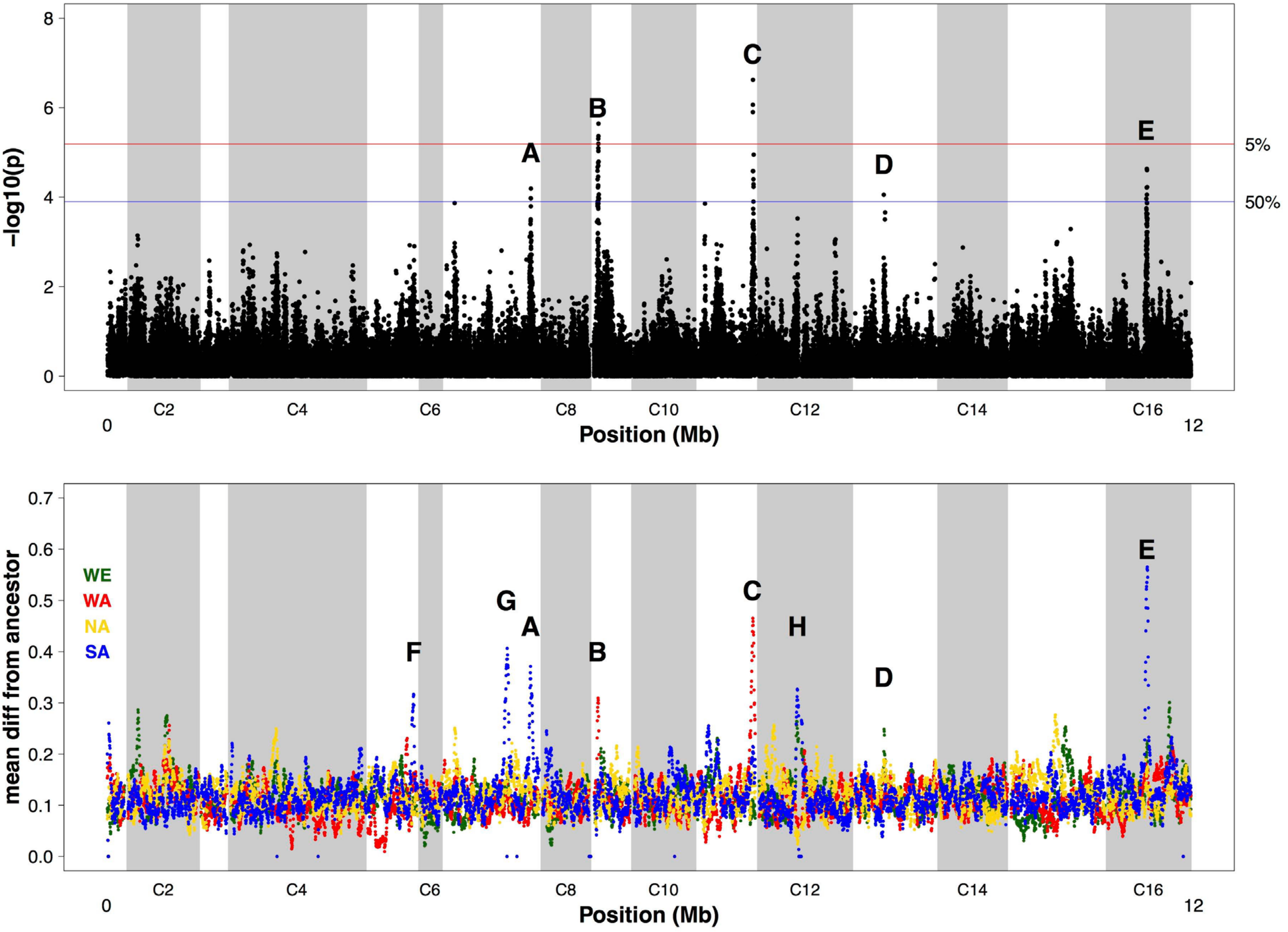
- fit linear models
- permutation tests



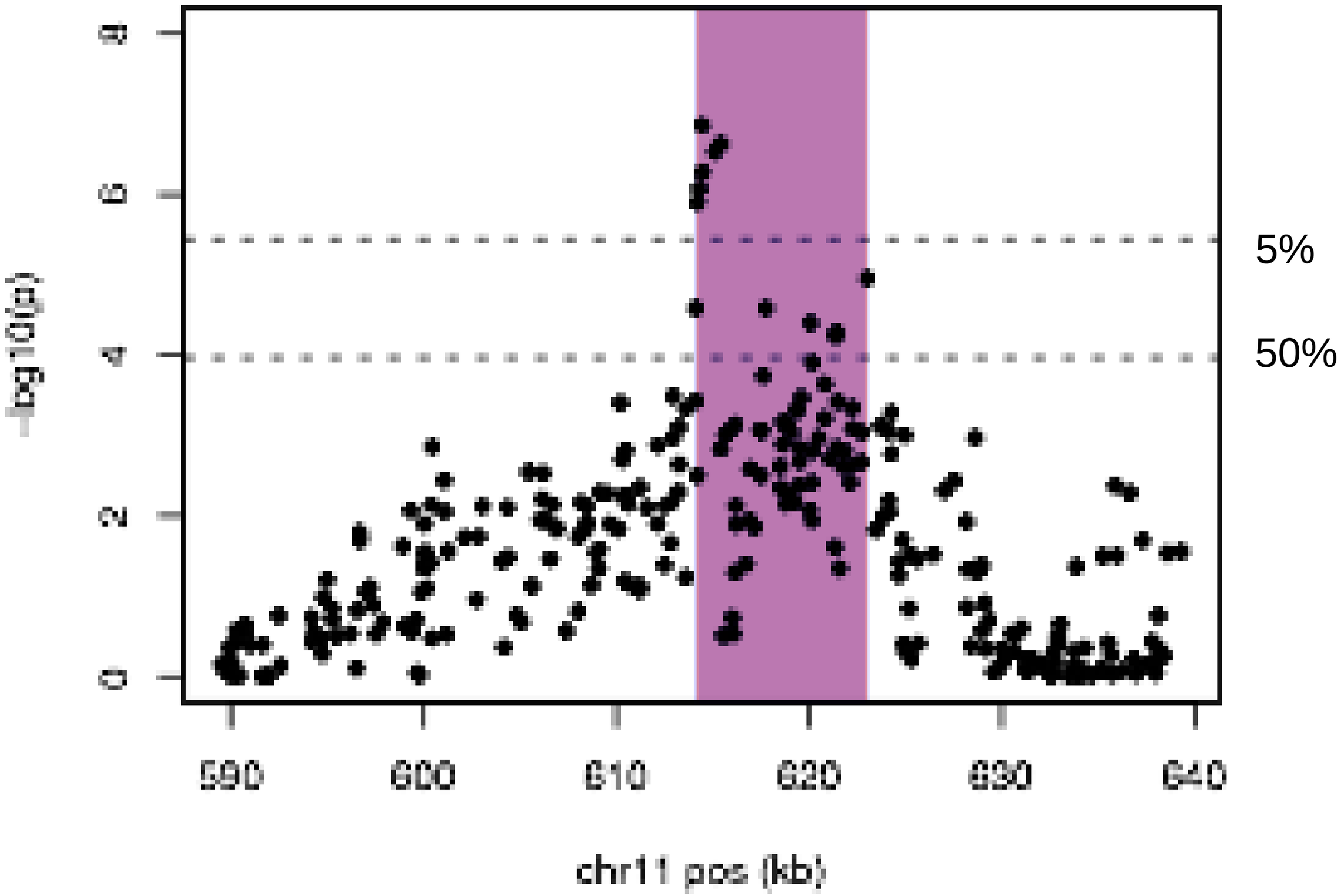
haplotype frequencies

- founder alleles known
- enrichment in evolved lines

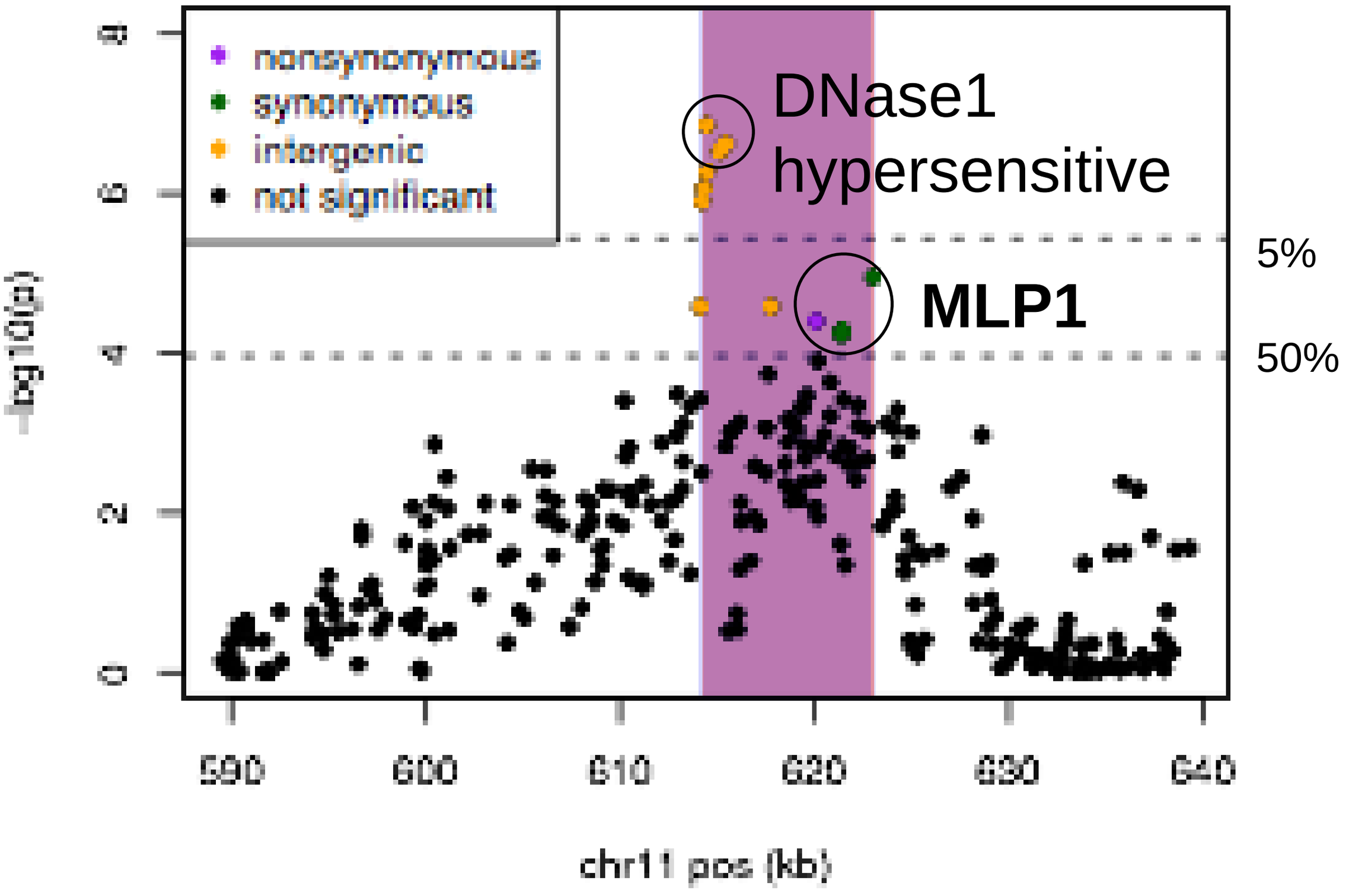


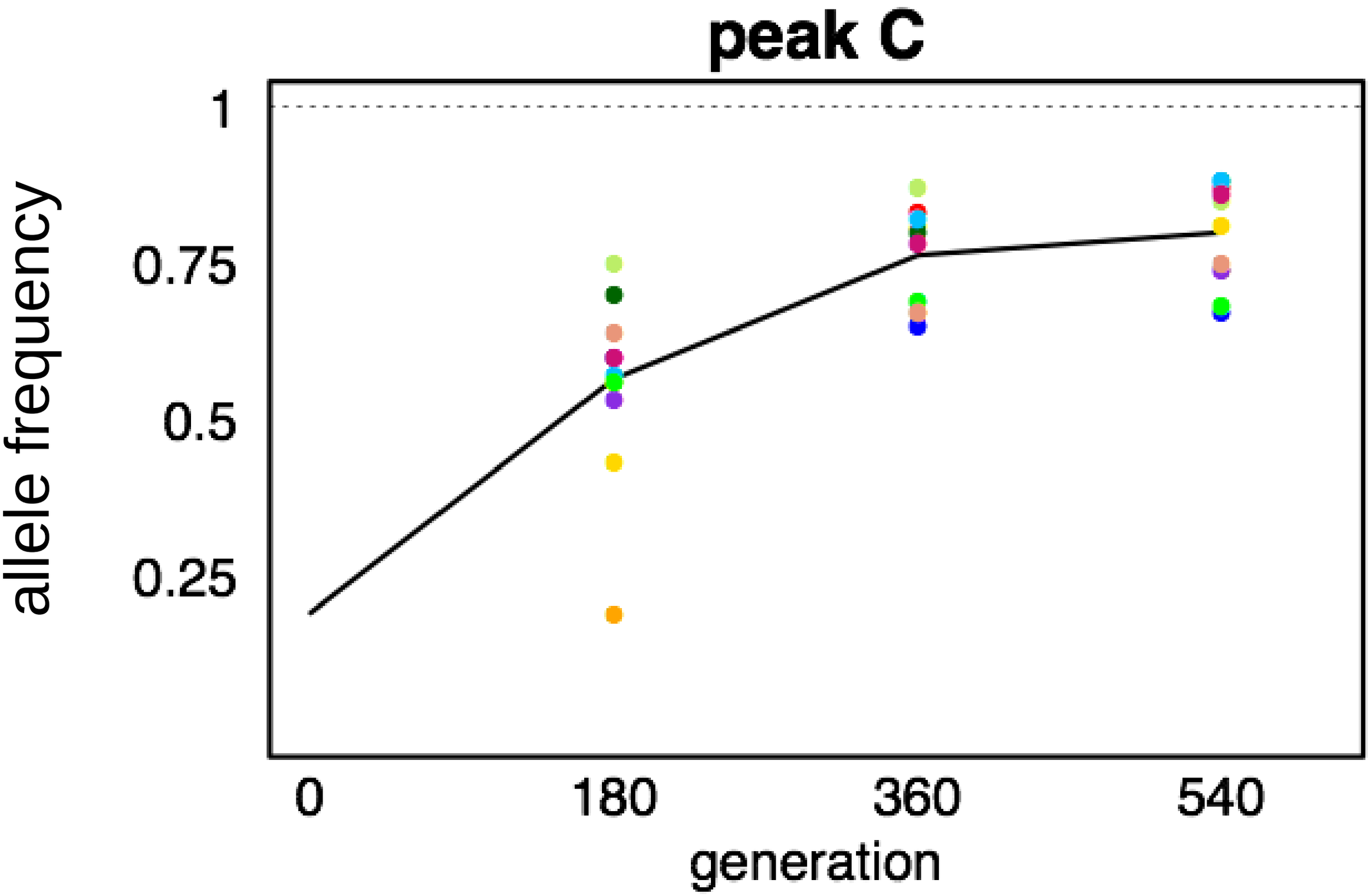

−log10(p)
8
6
4
2
0
A
B
C
D
E
5%
50%
C2
C4
C6
C8
C10
C12
C14
C16
0
12
Position (Mb)
mean diff from ancestor
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
WE
WA
NA
SA
F
G
A
B
C
H
D
E
C2
C4
C6
C8
C10
C12
C14
C16
0
12
Position (Mb)

peak C



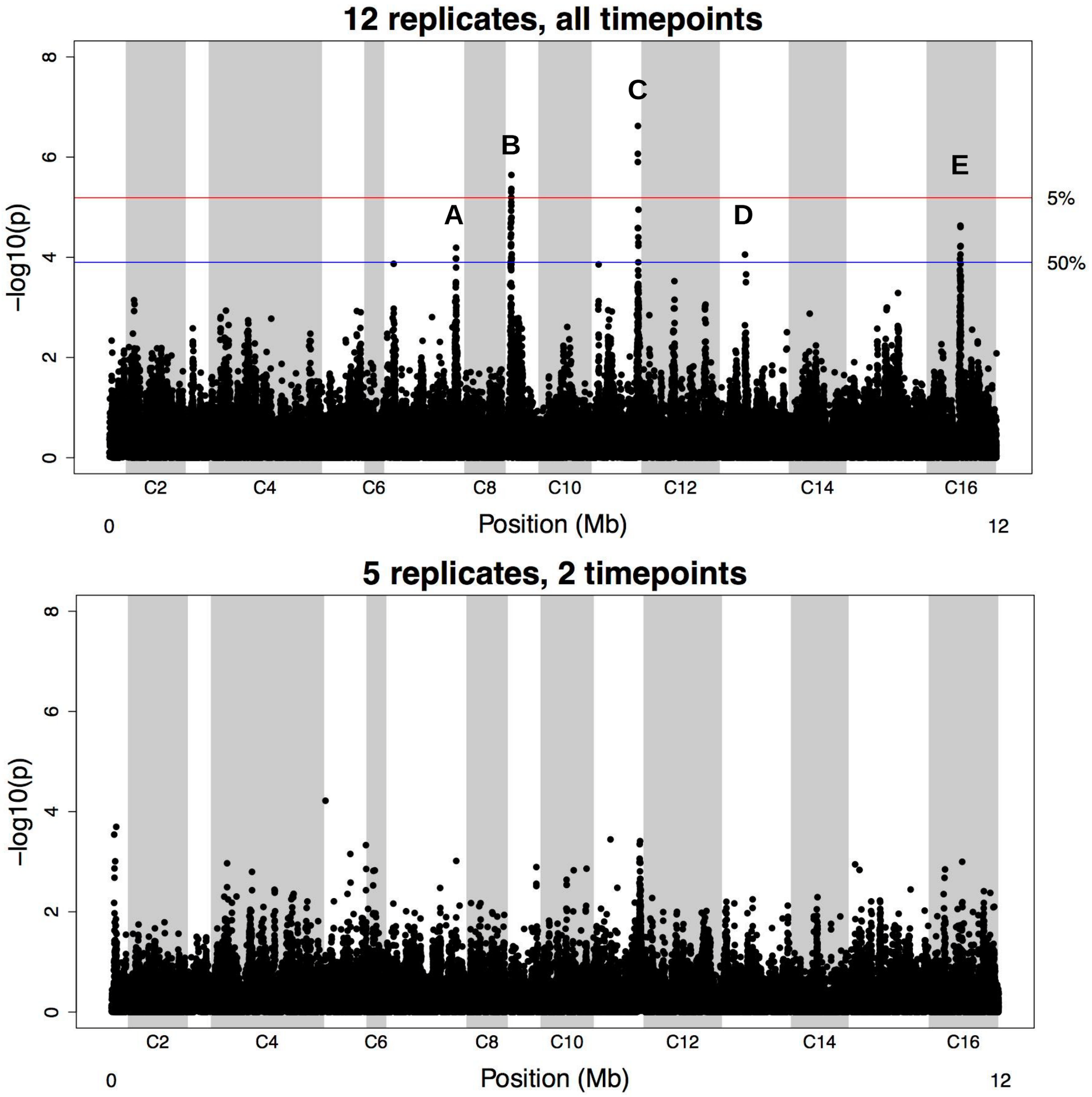
peak C


nonsynonymous
synonymous
intergenic
not significant
DNase1
hypersensitive
MLP1
5%
50%
8
6
4
2
0
-log10(p)
590
600
610
620
630
640
chr11 pos (kb)


peak C
1
0.75
0.5
0.25
allele frequency
0
180
360
540
generation

conclusions

- how repeatable is evolution?
- how well can we localize/identify QTL?
- what are the origins and fates of adaptive alleles?


12 replicates, all timepoints
−log10(p)
0
2
4
6
8
A
B
C
D
E
5%
50%
C2
C4
C6
C8
C10
C12
C14
C16
0
Position (Mb)
12
5 replicates, 2 timepoints
−log10(p)
0
2
4
6
8
C2
C4
C6
C8
C10
C12
C14
C16
0
Position (Mb)
12

A Guide for the Design of Evolve and Resequencing Studies

Robert Kofler[1] and Christian Schlötterer*[1]
[1]Institut für Populationsgenetik, Vetmeduni Vienna, Vienna, Austria
*Corresponding author: E-mail: christian.schloetterer@vetmeduni.ac.at.
Associate editor: John Parsch

Abstract

Standing genetic variation provides a rich reservoir of potentially useful mutations facilitating the adaptation to novel environmen
tion can als
experiments
about the d
evaluate dif
linkage dise
replicates ar
is more imp
beneficial lo
number of l
positives. O
to identify a
trajectories
Key words: e

Introducti

The importa
tion of natur
Pennings 20
Experimental
strated that
notypic resp
Turner et al.
of next-gene
interest in ex
sible to iden
et al. 2010;
Remolina et
uniformly ev
promises to
separated fo
and populati
the potencia
(molecular g
these loci du
little is know
optimized to
vored loci (B
whether a fe
of neutrally
whether E&
these questi
and evaluate

Article

Article

The Power to Detect Quantitative Trait Loci Using Resequenced, Experimentally Evolved Populations of Diploid, Sexual Organisms

James G. Baldwin-Brown,*[1] Anthony D. Long,[1] and Kevin R. Thornton[1]
[1]Department of Ecology and Evolutionary Biology, University of California, Irvine
*Corresponding author: E-mail: jbaldwi1@uci.edu.
Associate editor: John Parsch

Abstract

A novel approach for dissecting complex traits is to experimentally evolve laboratory populations under a controlled environment shift, resequence the resulting populations, and identify single nucleotide polymorphisms (SNPs) and/or genomic regions highly diverged in allele frequency. To better understand the power and localization ability of such an evolve and resequence (E&R) approach, we carried out forward-in-time population genetics simulations of 1 Mb genomic regions under a large combination of experimental conditions, then attempted to detect significantly diverged SNPs. Our analysis indicates that the ability to detect differentiation between populations is primarily affected by selection coefficient, population size, number of replicate populations, and number of founding haplotypes. We estimate that E&R studies can detect and localize causative sites with 80% success or greater when the number of founder haplotypes is over 500, experimental populations are replicated at least 25-fold, population size is at least 1,000 diploid individuals, and the selection coefficient on the locus of interest is at least 0.1. More achievable experimental designs (less replicated, fewer founder haplotypes, smaller effective population size, and smaller selection coefficients) can have power of greater than 50% to identify a handful of SNPs of which one is likely causative. Similarly, in cases where $s \geq 0.2$, less demanding experimental designs can yield high power.

Key words: **simulation, QTL detection, genomics, adaptive evolution, experimental evolution, evolve and resequence.**

Introduction

Quantitative traits are of special interest to biologists. The variation in many traits of medical, agricultural, and evolutionary relevance is due to the concerted action of several genes and the environment. Quantitative trait locus (QTL) mapping has been effective at explaining the majority of the heritability of a trait but is poorly suited to resolving the location of QTL beyond several cM (Mackay et al. 2009). More recently, several groups have attempted to increase the resolution of QTL mapping using advanced generation recombinant inbred lines (*cf.* Kover et al. 2009; Aylor et al. 2011; King et al. 2012), but resolution is still limited to cM scales. Recently, genome wide association studies (GWAS) have become a major method for investigating the genetic basis for quantitative traits (The Wellcome Trust Case Control Consortium, 2007a, 2007b; Craddock et al. 2010). Although GWAS studies have identified replicable associations between SNPs and complex traits, associated SNPs tend to explain only a small fraction of the heritable variation in the study trait (Manolio et al. 2009), a problem that cannot be solved by increasing sample sizes to tens of thousands of individuals (Signer-Hasler et al. 2012) or replacing SNPchips with complete resequenced genomes (Spencer et al. 2009). Clearly, it is of value to explore novel methods for dissecting complex traits.

In systems that have short generation times and that can easily be reared in the lab in large numbers, an alternative experimental approach to dissecting complex traits has been to "evolve and resequence" (E&R) populations of organisms. E&R studies have been performed with both asexual (Riehle et al. 2001; Barrick et al. 2009; Kishimoto et al. 2010; Tenaillon et al. 2012; Parts et al. 2011) and sexual (Teotónio et al. 2009; Burke et al. 2010; Johansson et al. 2010; Turner et al. 2011; Orozco-Terwengel et al. 2012; Turner and Miller 2012) populations. Because asexual experimental evolution lacks recombination and standing variation in the base population, the footprints of selection in the genome and the means by which an investigator may hope to identify causal variants are different in sexual and asexual systems. Thus, we limit our focus to E&R studies in sexual systems. Under the E&R paradigm, a base population is divided into several replicate populations, half of which are subjected to a well-defined selection pressure, and the other half of which are maintained without selection. Next, the DNA pools from each population are resequenced using NextGen technology and allele frequencies in each pool are estimated. SNPs and/or genomic regions showing consistent differentiation between selected and control population are candidates for harboring causative variants. Studies using this design have claimed to detect numbers of candidate causative sites (CS) from 662 (Burke et al. 2010) to almost 5,000 (Orozco-Terwengel et al. 2012) for various quantitative traits. Currently, the CSs detected by E&R methods have not been validated.

To date, the field of E&R has been almost entirely empirically motivated. Study designs have varied greatly in terms of

Open Access
31(4):1040–1055 doi:10.1093/molbev/msu048 Advance Access publication January 18, 2014

2014 **MBE** 31(4)

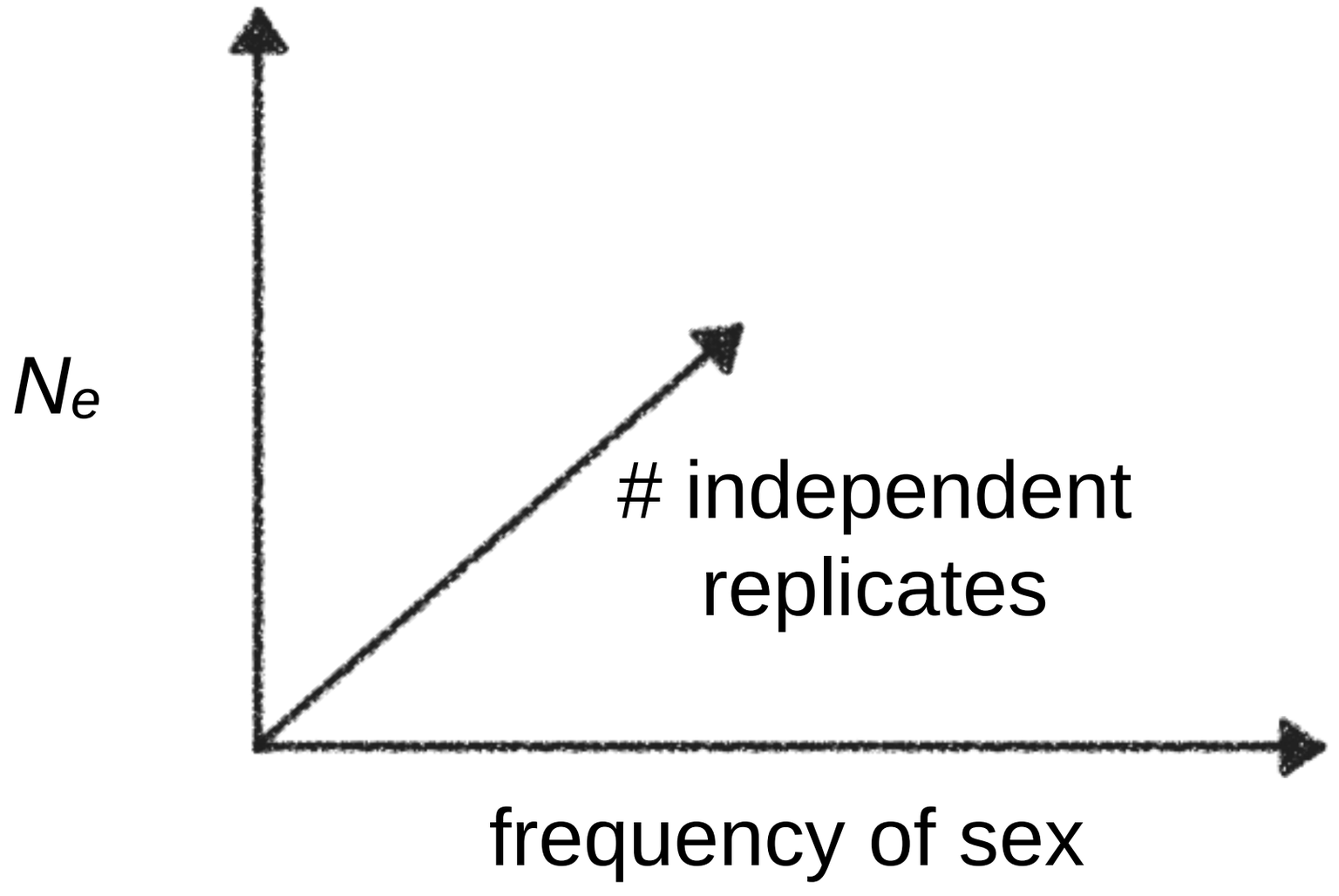
best practices:

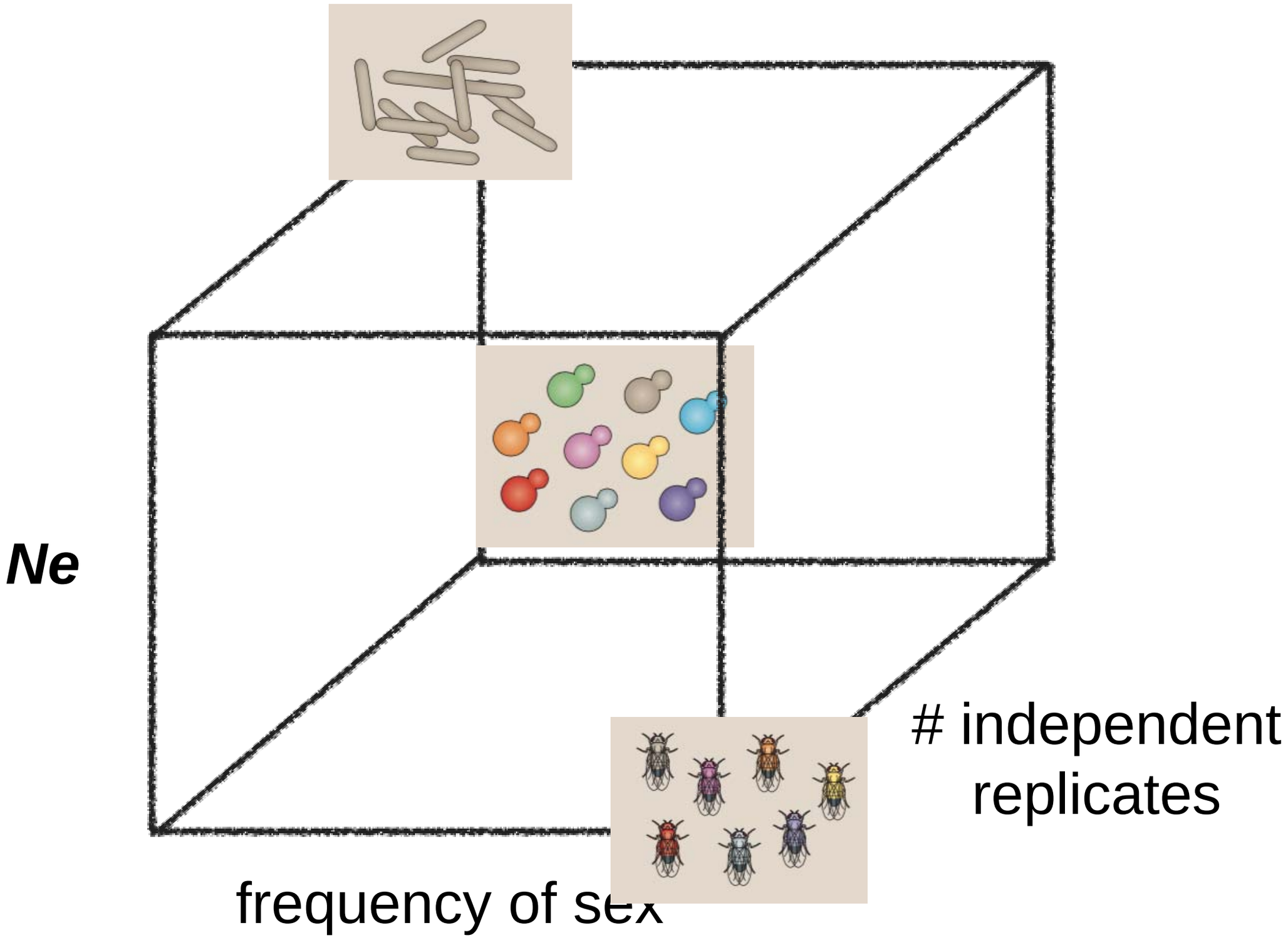
N_e > 1000

replicates > 25

generations > 500

experimental evolution parameter space




Ne
frequency of sex
independent replicates

acknowledgements

UC Irvine:
Michael Rose
Tony Long
Kevin Thornton
Laurence Mueller
Parvin Shahrestani
Mark Phillips
Larry Cabral
Thomas Barter

USC:
Joseph Dunham

NC A&T State U:
Joseph Graves

UT Tyler:
Kate Hertweck

UNLV:
Mira Han

IRCAN Nice:
Gianni Liti

OSU:
Adelaide Rhodes