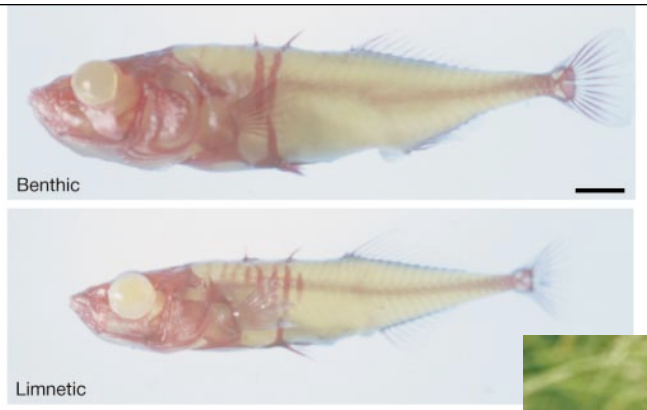
Genome-wide association studies in *Arabidopsis thaliana*

Magnus Nordborg
University of Southern California



The genetic basis of evolutionary change

Premise

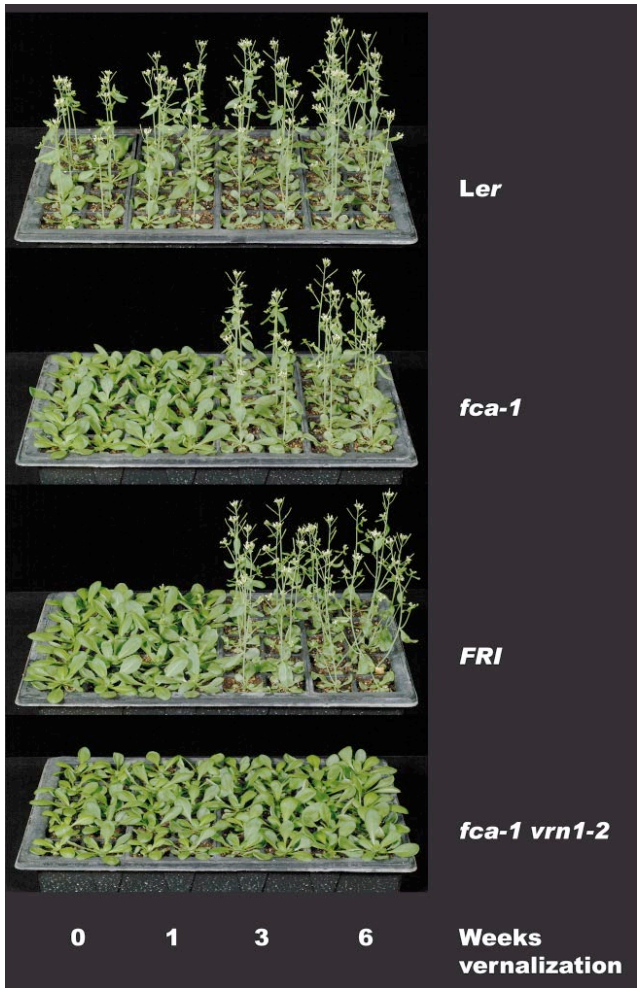
- Technology is revolutionizing many areas of biology — in particular, sequencing and genotyping technologies are advancing at an astonishing rate
- Genome sequencing is routine, and it is now possible to study genetic variation by sequencing large numbers of individuals
- Systems biology will meet natural variation

It doesn't get any better than this...

- Model plant with compact (120 Mb), sequenced genome
- Widely distributed, highly variable, and locally adapted
- Highly selfing:
  - Increased linkage disequilibrium

!

  - Naturally occurring inbred lines



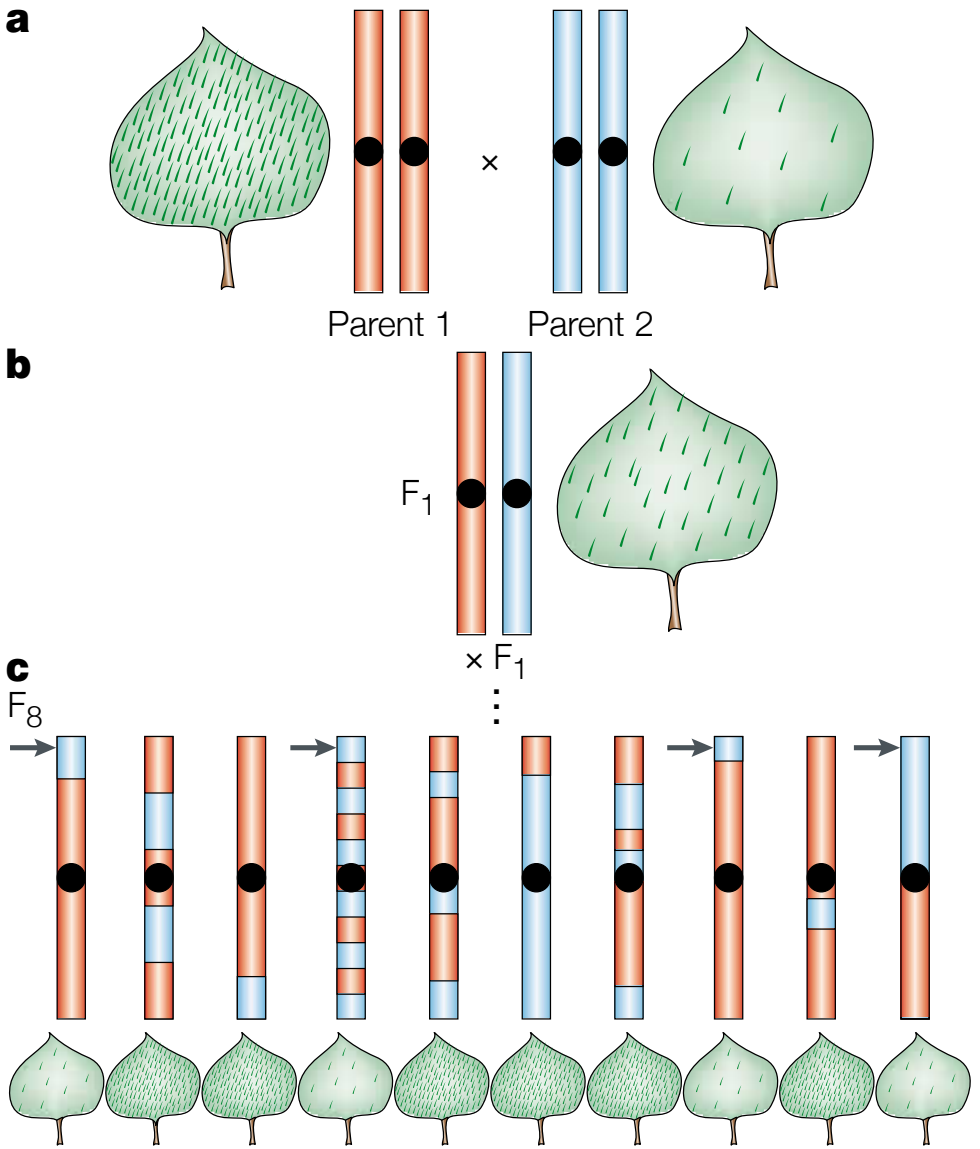
(from Henderson *et al.* 2003, *Annu. Rev. Genet.*)

Plants “stand still to be counted”

Genome-wide association studies (GWAS)

- Finding genes responsible for variation in complex traits is still very difficult
- The decreasing costs of genotyping has raised the promise of so-called "association mapping"

The basic principle behind mapping



(from Mauricio 2001, *Nature Rev. Genet.*)

Linkage vs association mapping

- *Linkage mapping* – look for associations in populations with known relationships (*e.g.* the members of a pedigree, or the offspring of a cross)
- *Association mapping* – look for associations in populations of "unrelated" individuals

The HapMap Project

Vol 447|7 June 2007|**doi:10.1038/nature05911**

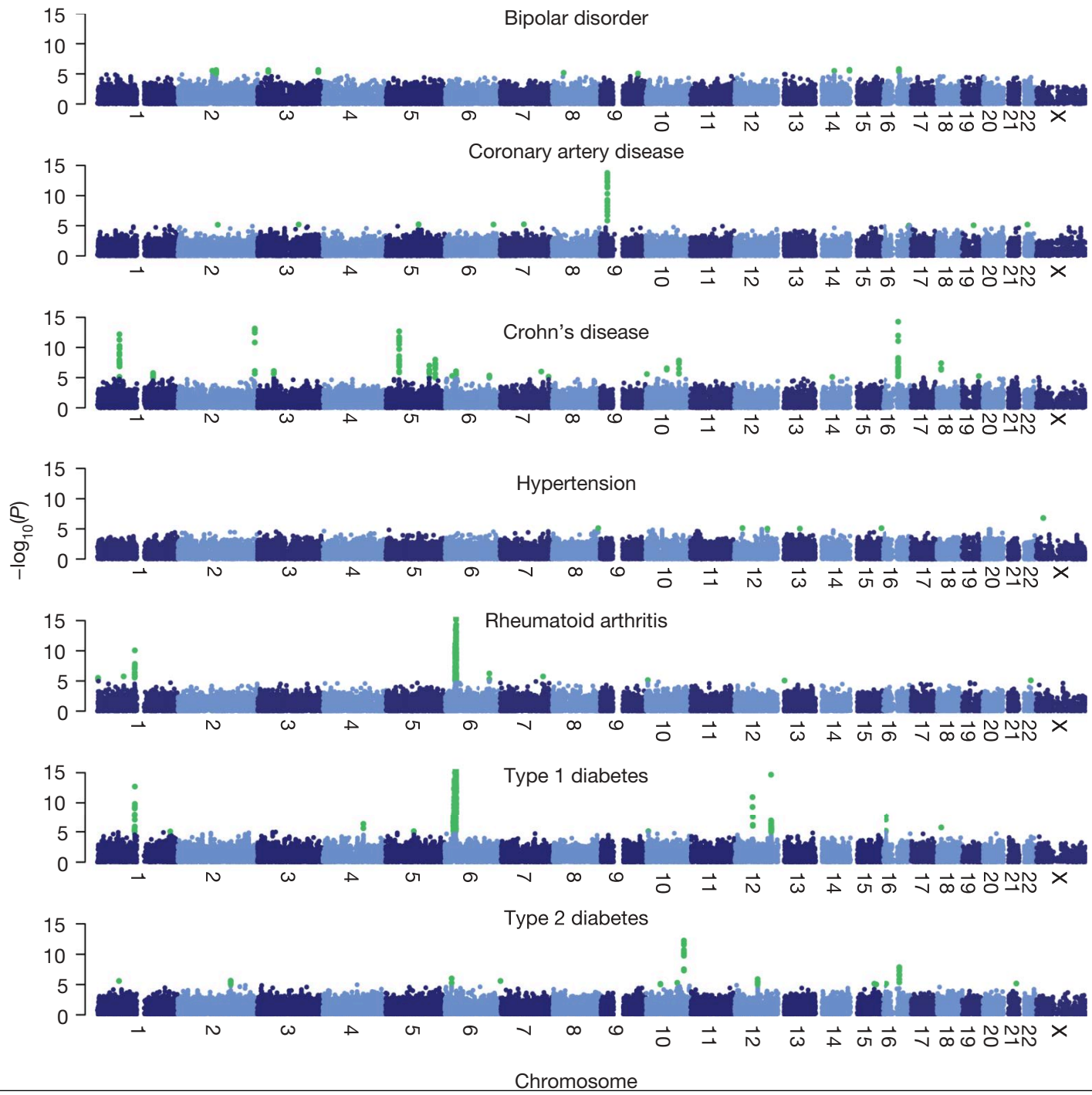
nature

ARTICLES

Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls

The Wellcome Trust Case Control Consortium*

There is increasing evidence that genome-wide association (GWA) studies represent a powerful approach to the identification of genes involved in common human diseases. We describe a joint GWA study (using the Affymetrix GeneChip 500K Mapping Array Set) undertaken in the British population, which has examined ~2,000 individuals for each of 7 major diseases and a shared set of ~3,000 controls. Case-control comparisons identified 24 independent association signals at $P < 5 \times 10^{-7}$: 1 in bipolar disorder, 1 in coronary artery disease, 9 in Crohn's disease, 3 in rheumatoid arthritis, 7 in type 1 diabetes and 3 in type 2 diabetes. On the basis of prior findings and replication studies thus-far completed, almost all of these signals reflect genuine susceptibility effects. We observed association at many previously identified loci, and found compelling evidence that some loci confer risk for more than one of the diseases studied. Across all diseases, we identified a large number of further signals (including 58 loci with single-point *P* values between 10^{-5} and 5×10^{-7}) likely to yield additional susceptibility loci. The importance of appropriately large samples was confirmed by the modest effect sizes


Bipolar disorder
Coronary artery disease
Crohn's disease
Hypertension
Rheumatoid arthritis
Type 1 diabetes
Type 2 diabetes
−log10(P)
15
10
5
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
Chromosome

Why association mapping?

+ Much higher resolution because of historical recombination
+ No crosses or pedigrees needed
— Results more difficult to interpret (false positives and negatives)
— Not guaranteed to work

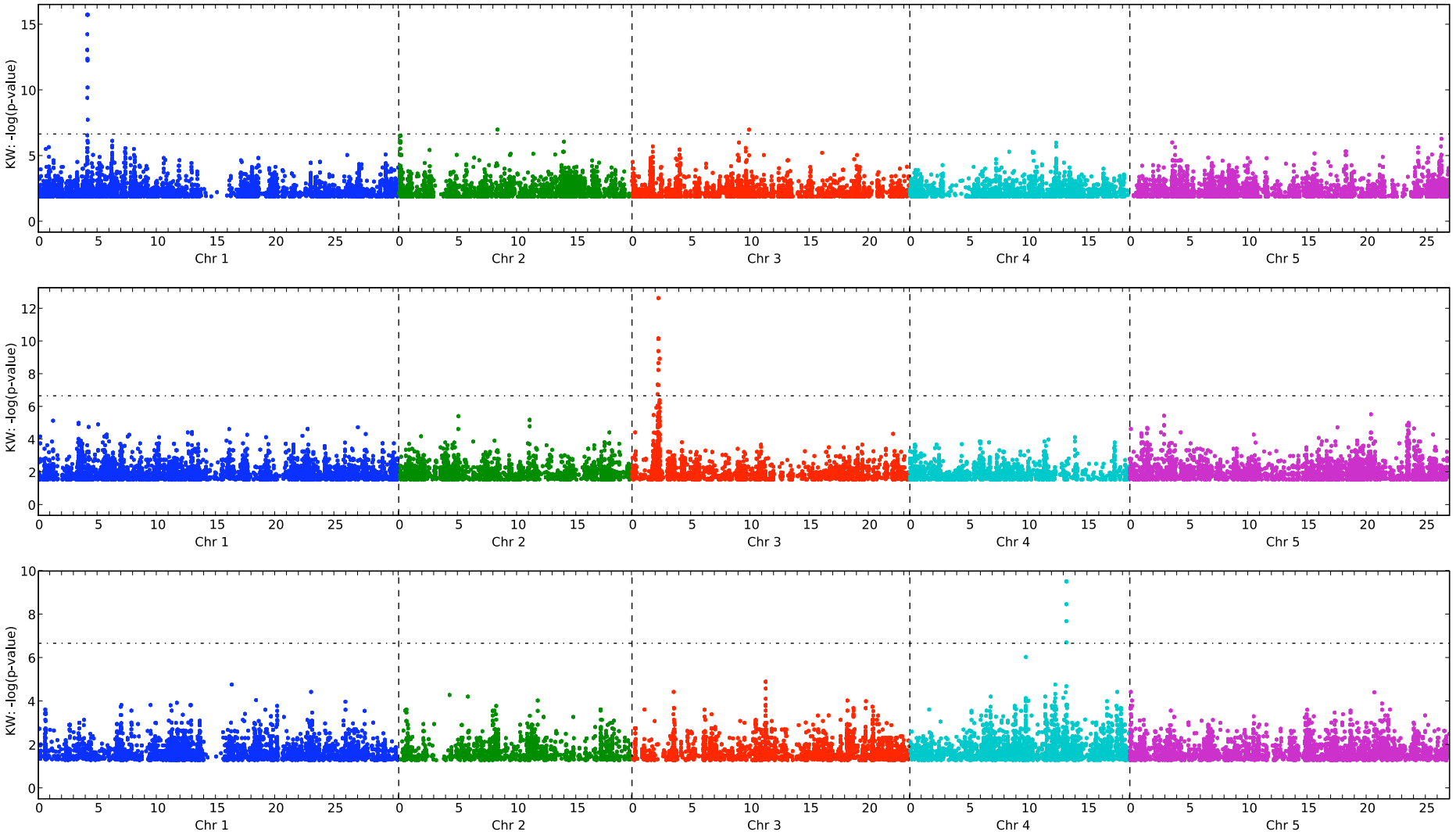
Towards genomic polymorphism data

1. Dideoxy re-sequencing of ~1,500 short fragments in a global sample of 96 lines
2. Whole-genome re-sequencing of 20 lines (a subset of the 96) using Perlegen technology
3. Genotyping of 250k SNPs in ~1,300 lines using custom Affymetrix chip (*in progress*)
4. Solexa shotgun sequencing of some or all of these lines (*in progress*)

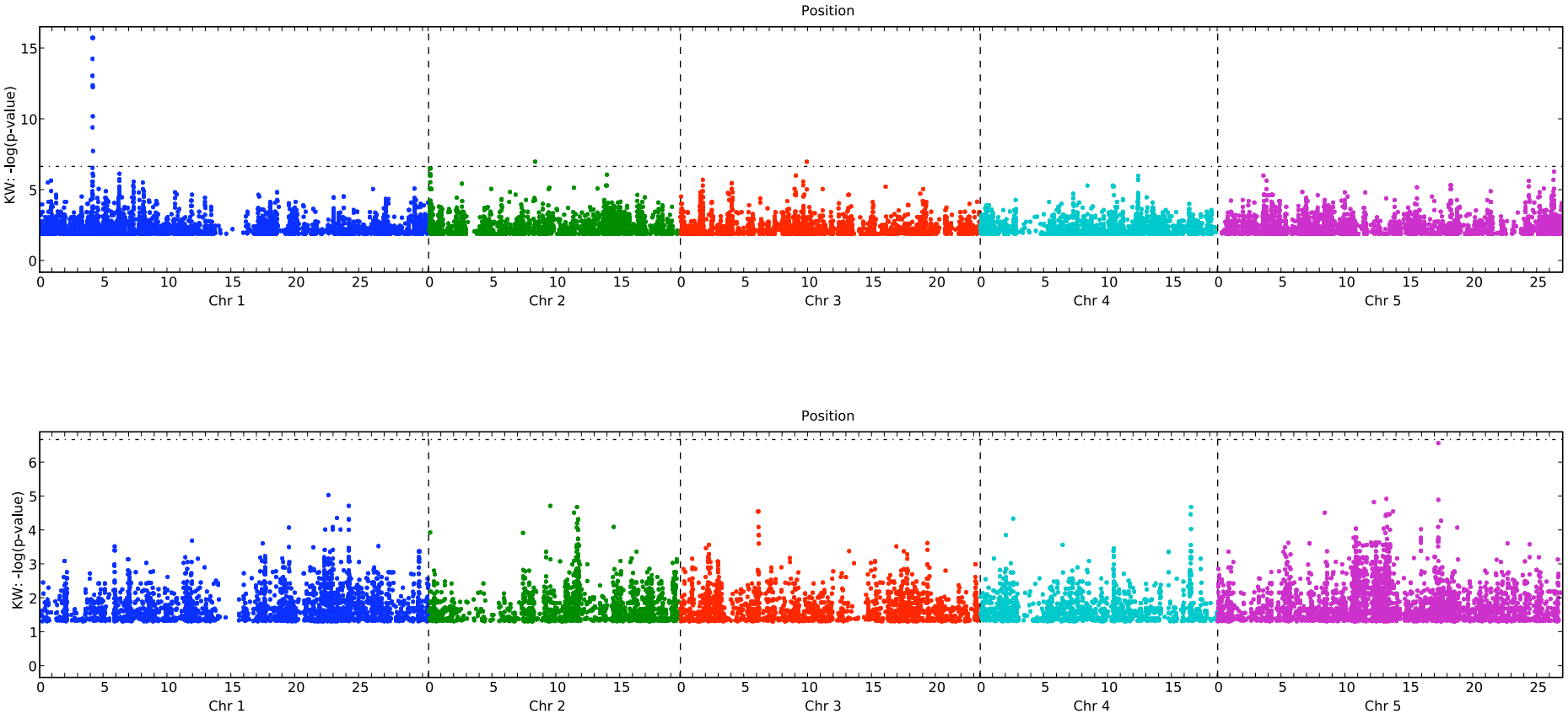
GWA in *A. thaliana*

- 186 different phenotypes, measured in the same 96 (or 192, or 384) lines;
- 4 different statistical methods;
- a total of almost 200 million association scores...

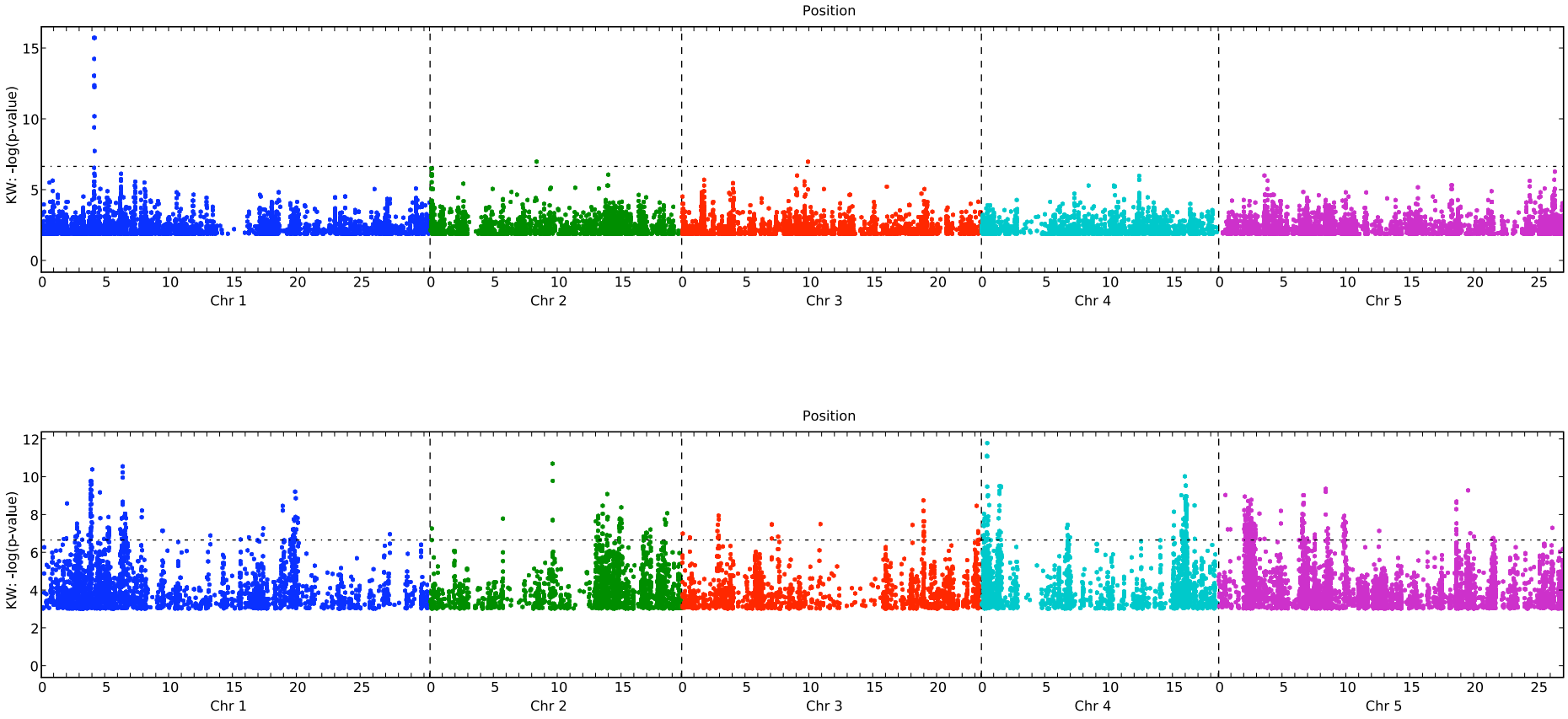
It works!



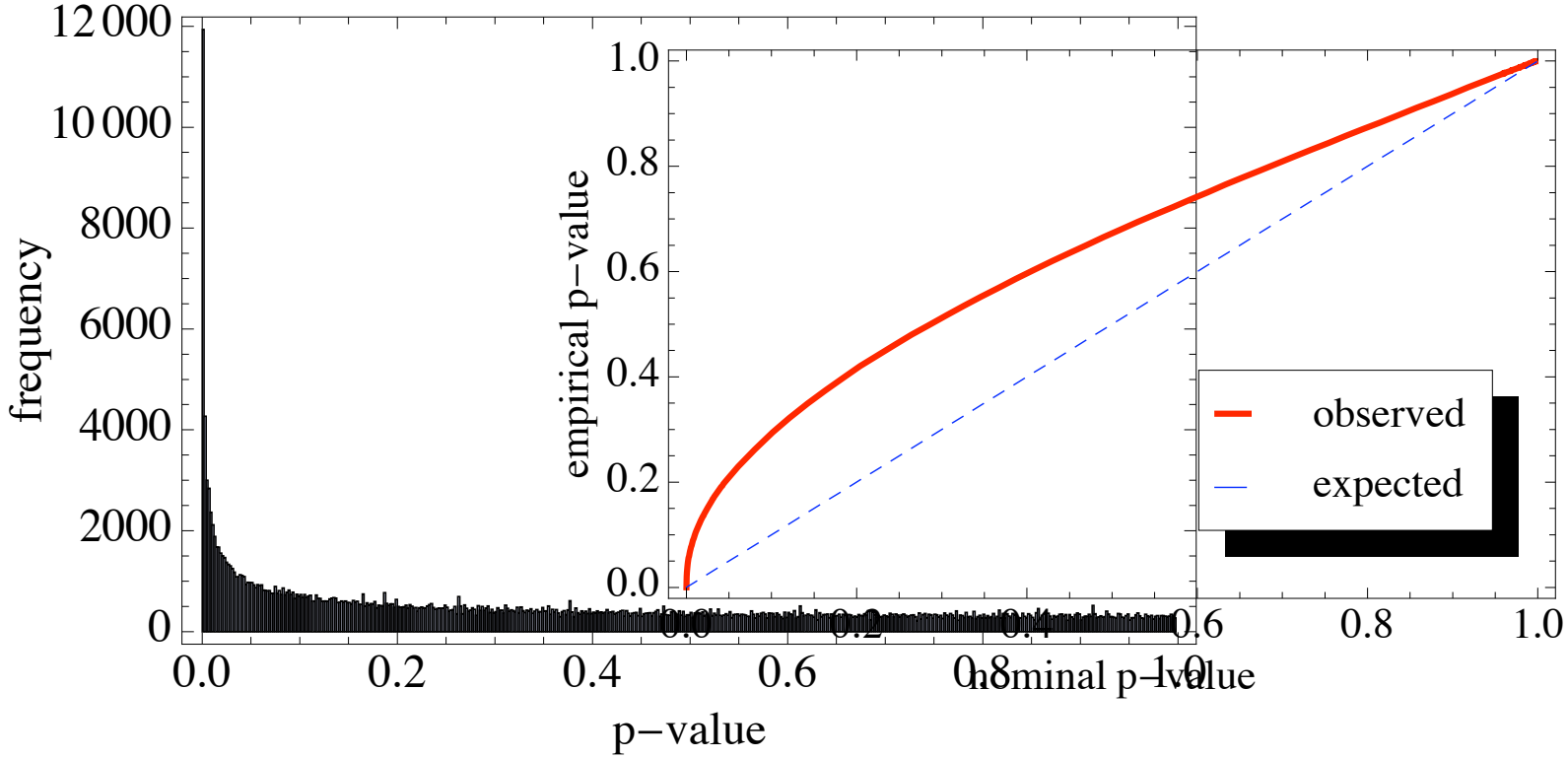
At least sometimes...



A much more “interesting” problem...



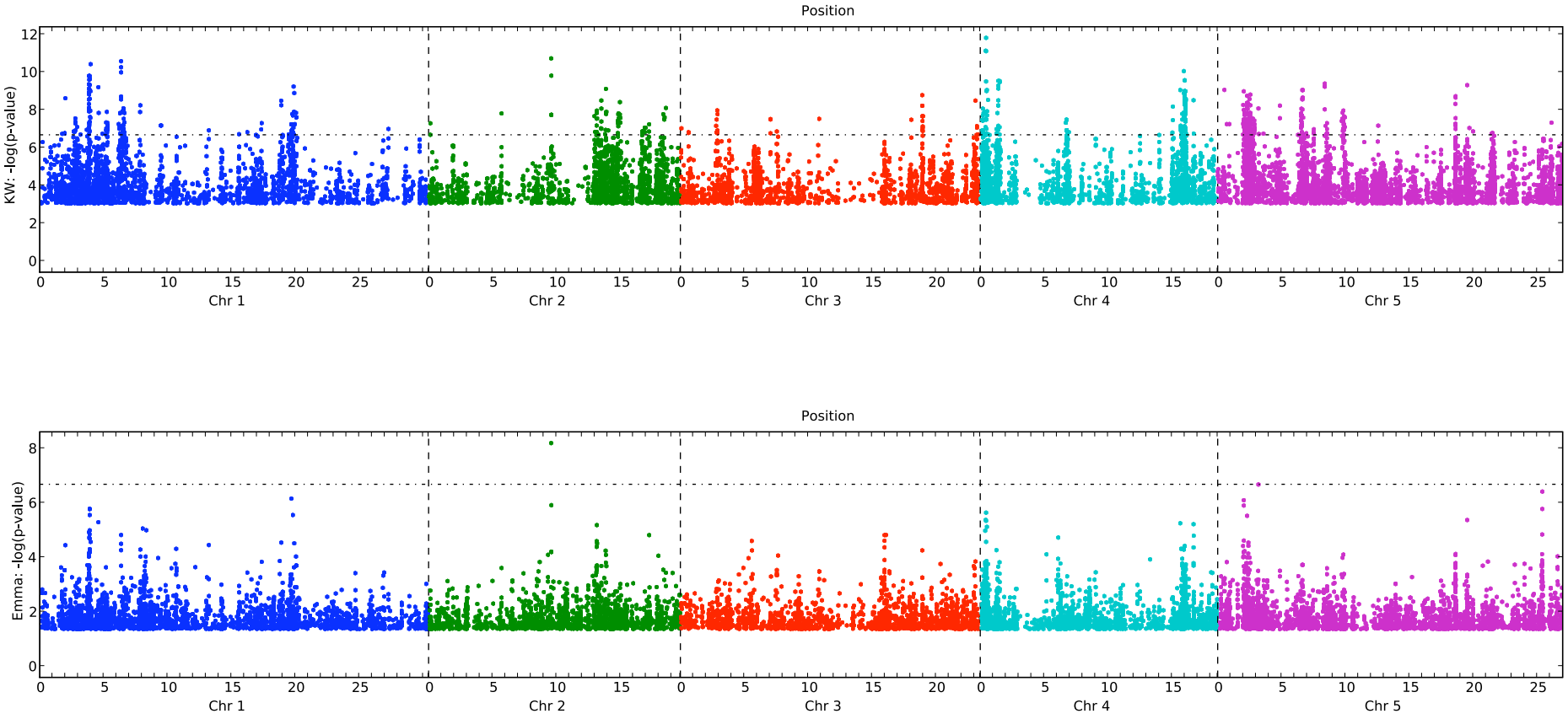
Confounding by population structure



What is to be done?

- Genomic control
  (Devlin & Roeder 1999, *Biometrics*)
- Structured association
  (Pritchard *et al.* 2000, *Am. J. Hum*
- Principal-components ap
  (Price *et al.* 2006, *Nature Gen*
- Mixed-model approach
  (Yu *et al.* 2006, *Nature Genet.*)

Correcting for population structure works...



No magic bullet

- Some confounding remains (false positives)
- Removing confounding removes true positives (introduces false negatives)

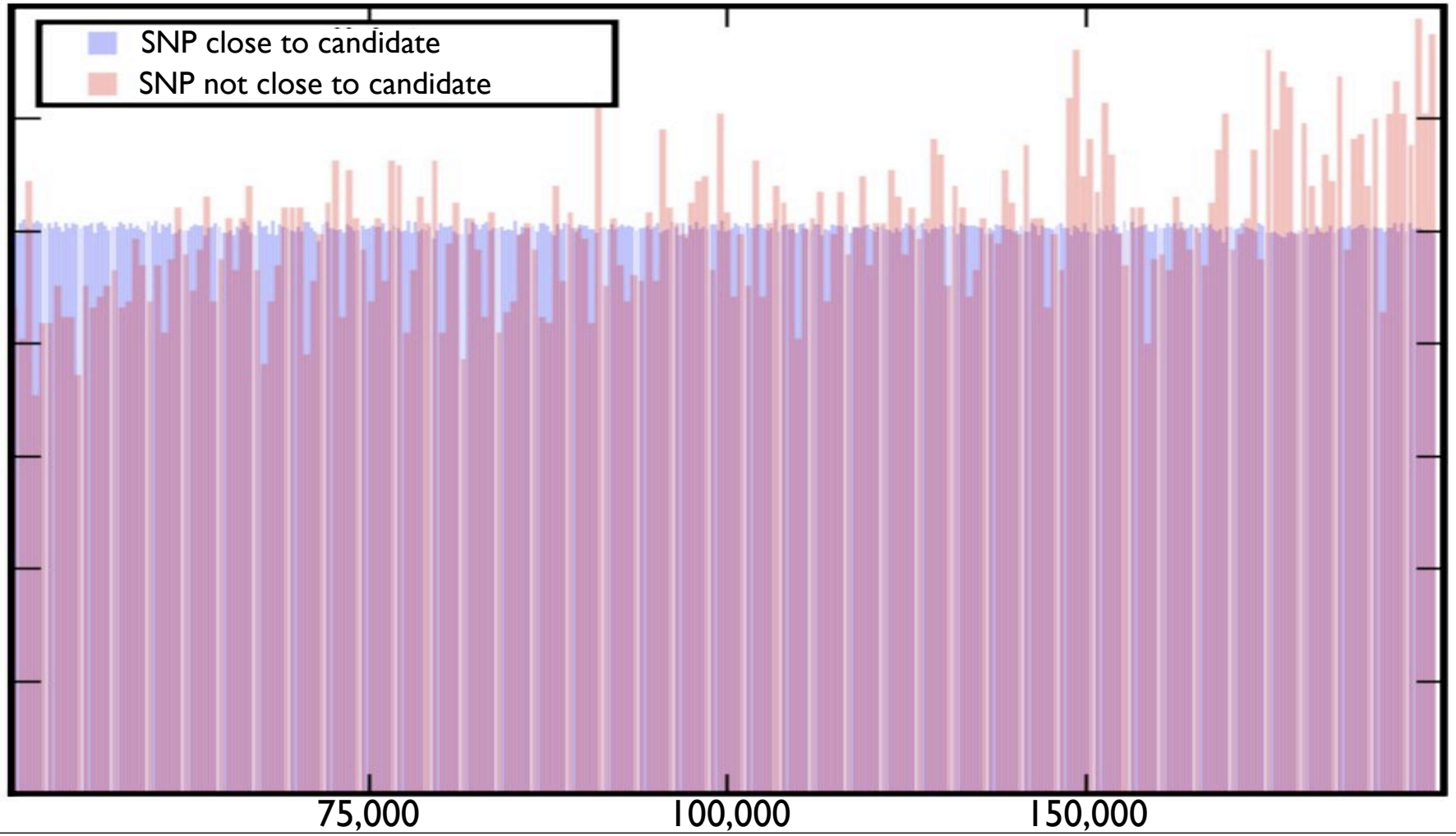
No magic bullet

- Some confounding remains (false positives)
- Removing confounding removes some true positives (introduces false negatives)
- There is no substitute for independent evidence!

Generating lists of interesting associations

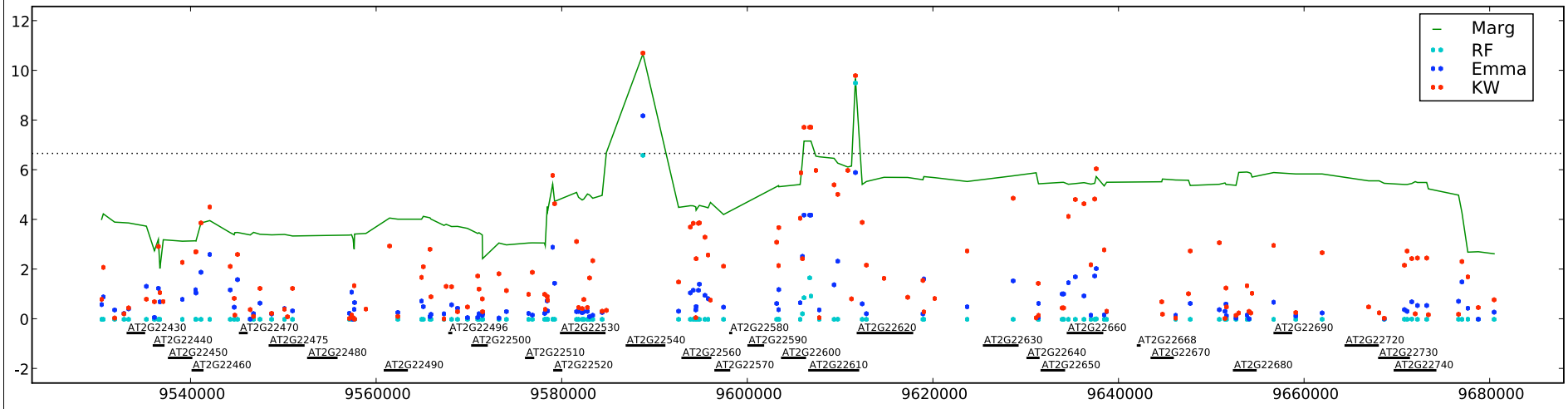
- Significant peaks — when p-values have meaning...
- Start from the top — when they don't...

There is signal

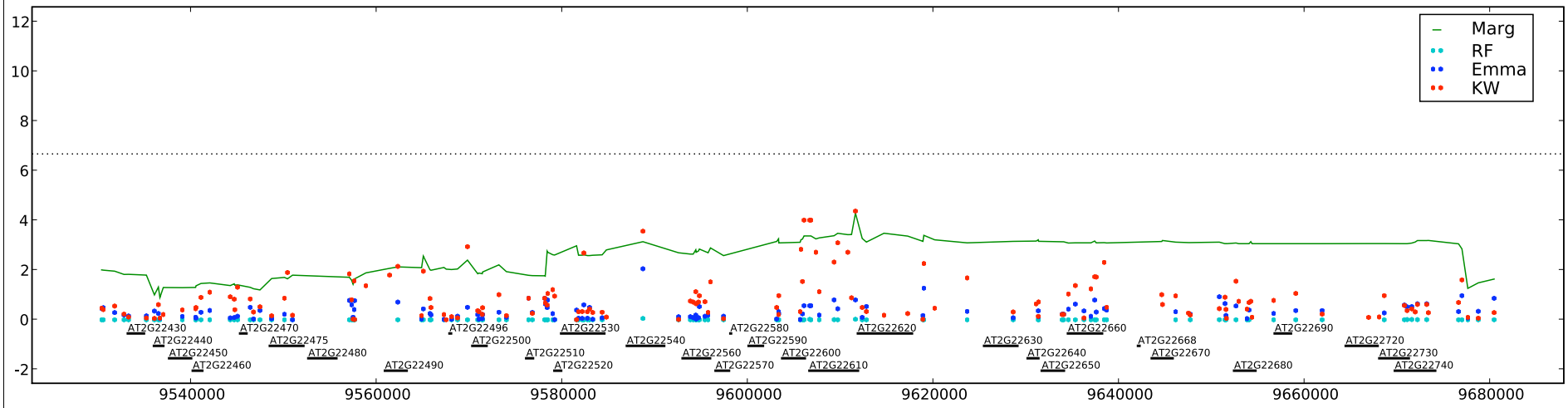


A nice example

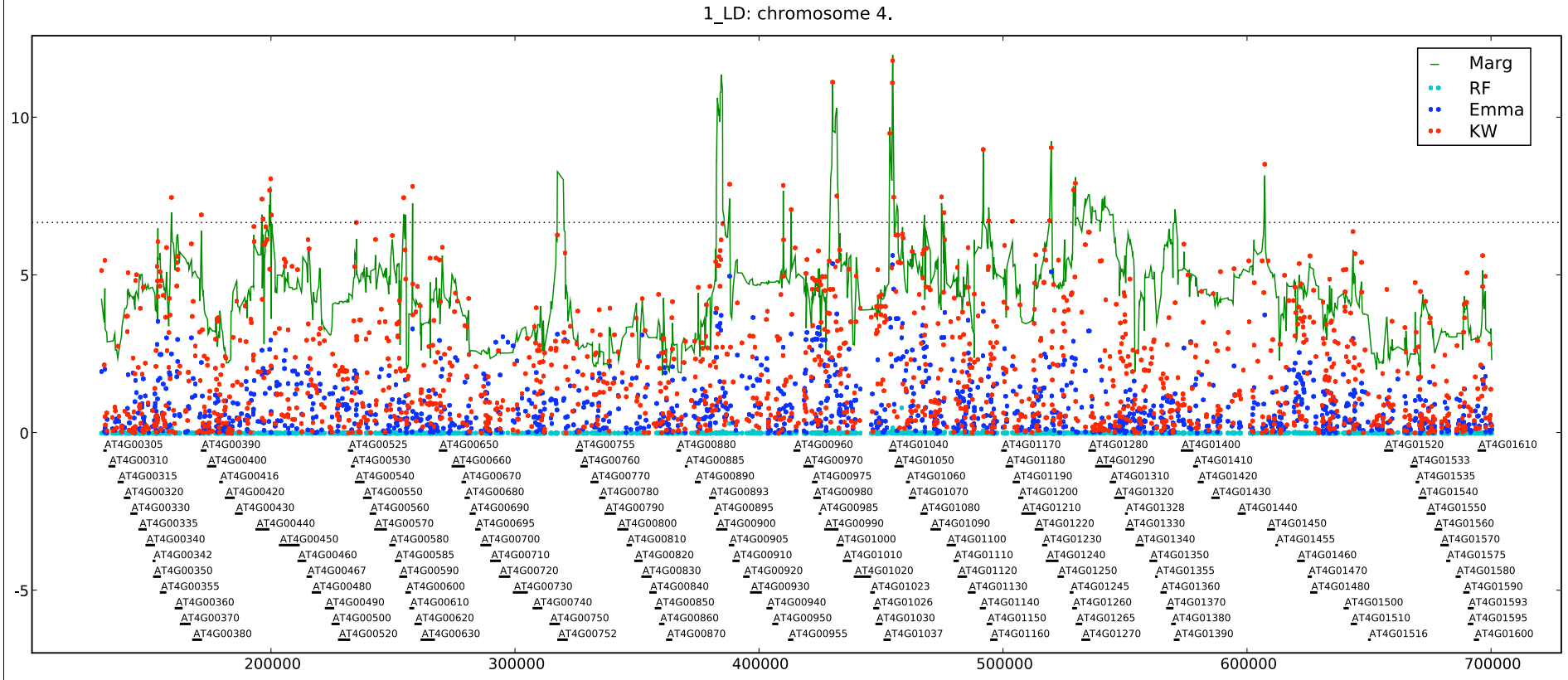
Long days, 18°C



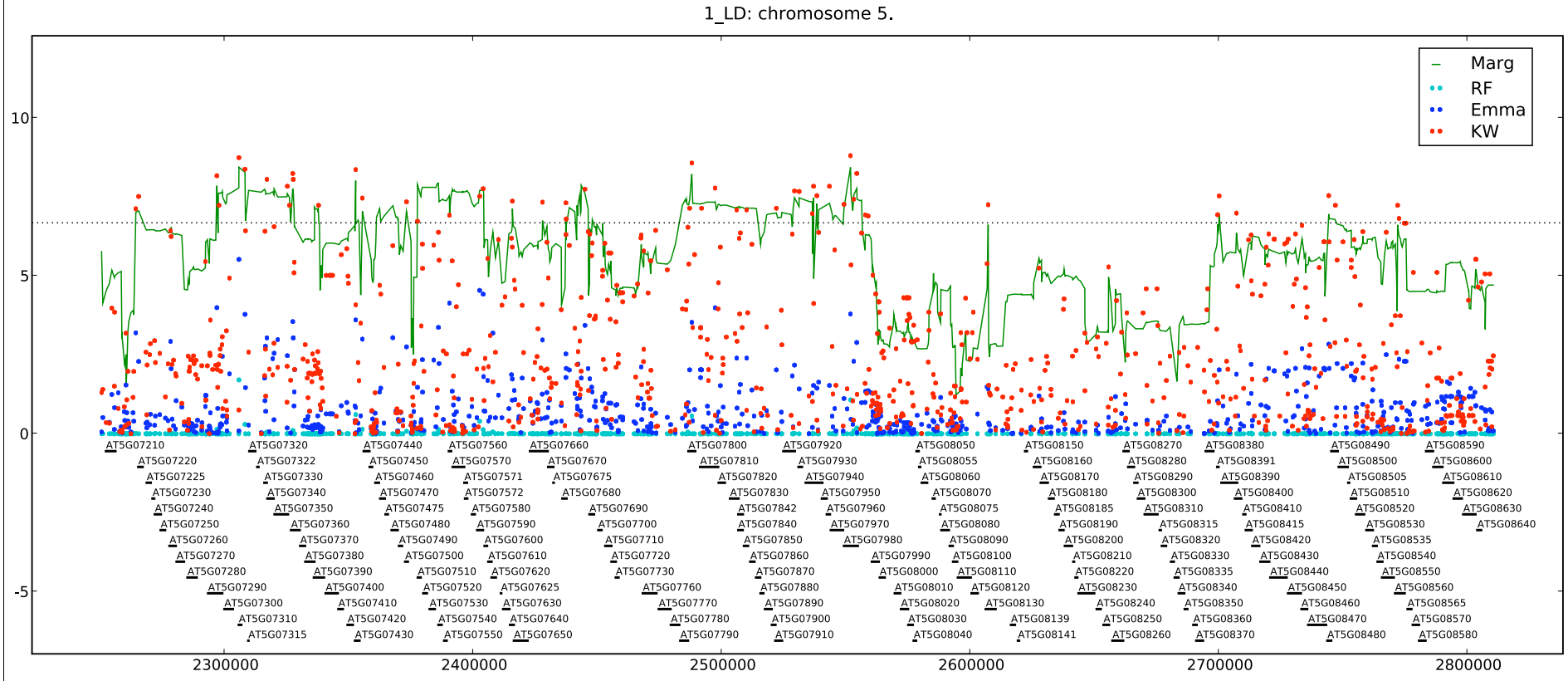
Short days, 18°C + vernalization

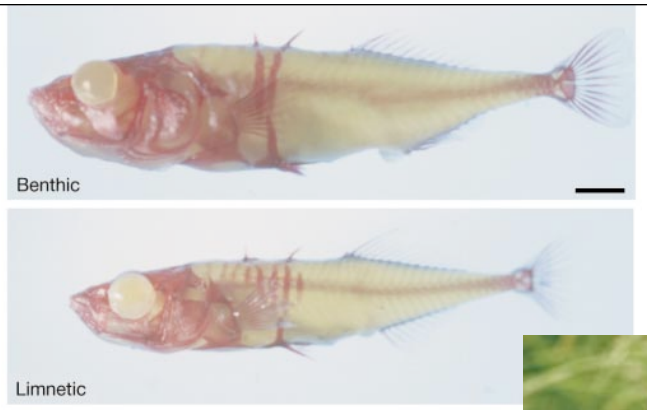


An interesting example



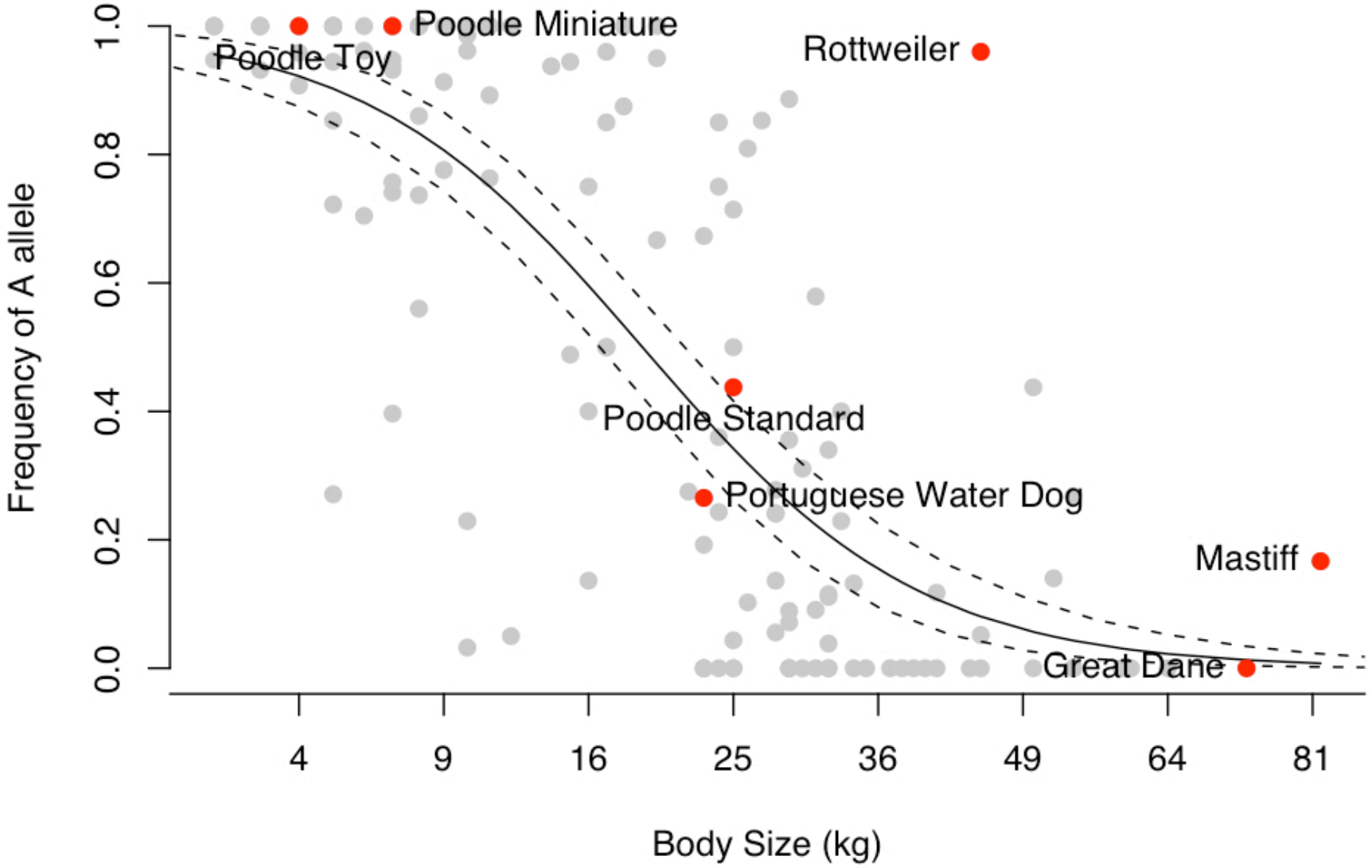
It gets worse (or better)





Confounding *will* be a problem!

A shared allele for small size across dog breeds



Aquilegia

formosa

pubescens

(photo by Scott Hodges, UCSB)

Footprints of selection

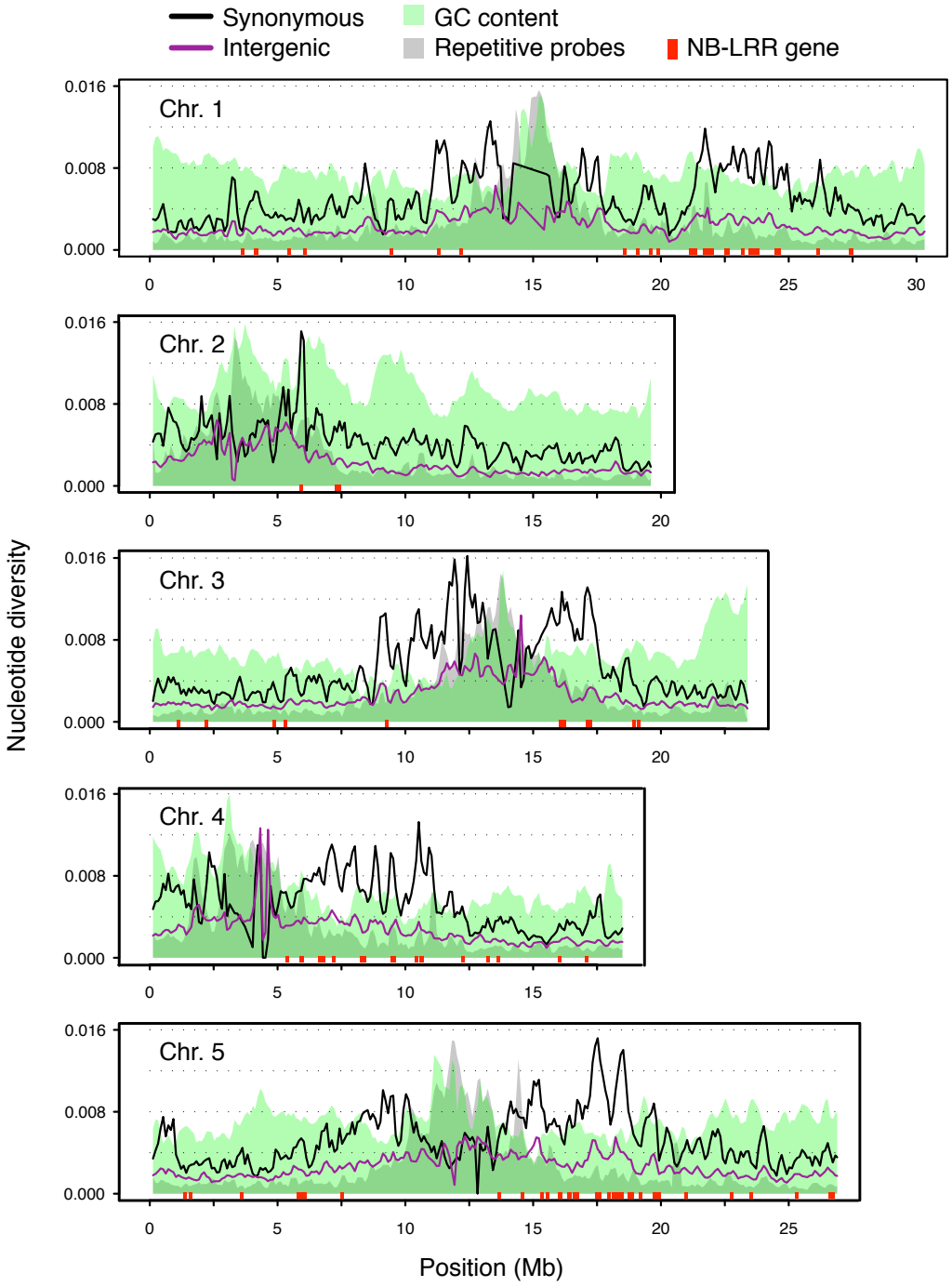
- Understanding the forces governing the genomic pattern of variation is a goal in and of itself
- Can help identify adaptively important polymorphisms

Death to models!

- The old sequencing data from 96 lines revealed a pattern of polymorphism that strongly deviated from that expected under the standard neutral model
- We were unable to fit *any* model to the data: the variability was too great
- The Perlegen data confirmed that our pessimism was not premature...

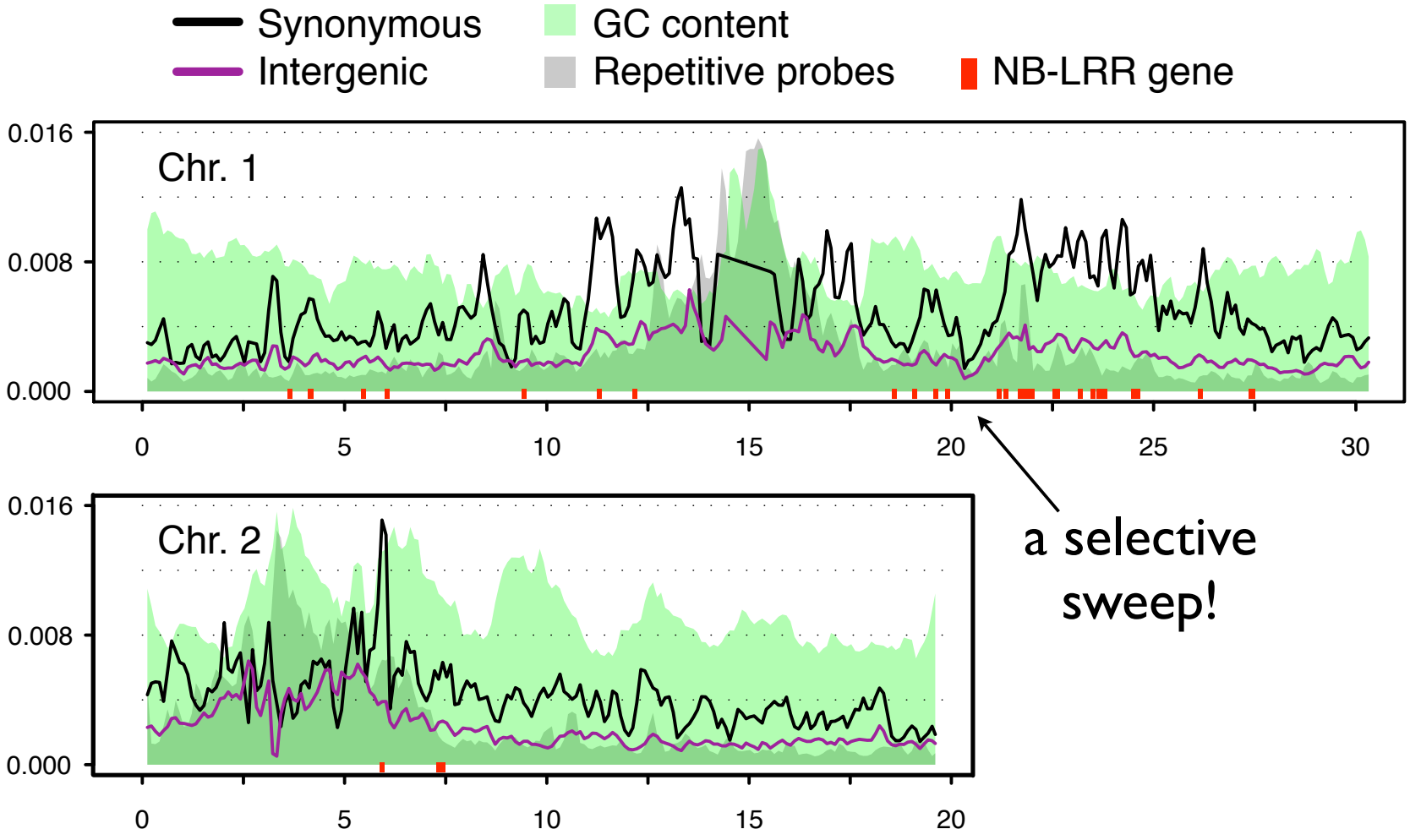
Peaks and valleys of polymorphism

- Selection on linked sites?
- Gene conversion?
- Mutation rate differences?
- Bias?



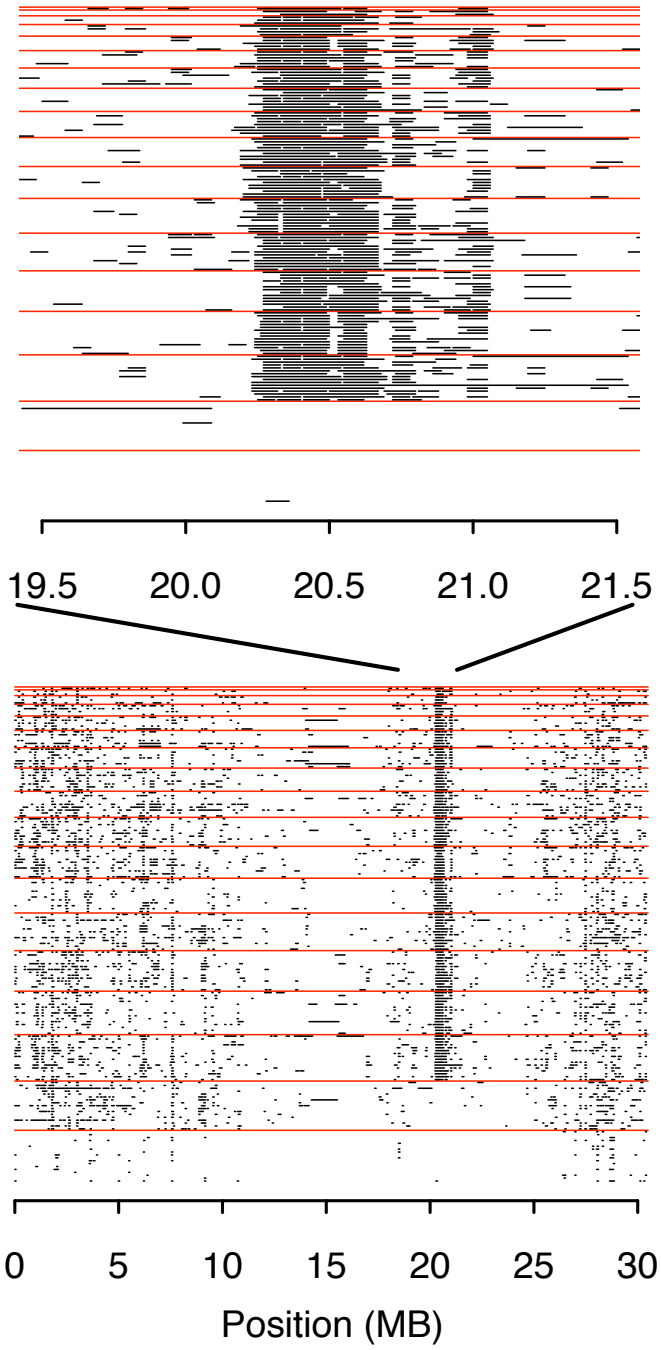
(from Clark *et al.* 2007, *Science*)

Selective sweeps are easier

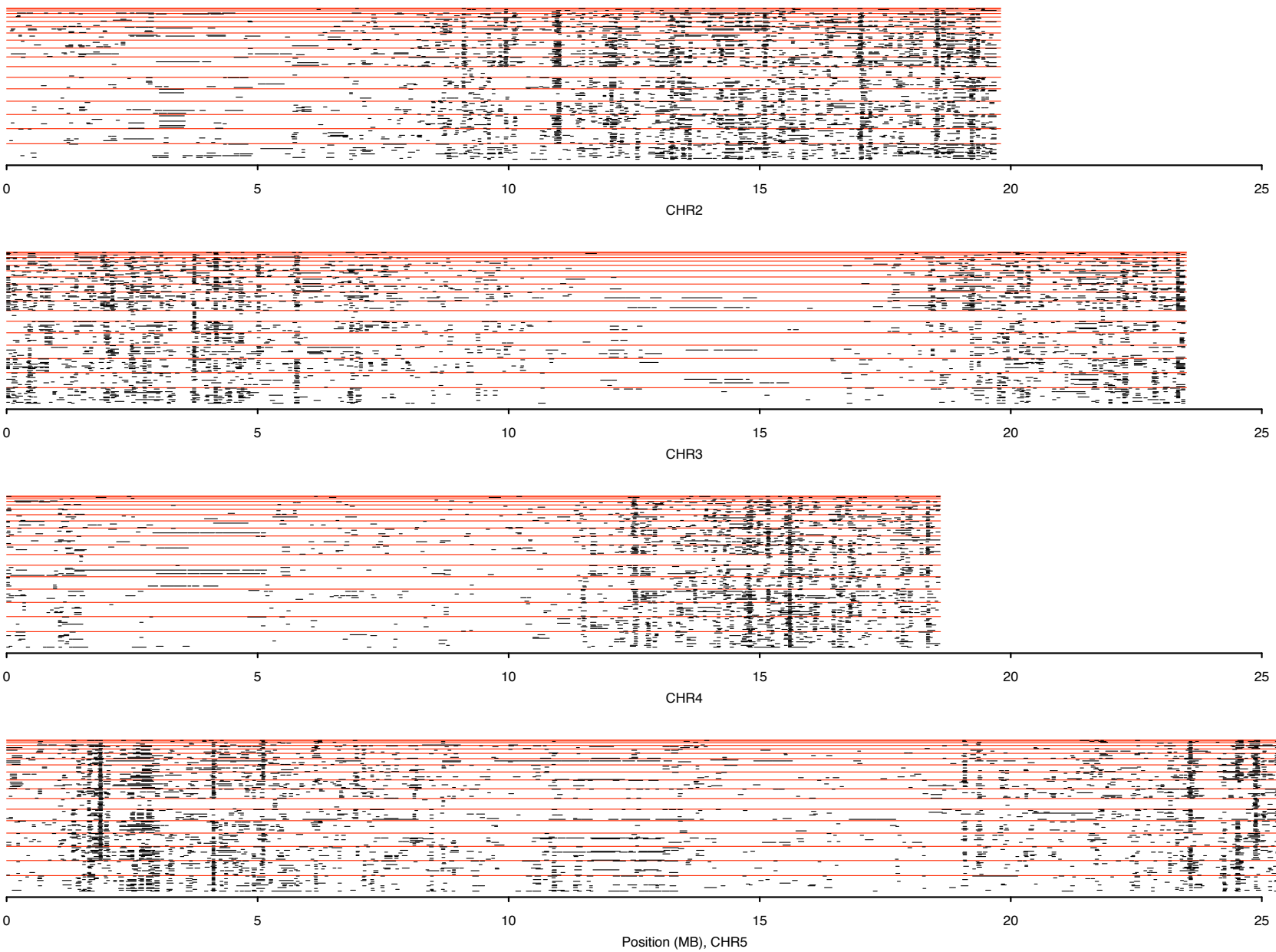


Selective sweep on chromosome 1

- 18/20 accessions share a haplotype that often extends up to 500 kb
- Only Löv-1 and Cvi-0 have escaped the sweep, which appears to be centered on *RPP-27*



(from Clark *et al.* 2007, *Science*)



(from Clark *et al.* 2007, *Science*)

A. thaliana

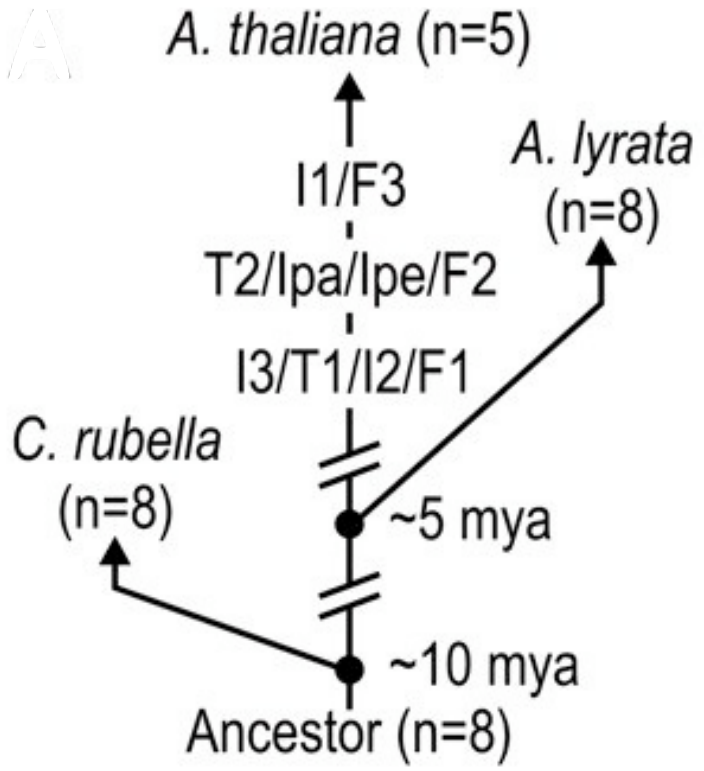
vs.

A. lyrata

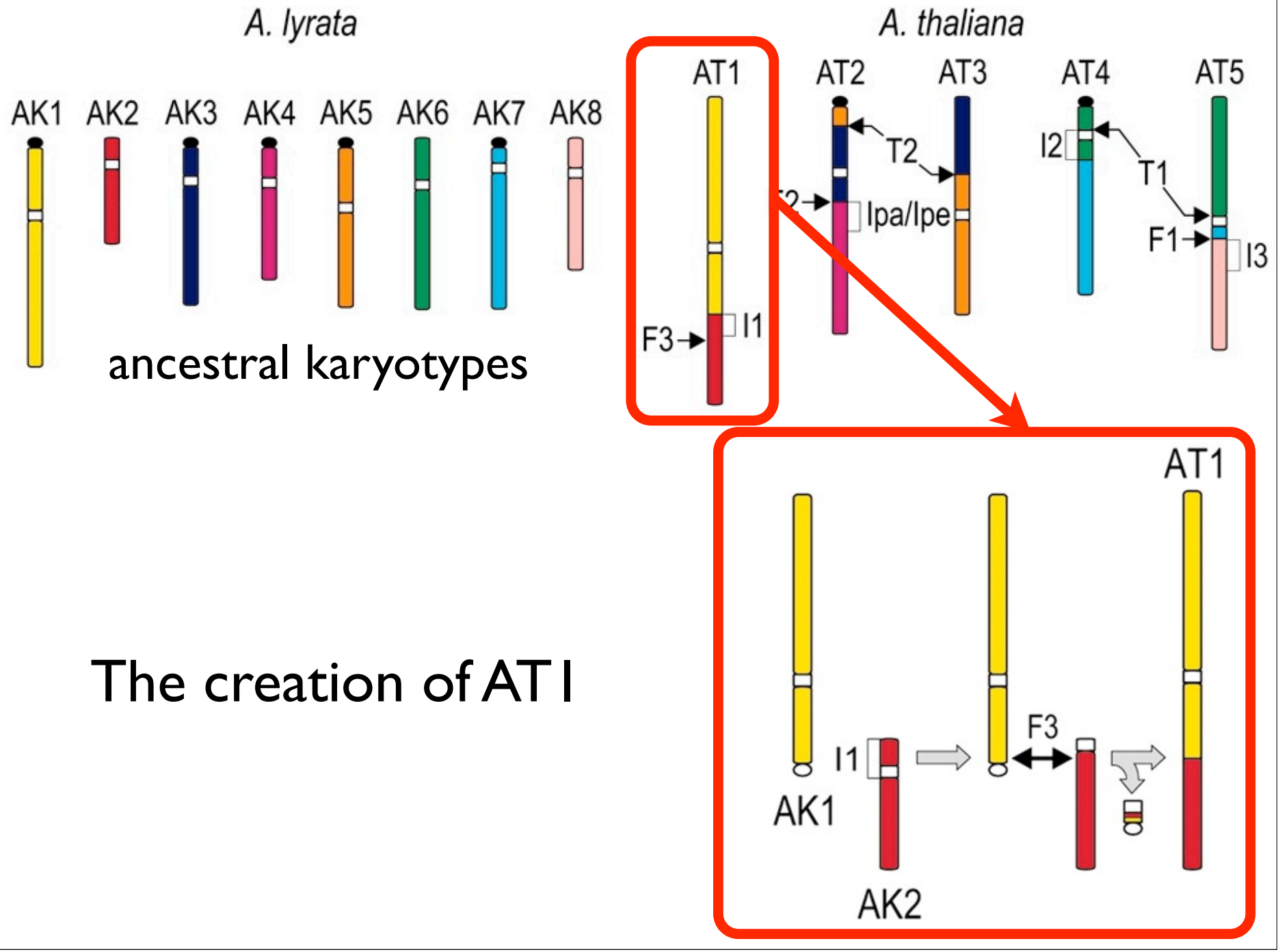
- *A. lyrata*'s genome is ~ 2 X larger than *A. thaliana*'s
- *A. lyrata* has 8 chromosomes; *A. thaliana* 5
- *A. lyrata* is self-incompatible; *A. thaliana* is a selfer

Transition from 8 to 5 chromosomes

- At least 5 inversions
- 2 reciprocal translocations
- 3 chromosomal fusions



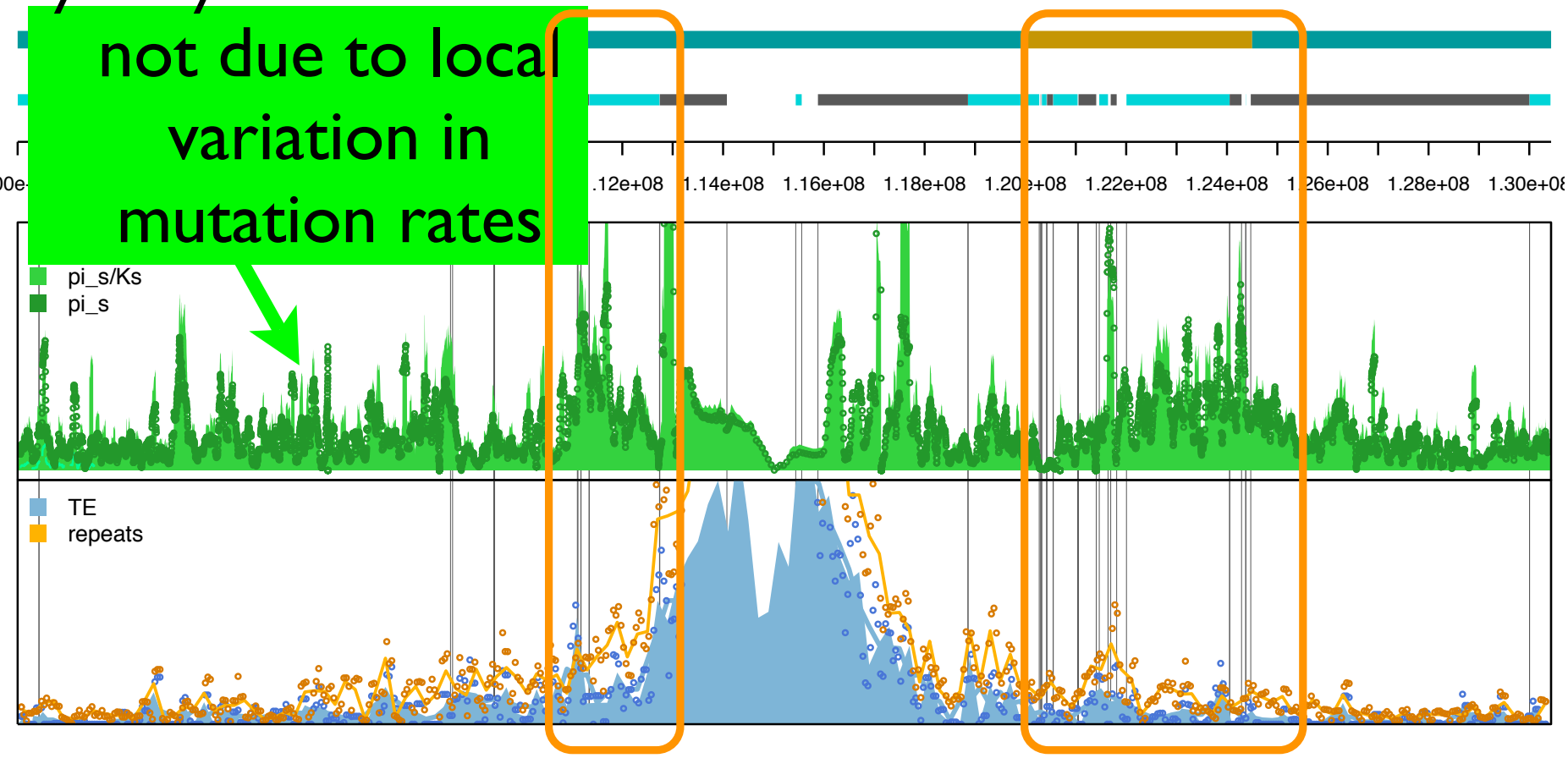
(from Lysak *et al.* 2006, *PNAS*)


A. lyrata
AK1
AK2
AK3
AK4
AK5
AK6
AK7
AK8
ancestral karyotypes
A. thaliana
AT1
AT2
AT3
AT4
AT5
F3
I1
T2
Ipa/Ipe
I2
T1
F1
I3
The creation of AT1
AT1
I1
F3
AK1
AK2

AT I

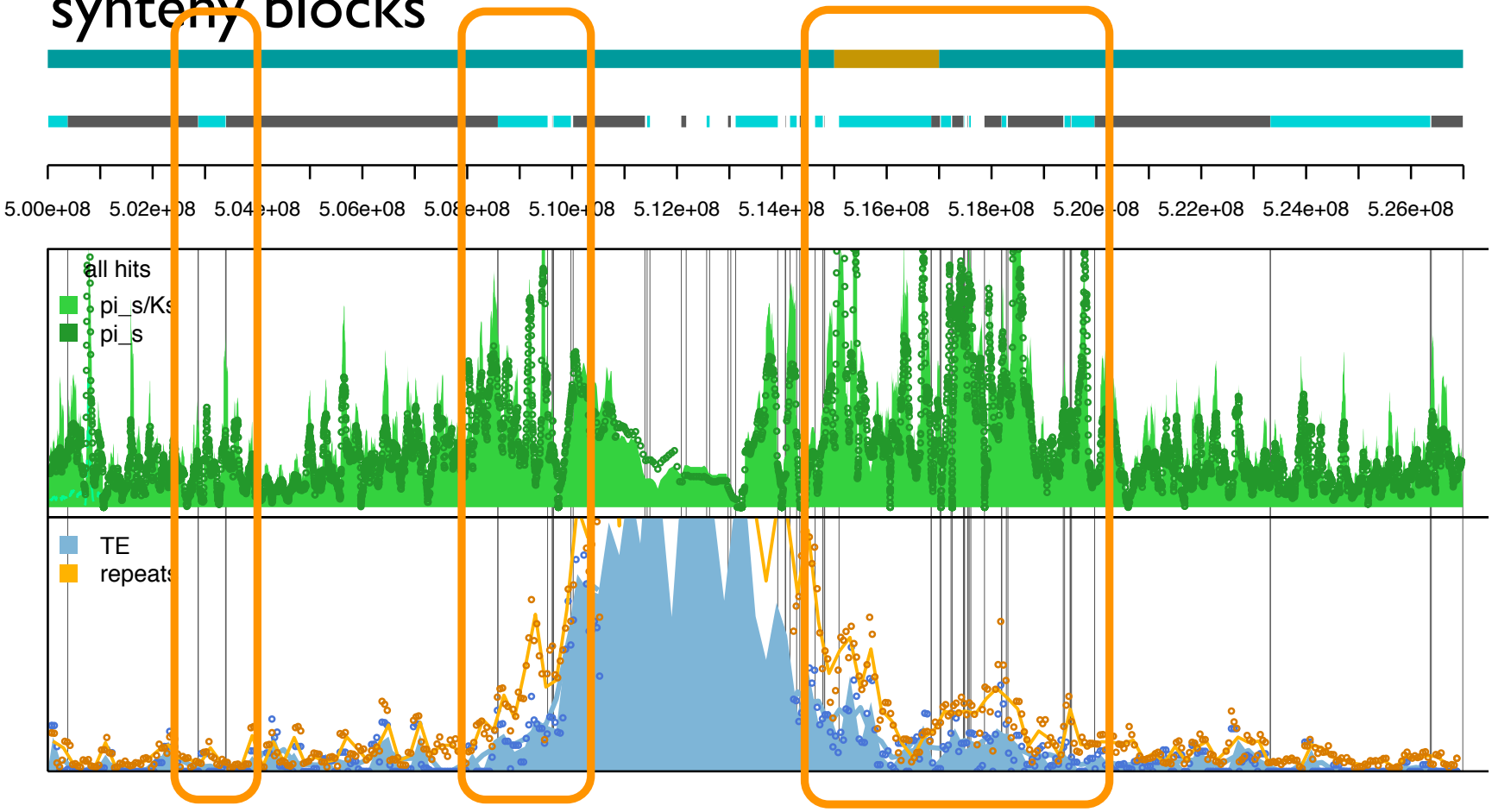
synteny blocks

not due to local variation in mutation rates

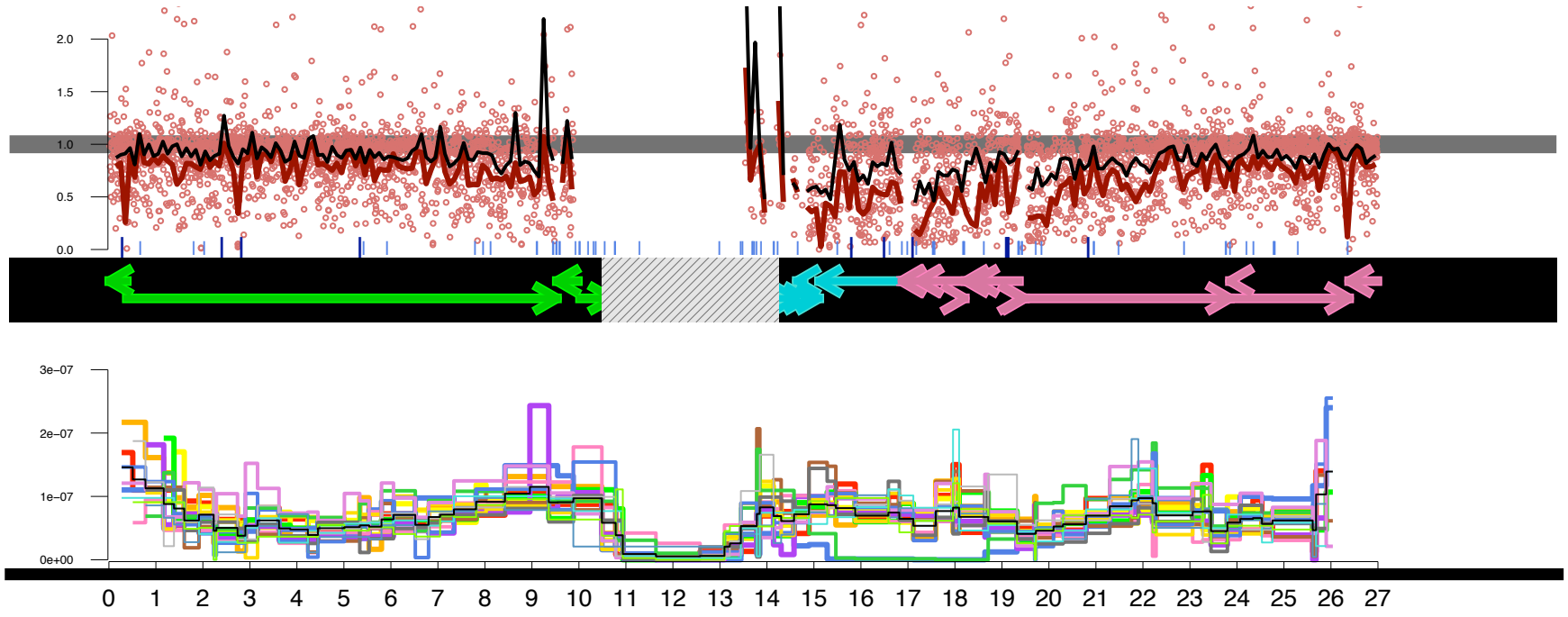


AT5

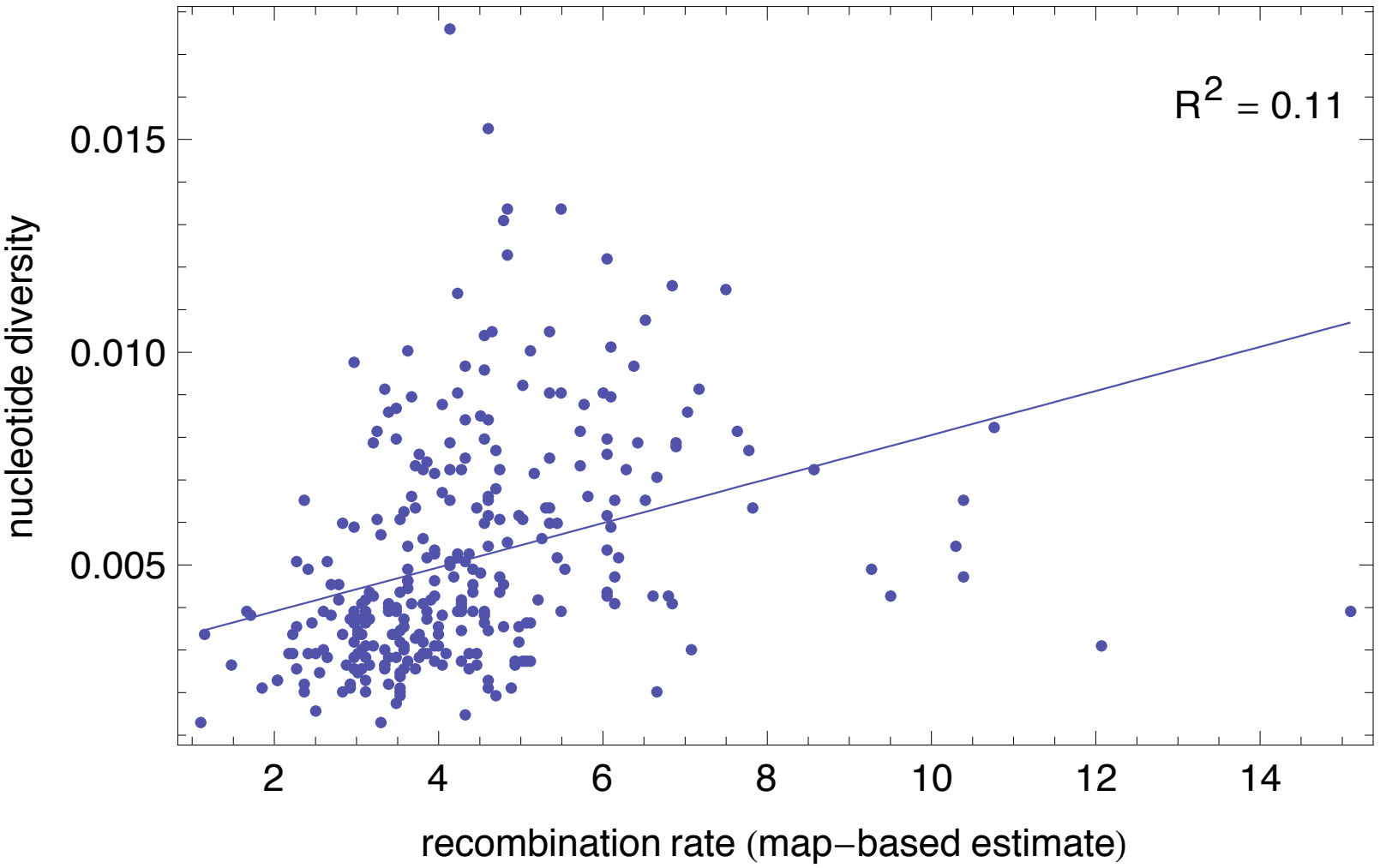
synteny blocks



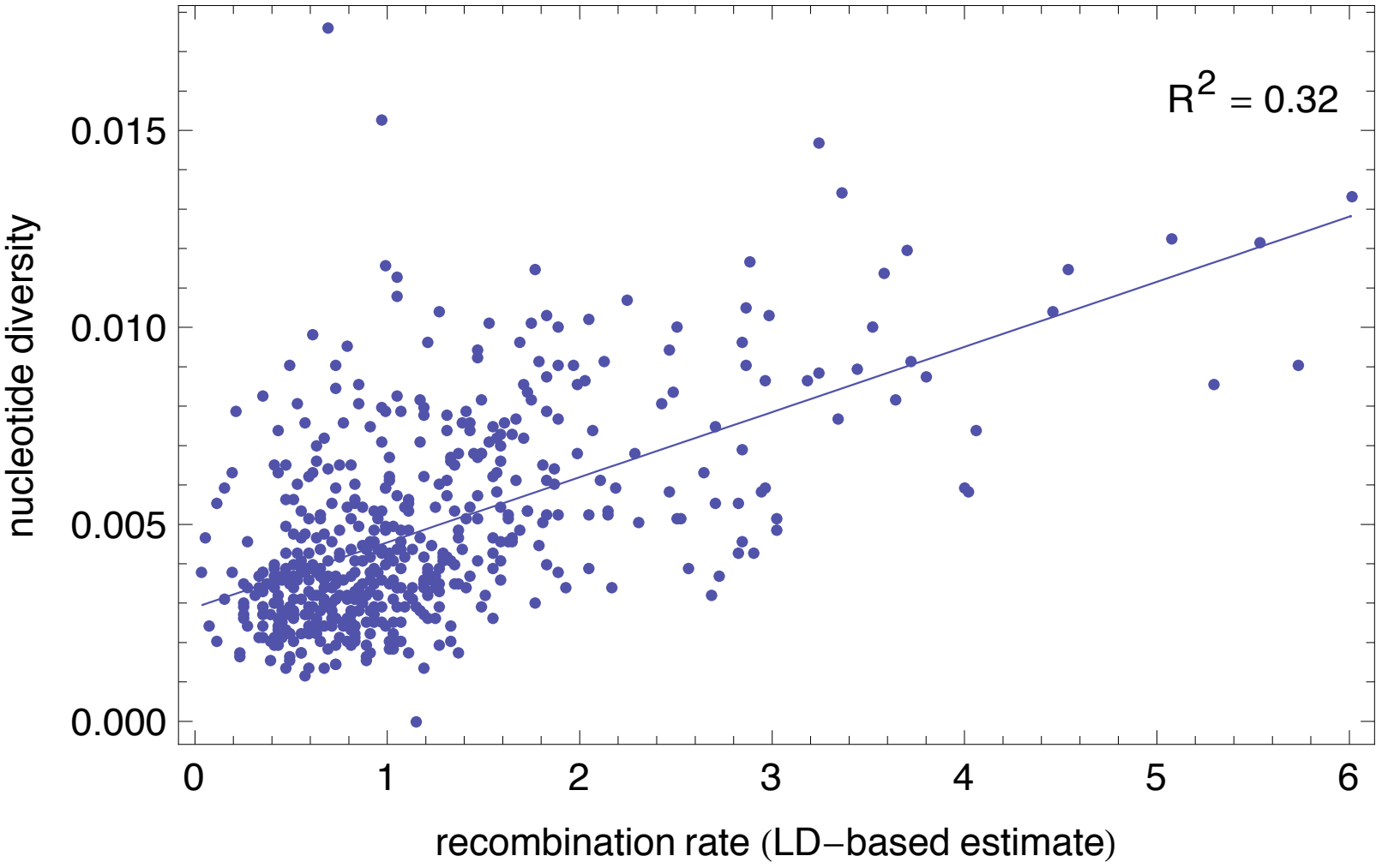
There have been changes at many levels — and the process is still under way



Death to the Neutral Theory!



Death to the Neutral Theory!



Prospect

- By the end of the year, over 1,000 lines will have been genotyped with the 250k chip
- In a couple of years, these lines will also have been sequenced
- We are systematically crossing lines
- Add "dynamic" information — epigenome, transcriptome, proteome — and system biology is finally meeting natural variation!