Quantile-based Permutation Thresholds for QTL Hotspots

Brian S Yandell and Elias Chaibub Neto

17 March 2012

Fisher on inference

We may at once admit that any inference from the particular to the general must be attended with some degree of uncertainty, but this is not the same as to admit that such inference cannot be absolutely rigorous, for the nature and degree of the uncertainty may itself be capable of rigorous expression.

Sir Ronald A Fisher(1935)

The Design of Experiments

Why study hotspots?

How do genotypes affect phenotypes?

genotypes = DNA markers for an individual

phenotypes = traits measured on an individual

(clinical traits, thousands of mRNA expression levels)

QTL hotspots = genomic locations affecting many traits

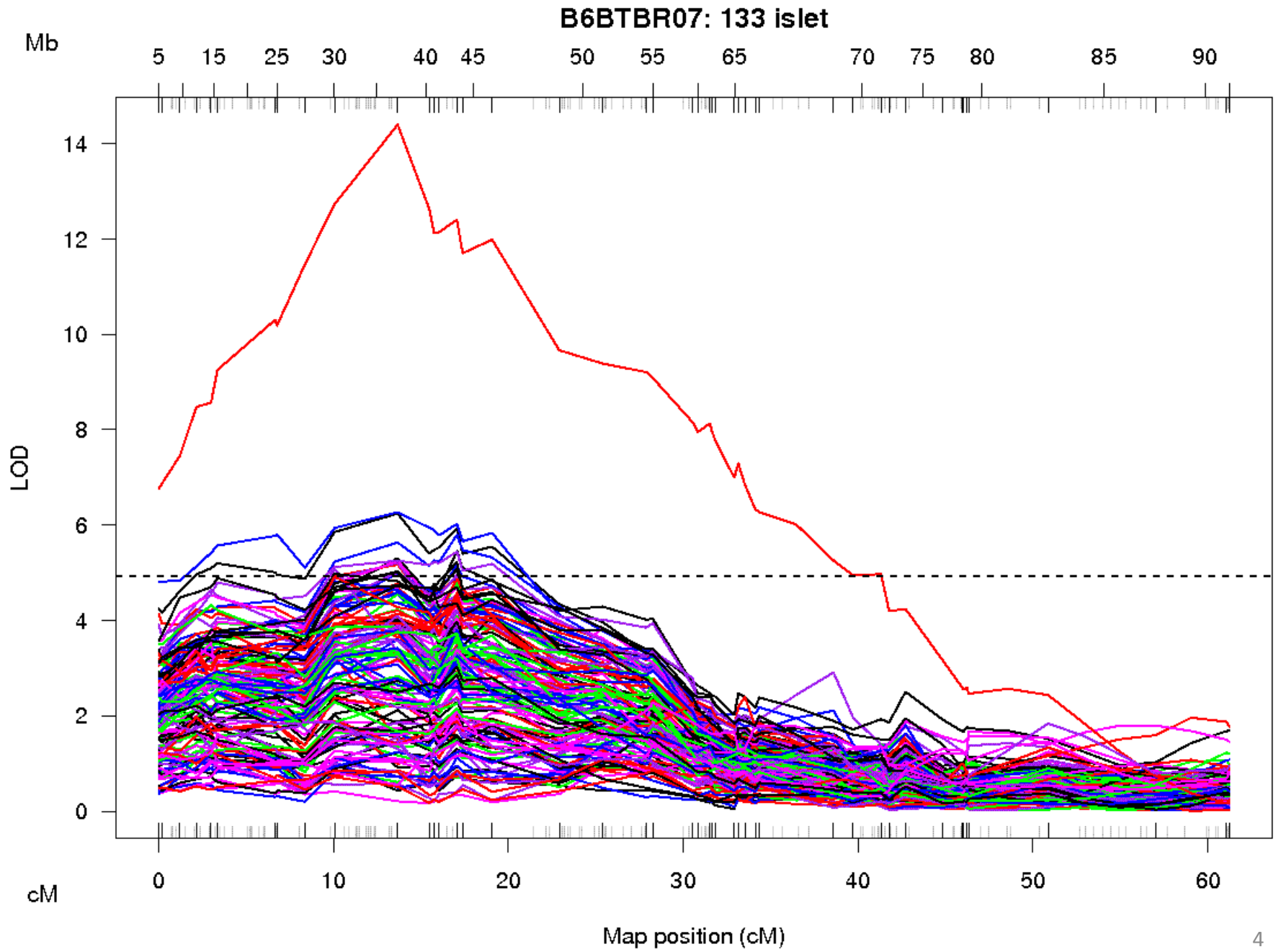
common feature in genetical genomics studies

biologically interesting--may harbor critical regulators

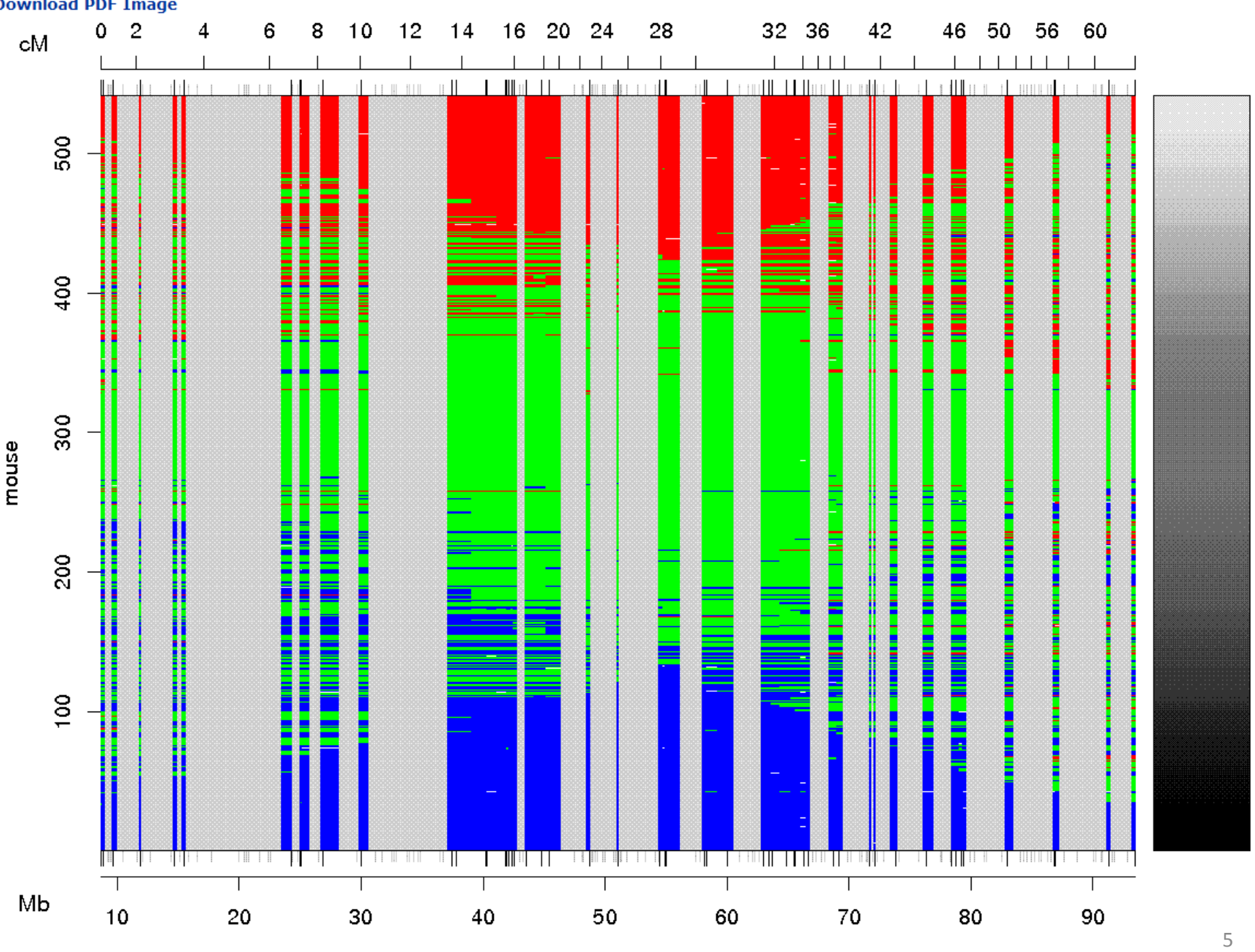
But are these hotspots real? Or are they spurious or random?

non-genetic correlation from other environmental factors

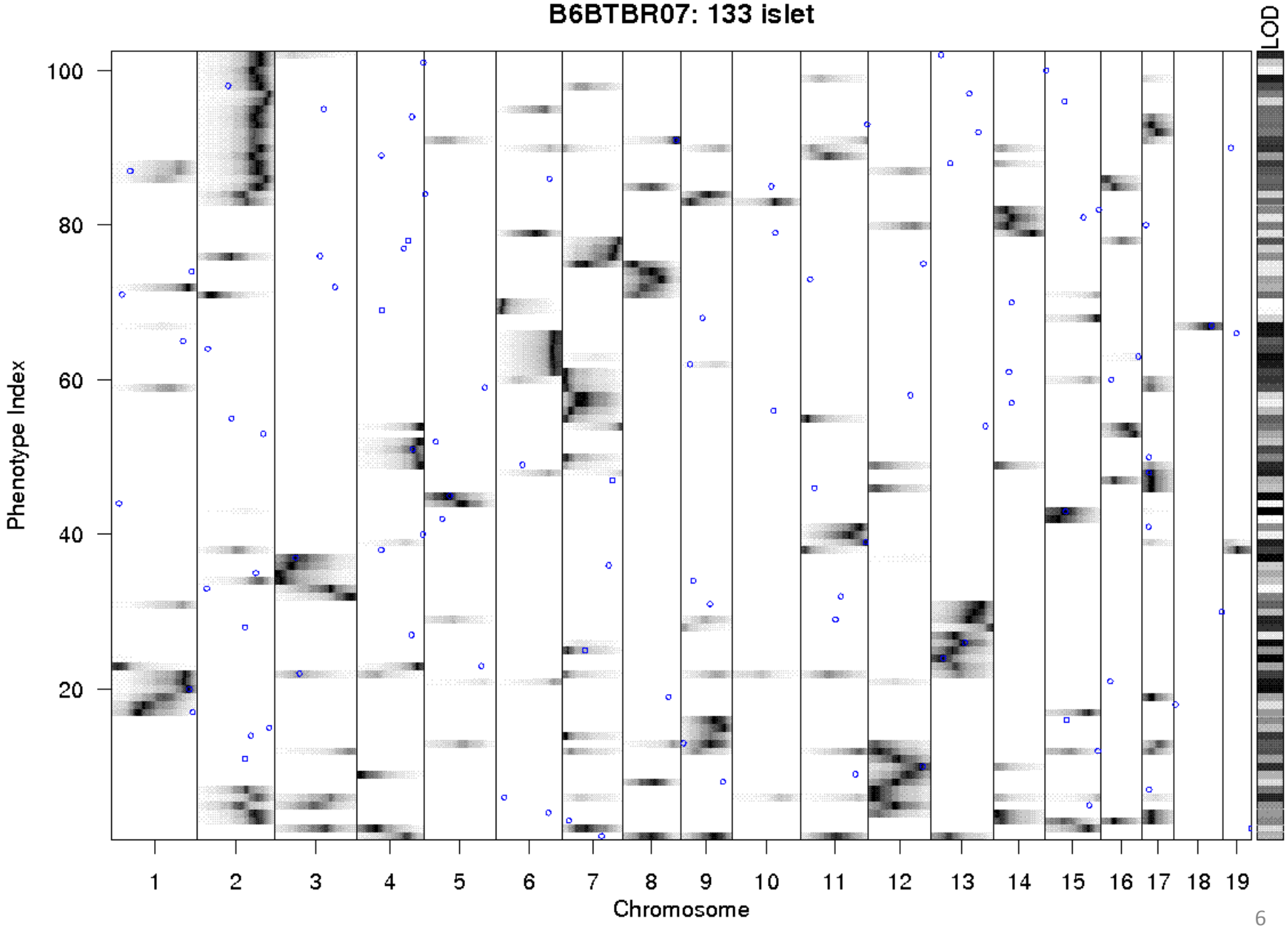
Download PDF Image

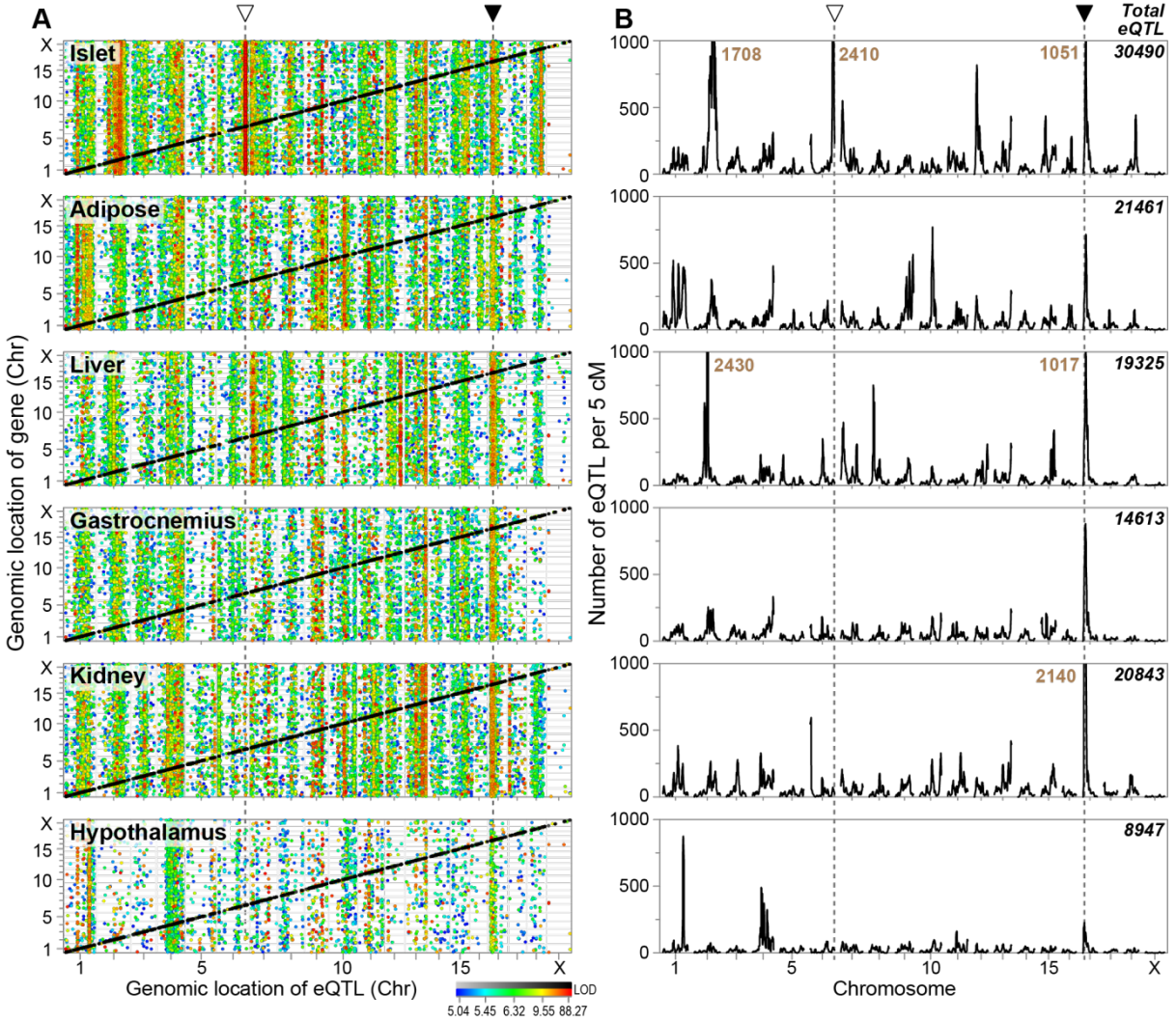


Download PDF Image



Download PDF Image





Genetic architecture of gene expression in 6 tissues.

A Tissue-specific panels illustrate the relationship between the genomic location of a gene (y-axis) to where that gene's mRNA shows an eQTL (LOD > 5), as a function of genome position (x-axis). Circles represent eQTLs that showed either cis-linkage (black) or trans-linkage (colored) according to LOD score. Genomic hot spots, where many eQTLs map in trans, are apparent as vertical bands that show either tissue selectivity (e.g., Chr 6 in the islet, ▽) or are present in all tissues (e.g., Chr 17, ▼). **B** The total number of eQTLs identified in 5 cM genomic windows is plotted for each tissue; total eQTLs for all positions is shown in upper right corner for each panel. The peak number of eQTLs exceeding 1000 per 5 cM is shown for islets (Chrs 2, 6 and 17), liver (Chrs 2 and 17) and kidney (Chr 17).

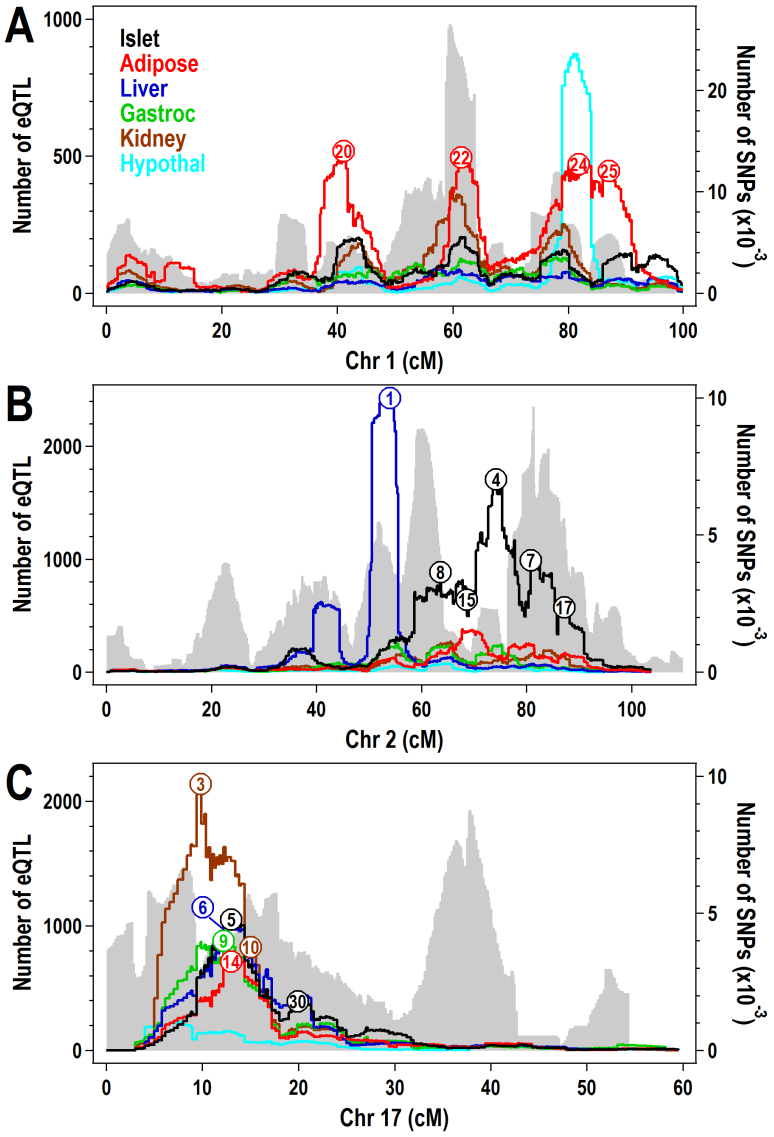


Figure 4 Tissue-specific hotspots with eQTL and SNP architecture for Chrs 1, 2 and 17.

The number of eQTLs for each tissue (left axis) and the number of SNPs between B6 and BTBR (right axis) that were identified within a 5 cM genomic window is shown for Chr 1 (**A**), Chr 2 (**B**) Chr 17 (**C**). The location of tissue-specific hotspots are identified by their number corresponding to that in Table 1. eQTL and SNP architecture is shown for all chromosomes in supplementary material.

How large a hotspot is large?

recently proposed empirical test

Brietling et al. Jansen (2008)

hotspot = count traits above LOD threshold

LOD = rescaled likelihood ratio ~ F statistic

assess null distribution with permutation test

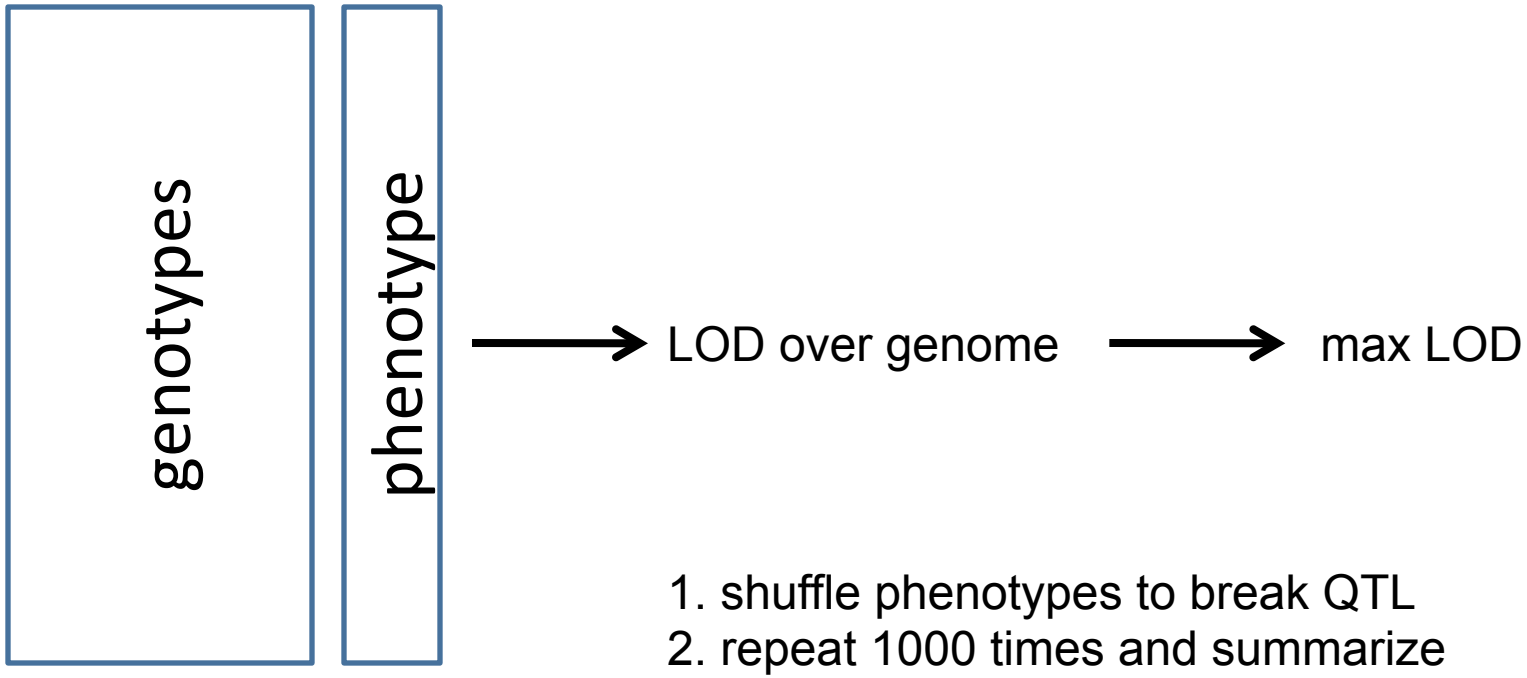
extension of Churchill and Doerge (1994)

extension of Fisher's permutation t-test

Single trait permutation threshold *T* Churchill Doerge (1994)

- Null distribution of max LOD
  - Permute single trait separate from genotype
  - Find max LOD over genome
  - Repeat 1000 times
- Find 95% permutation threshold *T*
- Identify interested peaks above *T* in data
- Controls genome-wide error rate (GWER)
  - Chance of detecting at least on peak above *T*

Single trait permutation schema

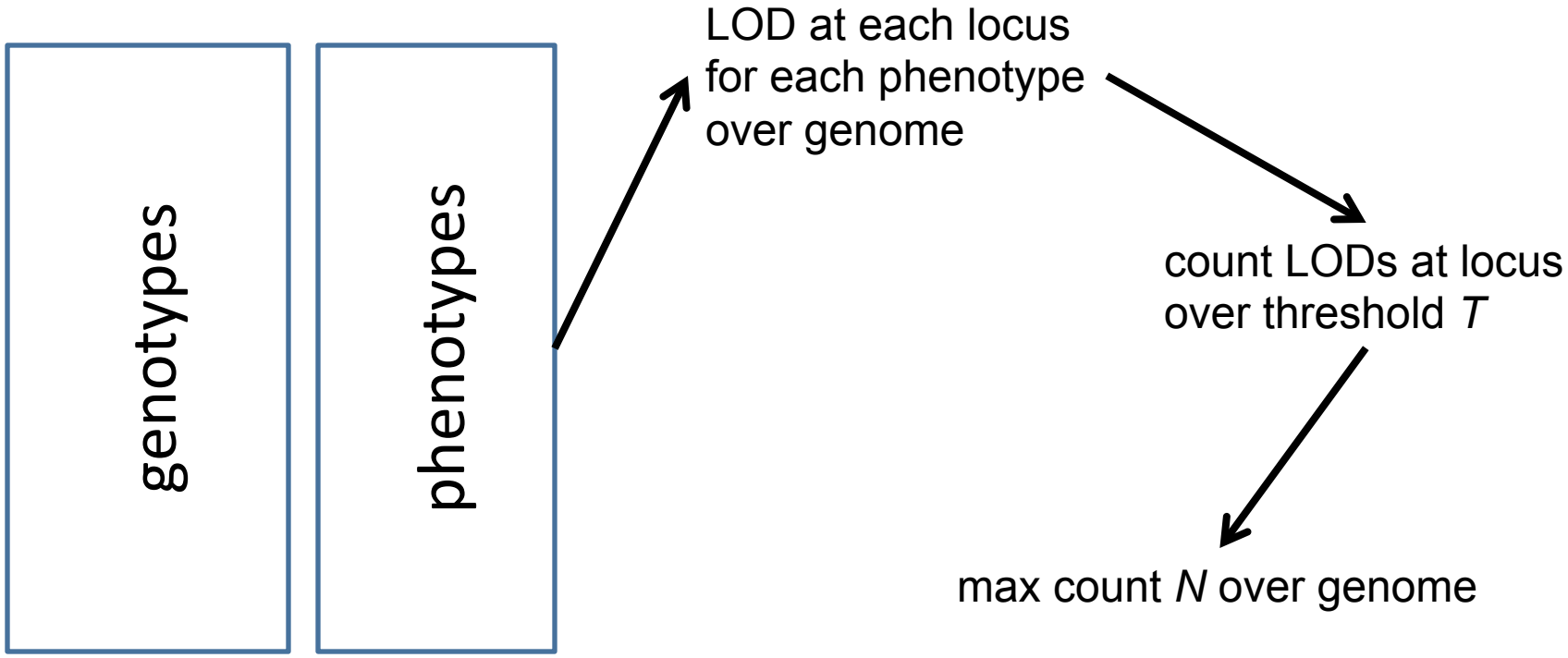


1. shuffle phenotypes to break QTL
2. repeat 1000 times and summarize

Hotspot count threshold *N*(*T*) Breitling et al. Jansen (2008)

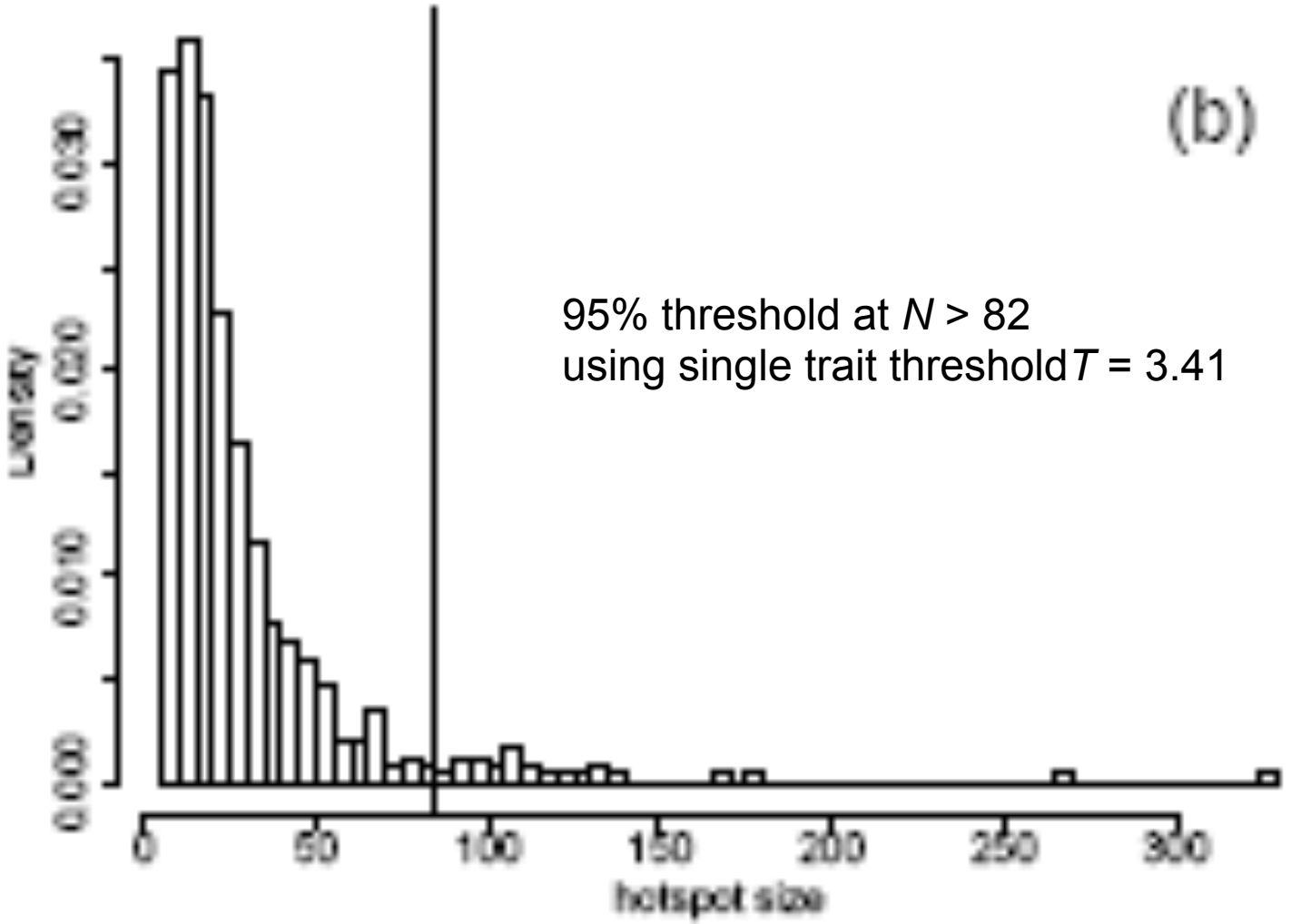
- Null distribution of max count above *T*
  - Find single-trait 95% LOD threshold *T*
  - Find max count of traits with LODs above *T*
  - Repeat 1000 times
- Find 95% count permutation threshold *N*
- Identify counts of LODs above *T* in data
  - Locus-specific counts identify hotspots
- Controls GWER in some way

Hotspot permutation schema

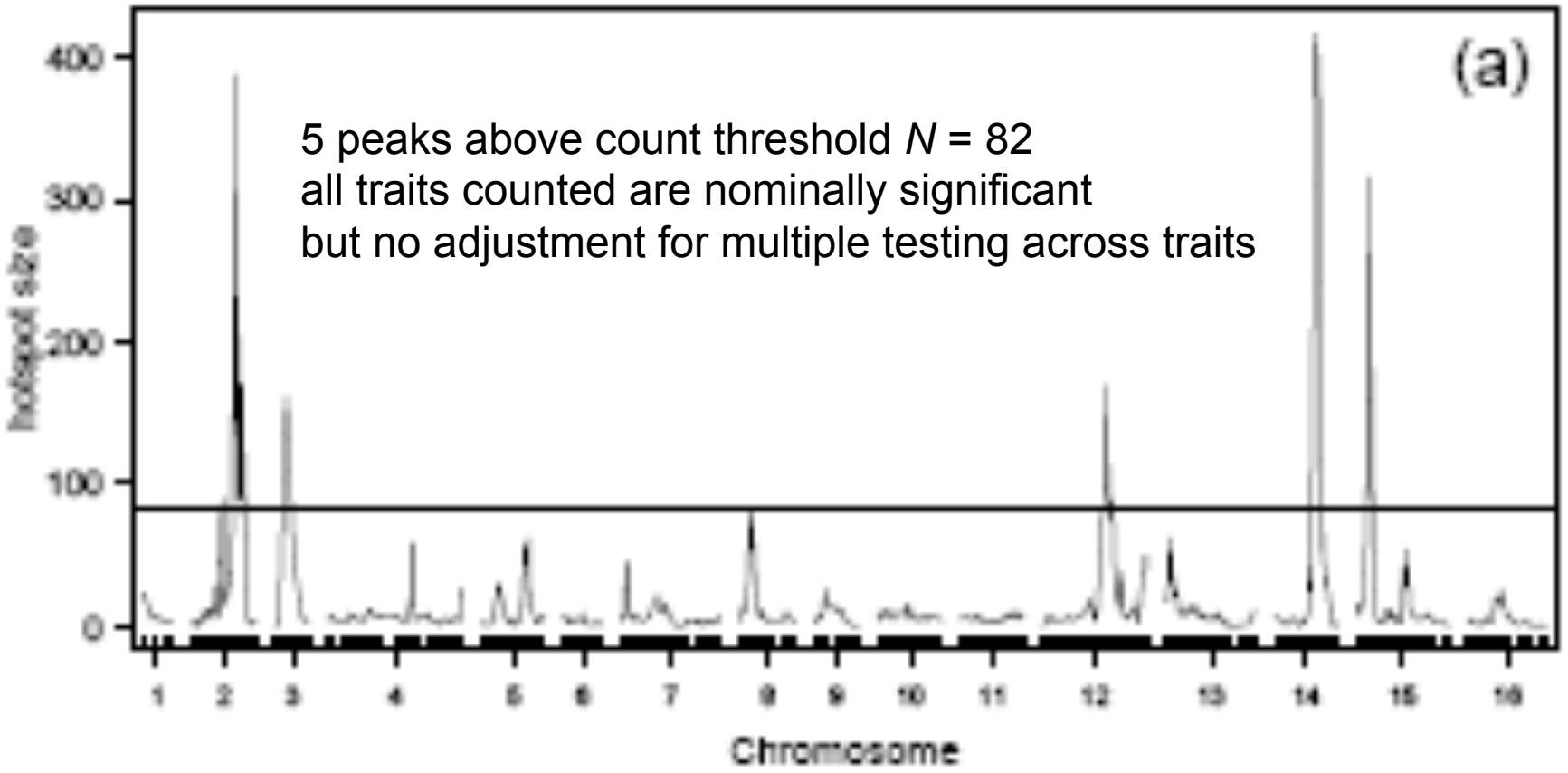


1. shuffle phenotypes by row to break QTL, keep correlation
2. repeat 1000 times and summarize

spurious hotspot permutation histogram for hotspot size above 1-trait threshold



Hotspot sizes based on count of LODs above single-trait threshold

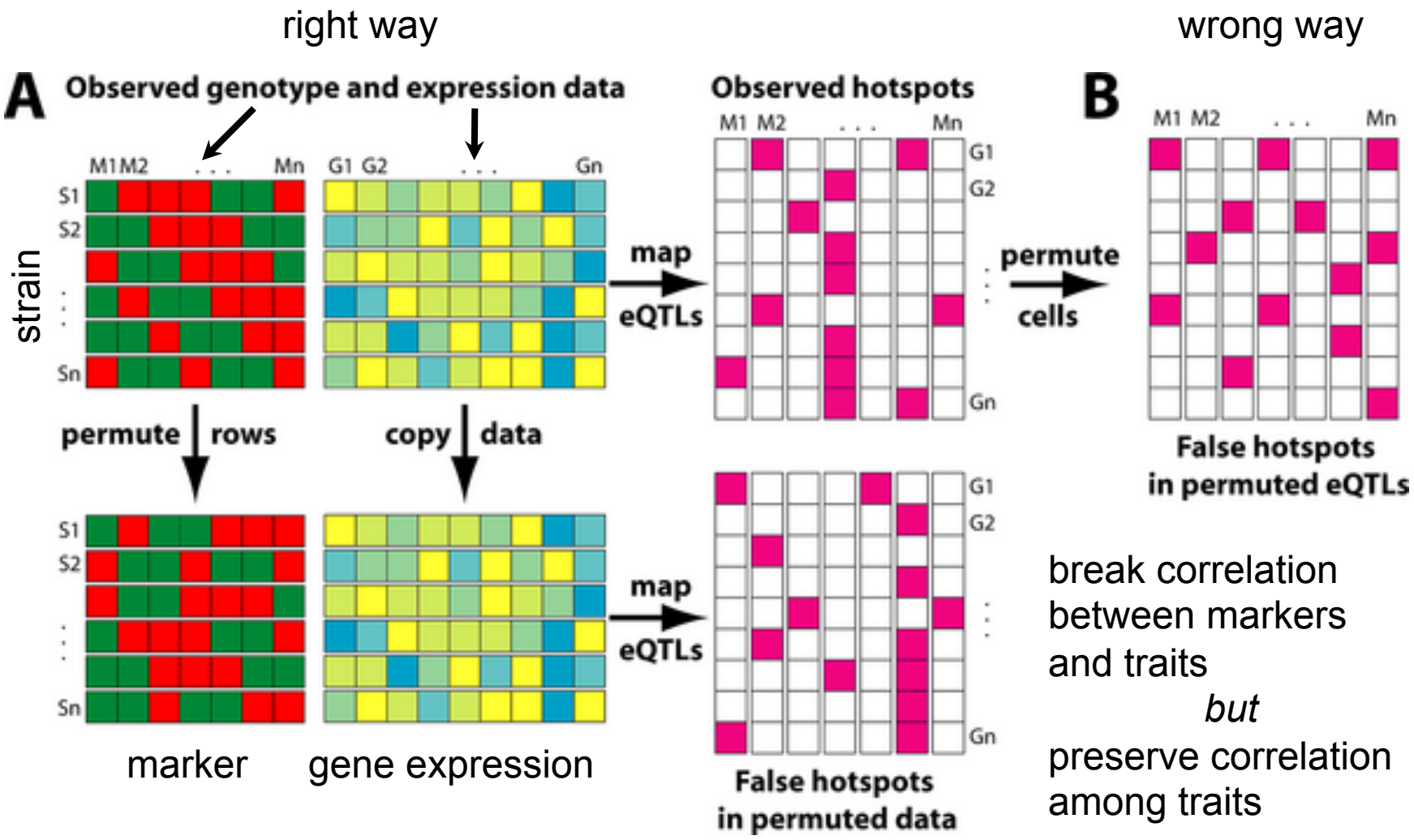


hotspot permutation test
(Breitling et al. Jansen 2008 *PLoS Genetics*)

- for original dataset and each permuted set:
  - Set single trait LOD threshold *T*
    - Could use Churchill-Doerge (1994) permutations
  - Count number of traits (*N*) with LOD above *T*
    - Do this at every marker (or pseudomarker)
    - Probably want to smooth counts somewhat
- find count with at most 5% of permuted sets above (critical value) as count threshold
- conclude original counts above threshold are real

permutation across traits

(Breitling et al. Jansen 2008 *PLoS Genetics*)



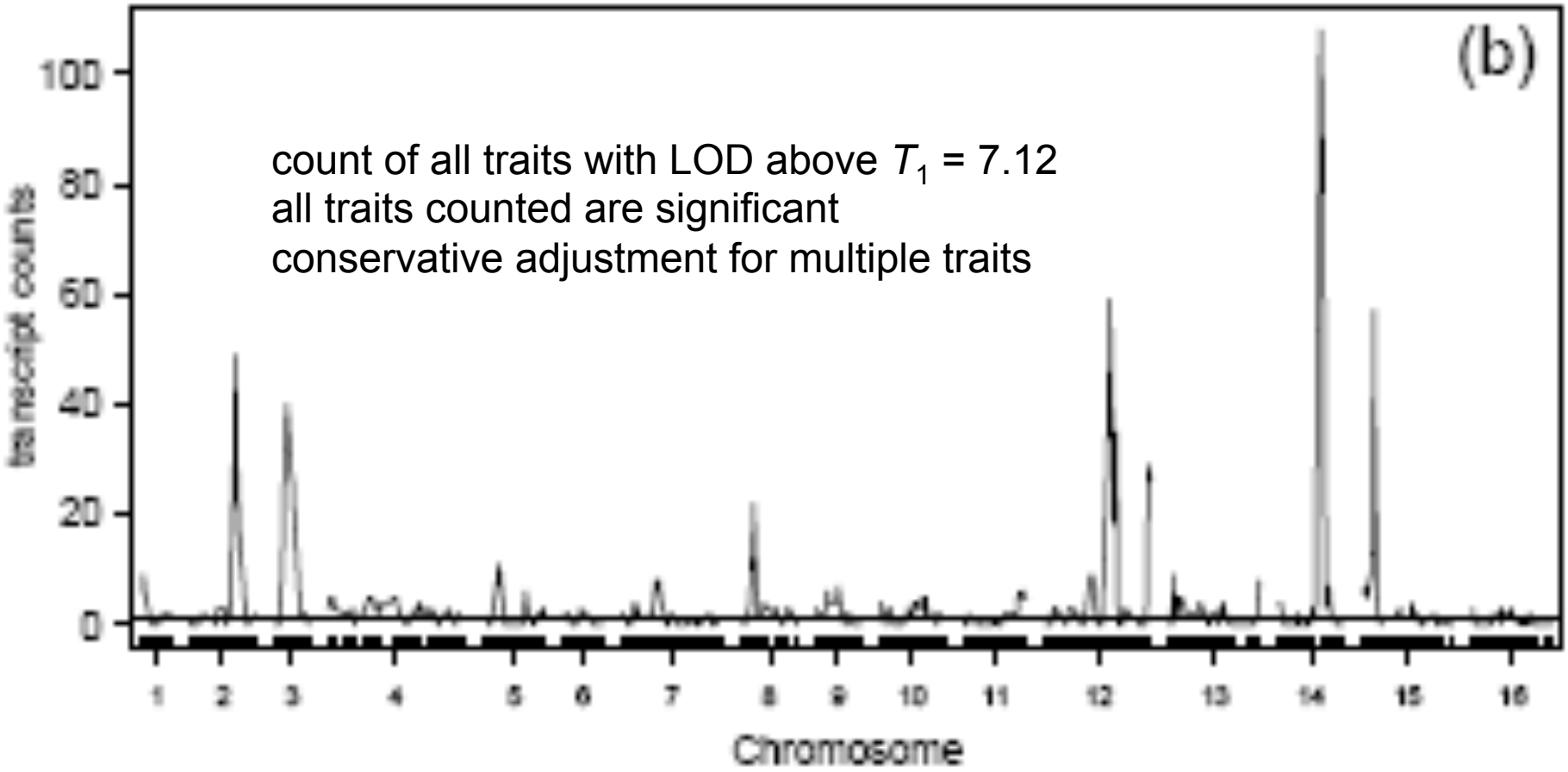
quality vs. quantity in hotspots (Chaibub Neto et al. in review)

- detecting single trait with very large LOD
  - control FWER across genome
  - control FWER across all traits
- finding small "hotspots" with significant traits
  - all with large LODs
  - could indicate a strongly disrupted signal pathway
- sliding LOD threshold across hotspot sizes

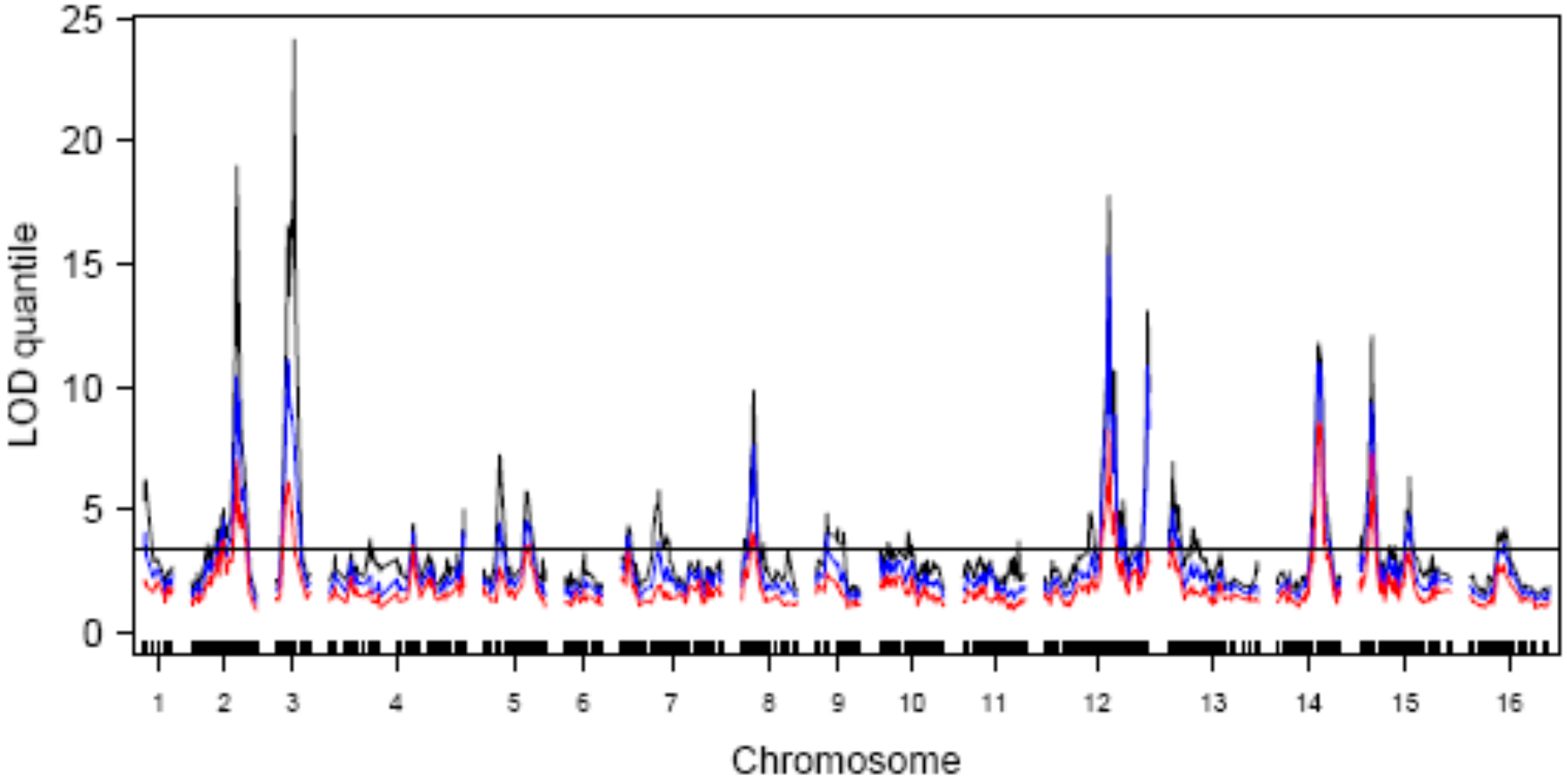
Rethinking the approach

- Breitling et al. depends highly on *T*
- Threshold *T* based on single trait
  - but interested in multiple correlated traits
- want to control hotspot GWER (hGWER$_N$)
  - chance of detecting at least one spurious hotspot of size *N* or larger
- *N = 1*
  - chance of detecting at least 1 peak above threshold across all traits and whole genome
  - Use permutation null distribution of maximum LOD scores across all transcripts and all genomic locations

Hotspot architecture using multiple trait GWER threshold (T_1=7.12)



locus-specific LOD quantiles in data for 10(black), 20(blue), 50(red) traits

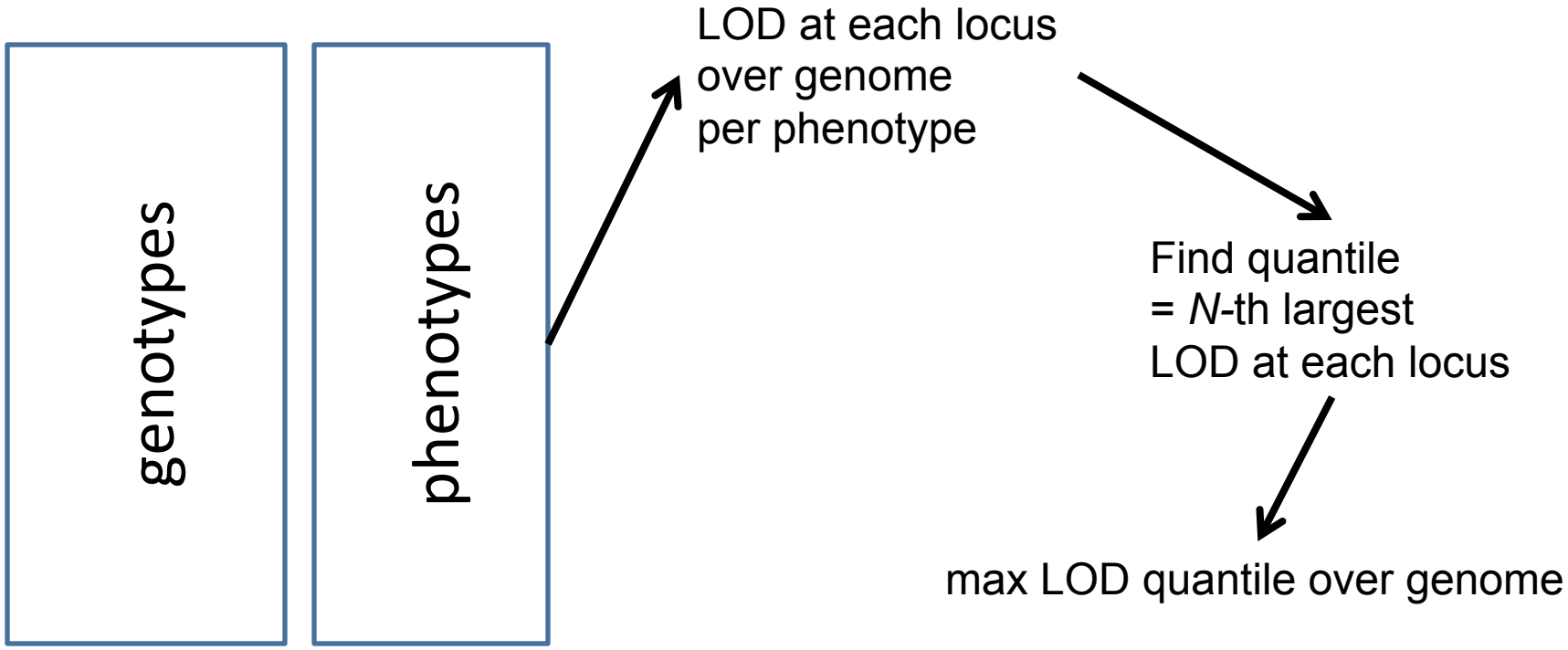


locus-specific LOD quantiles

- Quantile: what is LOD value for which at least 10 (or 20 or 50) traits are at above it?
- Breitling hotspots (chr 2,3,12,14,15)
  - have many traits with high LODs
- Chromosome max LOD quantile by trait count

color	count	chr 3	chr 8	chr 12	chr 14
black	10	24	10	18	12
blue	20	11	8	15	11
red	50	6	4	9	9

Hotspot permutation revisited

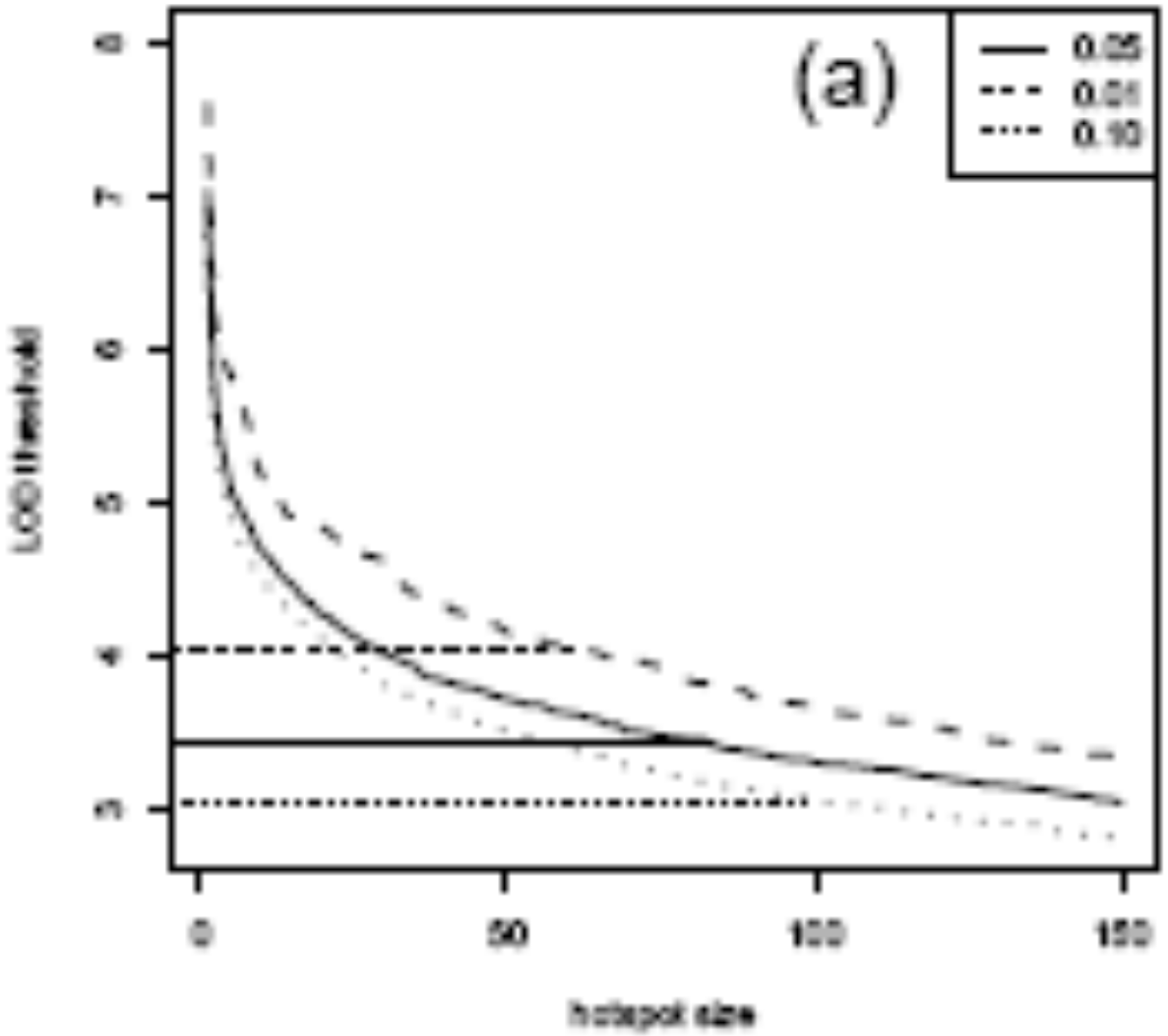


1. shuffle phenotypes by row to break QTL, keep correlation
2. repeat 1000 times and summarize

Tail distribution of LOD quantiles and size-specific thresholds

- What is locus-specific (spurious) hotspot?
  - all traits in hotspot have LOD above null threshold
- Small spurious hotspots have higher minimum LODs
  - min of 10 values > min of 20 values
- Large spurious hotspots have many small LODs
  - most are below single-trait threshold
- Null thresholds depending on hotspot size
  - Decrease with spurious hotspot size (starting at $N = 1$)
  - Be truncated at single-trait threshold for large sizes
- Chen Storey (2007) studied LOD quantiles
  - For multiple peaks on a single trait

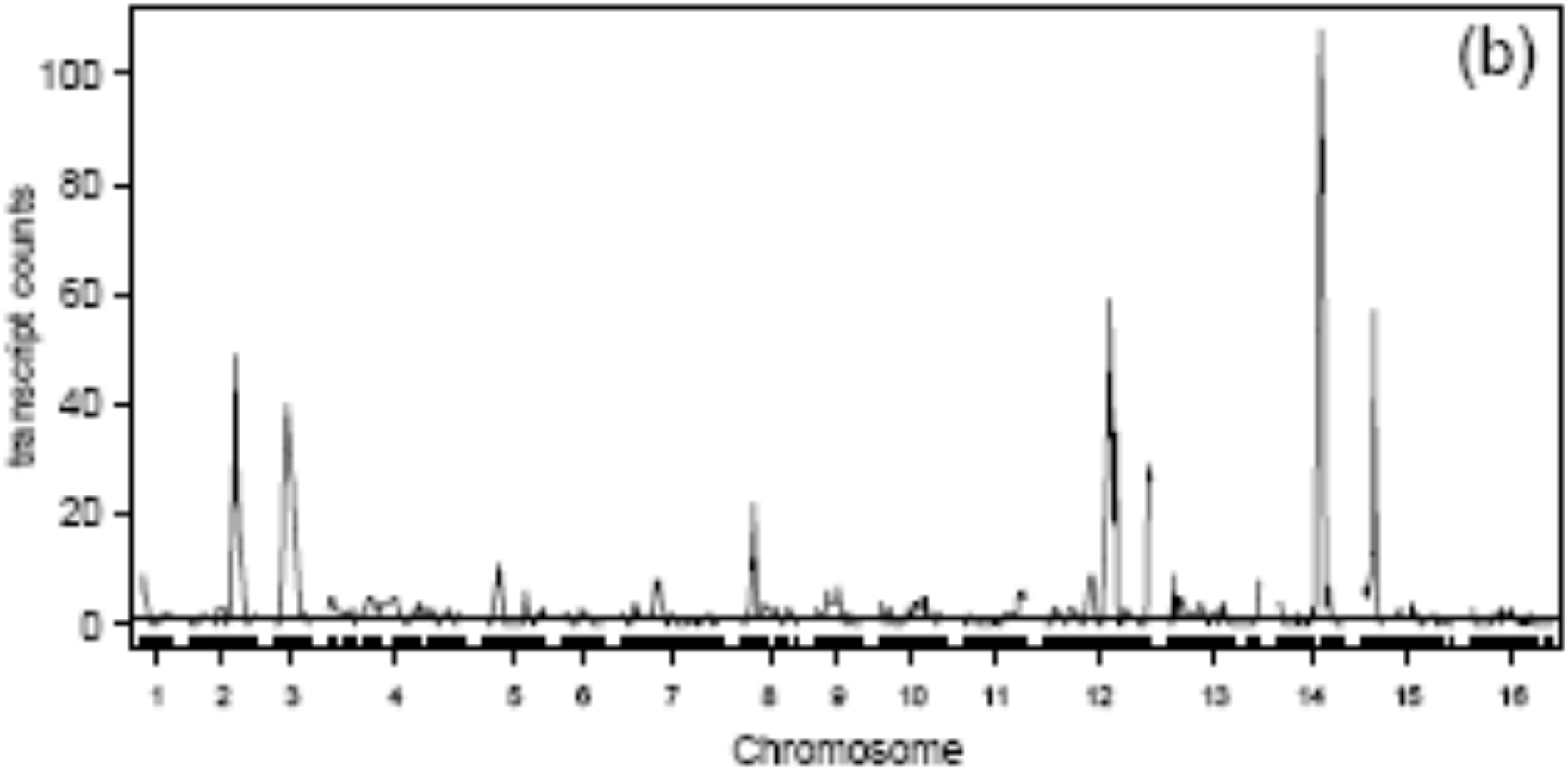
genome-wide LOD permutation threshold vs. spurious hotspot size



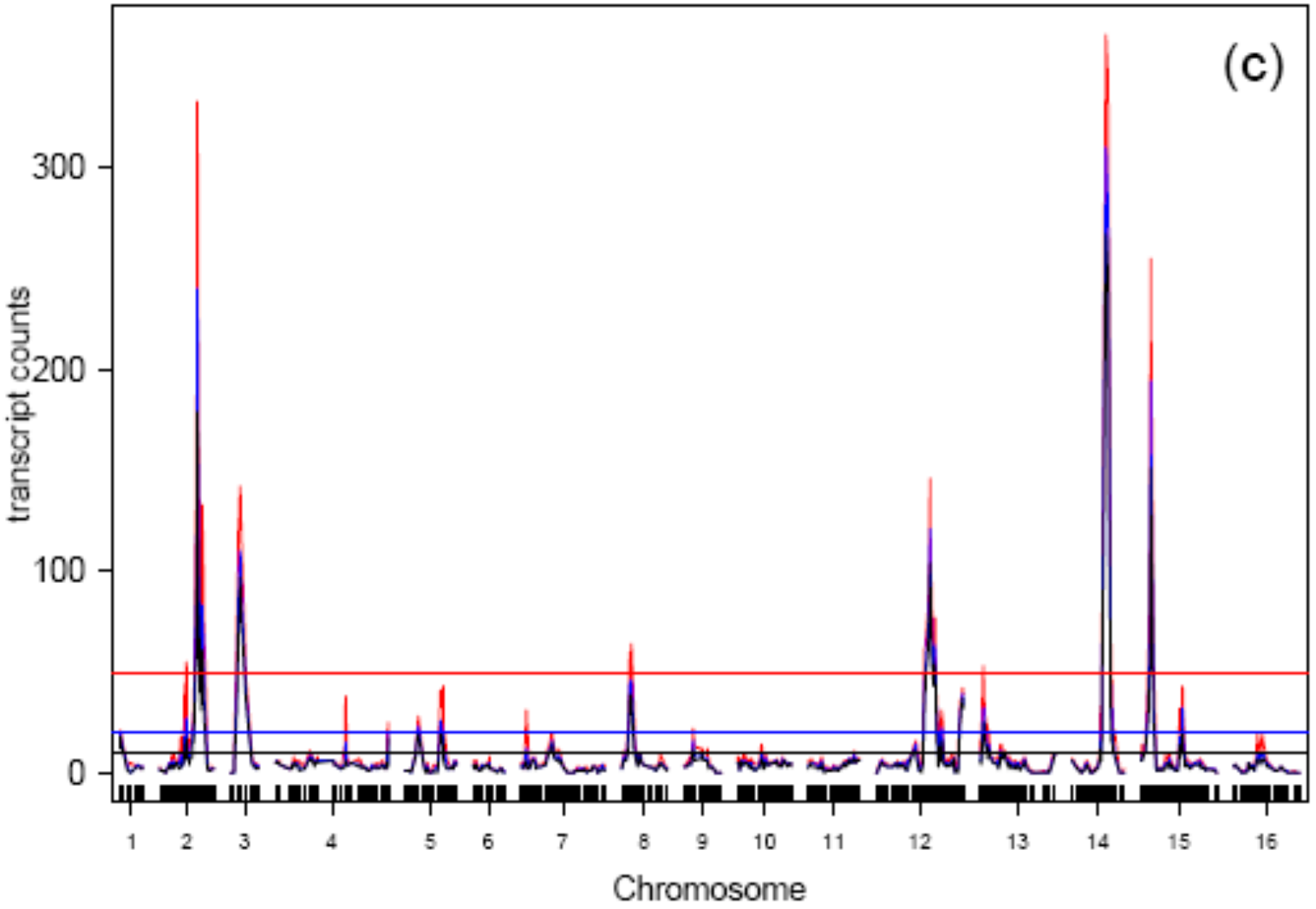
smaller spurious hotspots have higher LOD thresholds

larger spurious hotspots allow many traits with small LODS (below *T*=3.41)

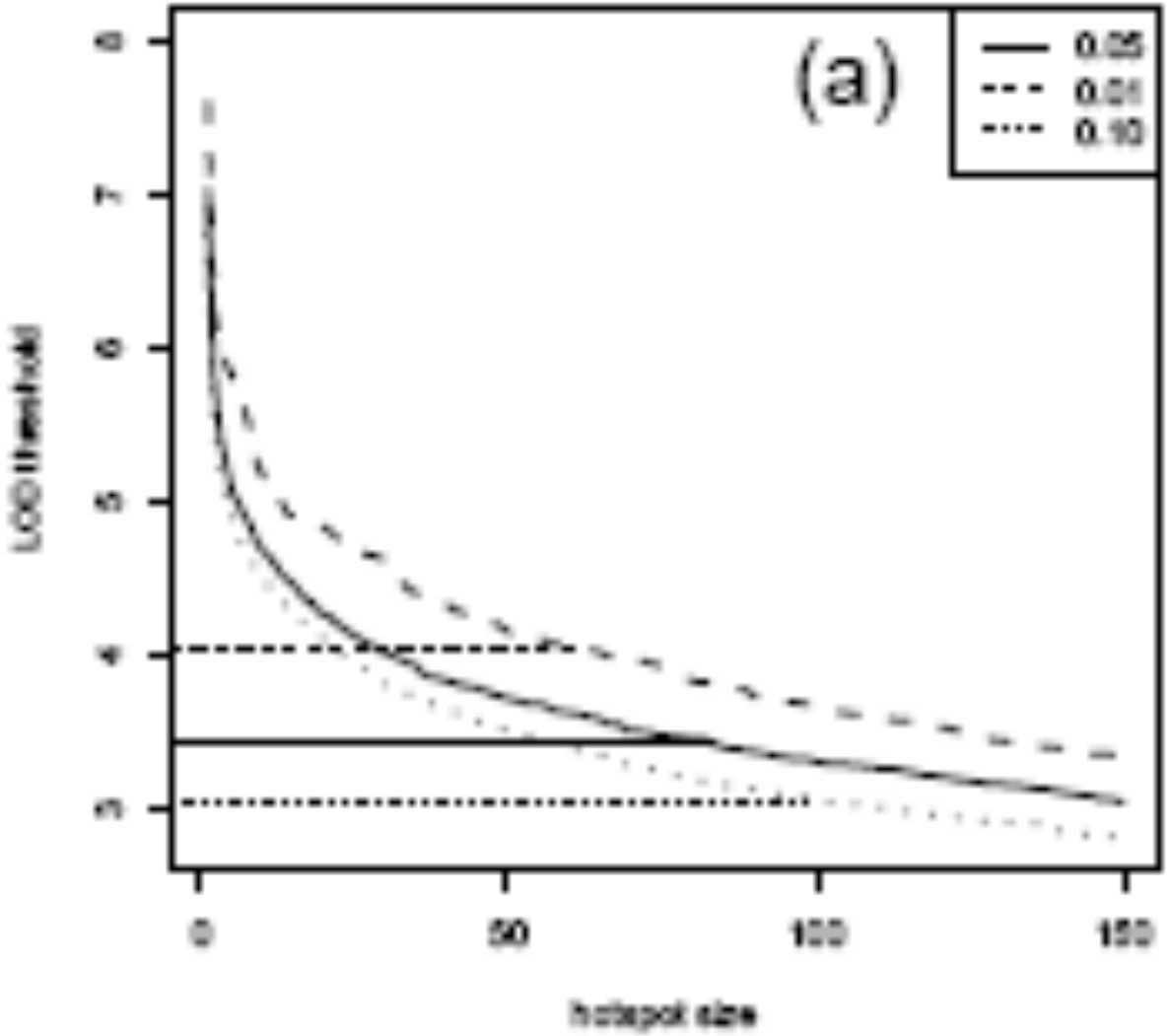
Hotspot architecture using multiple trait GWER threshold (T_1=7.12)



hotspot architectures using LOD thresholds for 10(black), 20(blue), 50(red) traits



Sliding threshold between multiple trait (T_1=7.12) and single trait (T_0=3.41) GWER



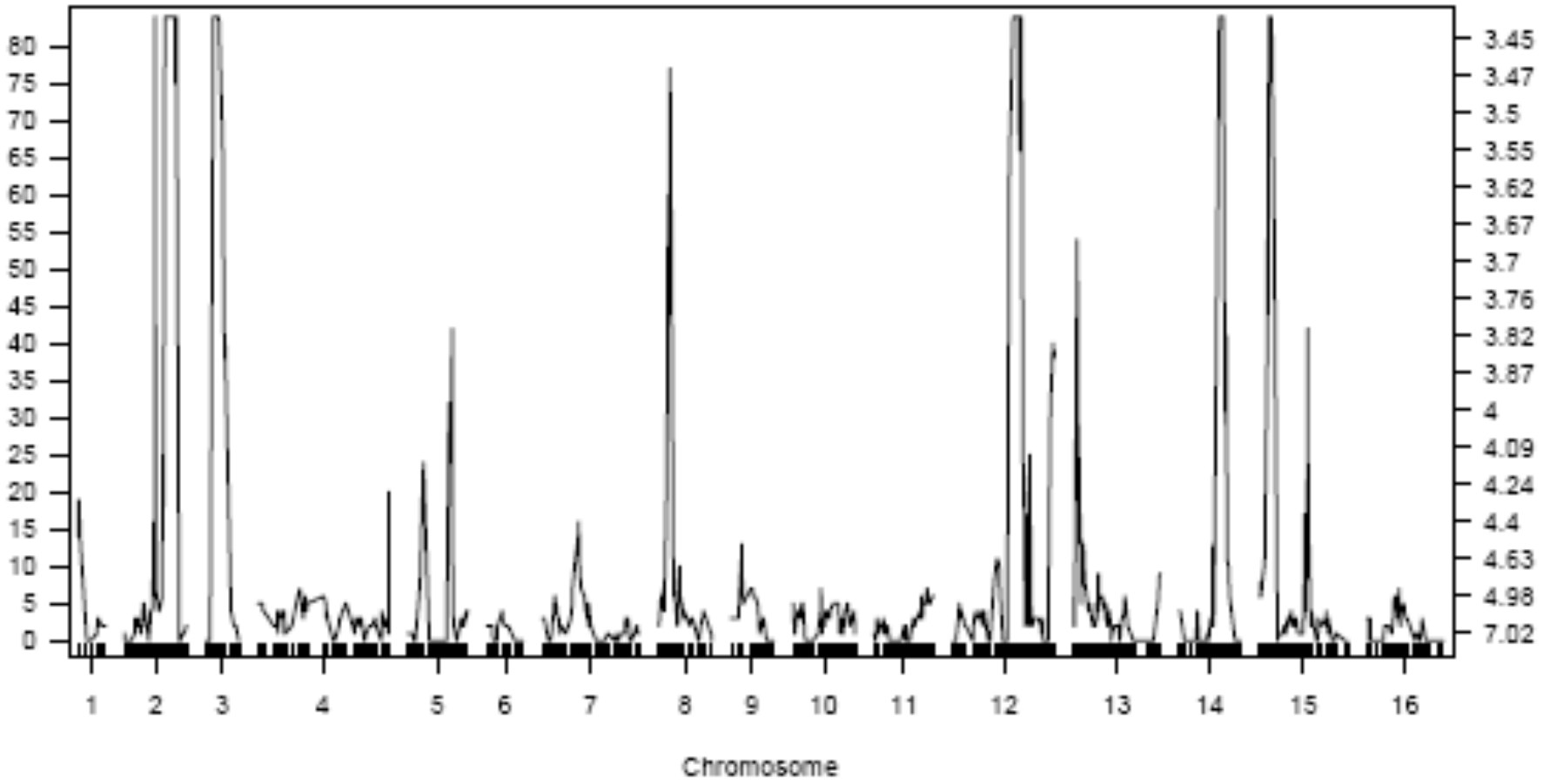
T_1=7.12 controls GWER across all traits

T_0=3.41 controls GWER for single trait

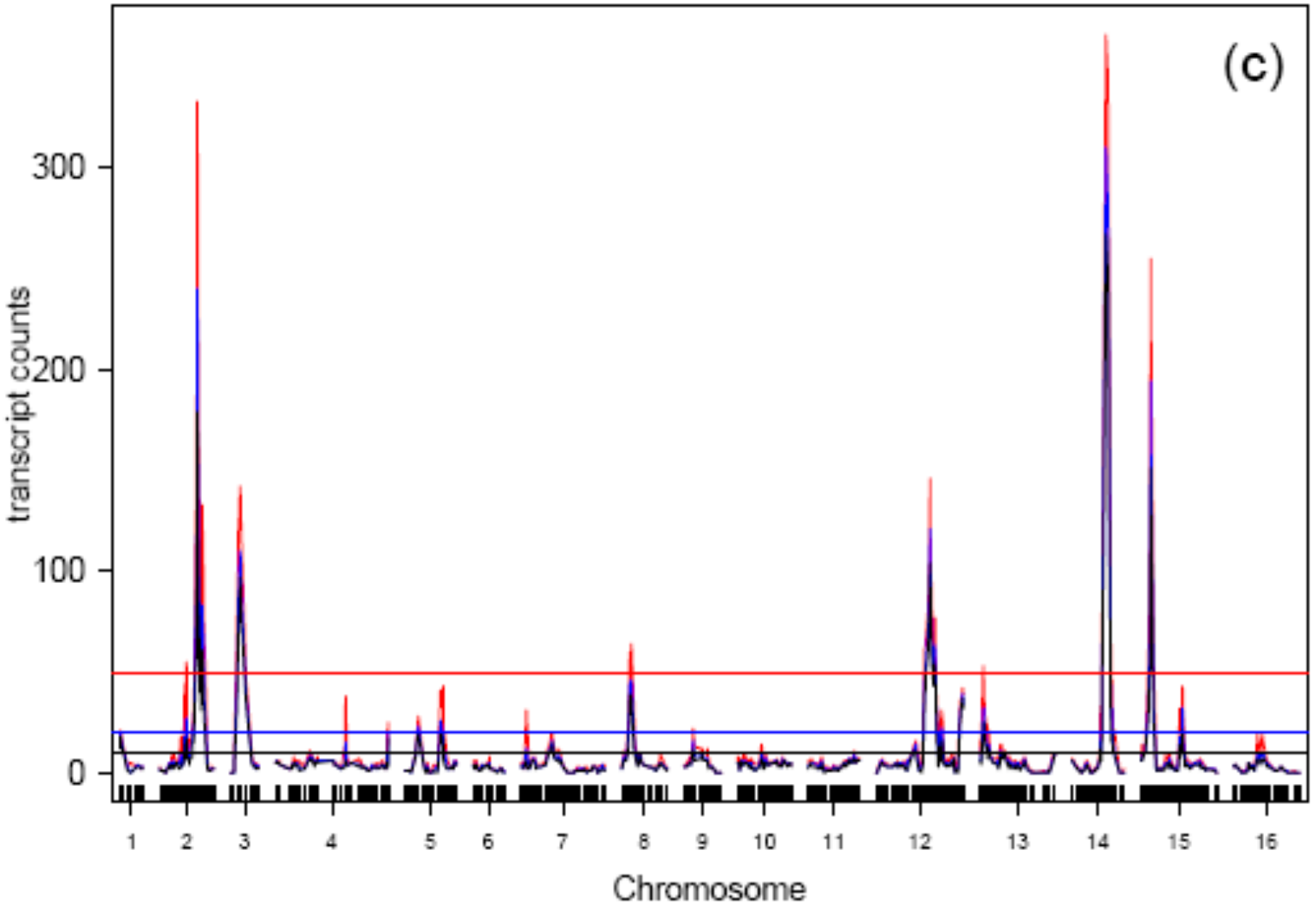
Hotspot size significance profile

- Construction
  - Fix significance level (say 5%)
  - At each locus, find largest hotspot that is significant using sliding threshold
  - Plot as profile across genome
- Interpretation
  - Large hotspots were already significant
  - Traits with LOD > 7.12 could be hubs
  - Smaller hotspots identified by fewer large LODs (chr 8)
  - Subjective choice on what to investigate (chr 13, 5?)

Hotspot size signifcance profile



hotspot architectures using LOD thresholds for 10(black), 20(blue), 50(red) traits



Yeast study

- 120 individuals
- 6000 traits
- 250 markers
- 1000 permutations
- 1.8 * 10^10 linear models

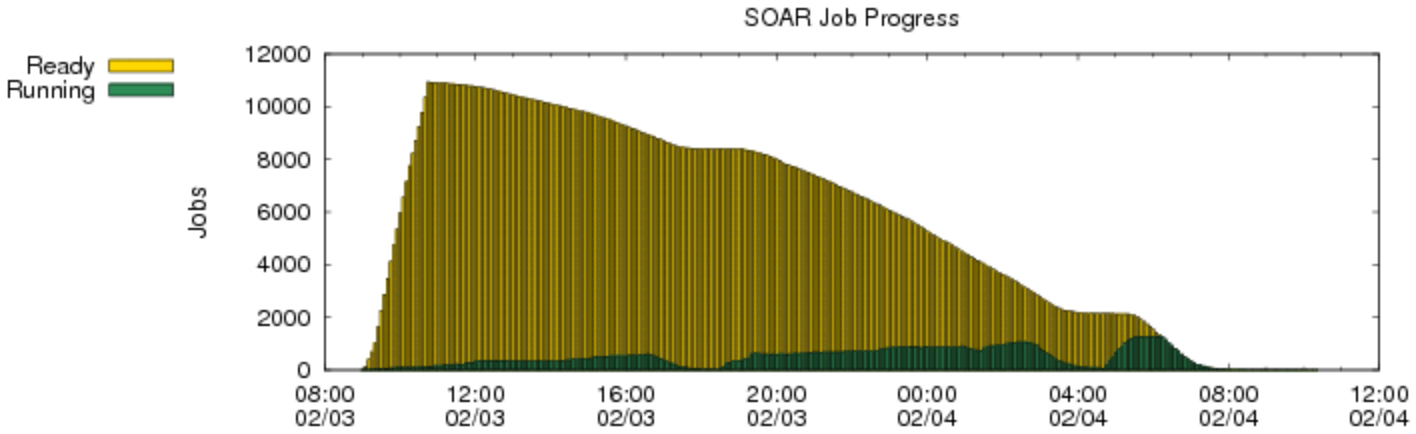
Mouse study

- 500 individuals
- 30,000 traits * 6 tissues
- 2000 markers
- 1000 permutations
- 1.8 * 10^13 linear models
- 1000 x more than yeast study

Scaling up permutations

- tremendous computing resource needs
  - Multiple analyses, periodically redone
    - Algorithms improve
    - Gene annotation and sequence data evolve
  - Verification of properties of methods
    - Theory gives easy cutoff values (LOD > 3) that may not be relevant
    - Need to carefully develop re-sampling methods (permutations, etc.)
  - Storage of raw, processed and summary data (and metadata)
    - Terabyte(s) of backed-up storage (soon petabytes and more)
    - Web access tools
- high throughput computing platforms (Condor)
  - Reduce months or years to hours or days
  - Free up your mind to think about science rather than mechanics
  - Free up your desktop/laptop for more immediate tasks
  - Need local (regional) infrastructure
    - Who maintains the machines, algorithms?
    - Who can talk to you in plain language?

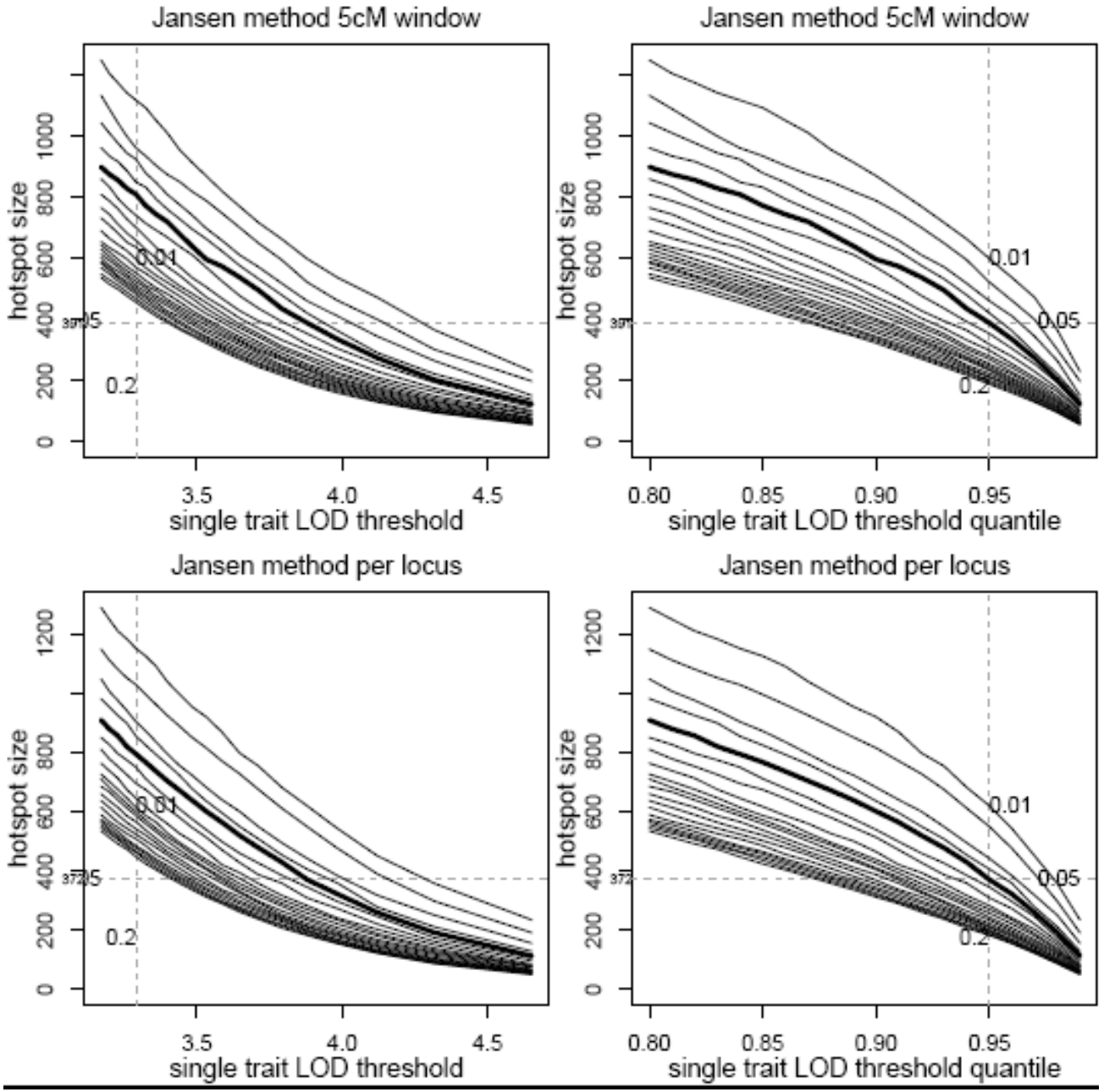
CHTC use: one “small” project

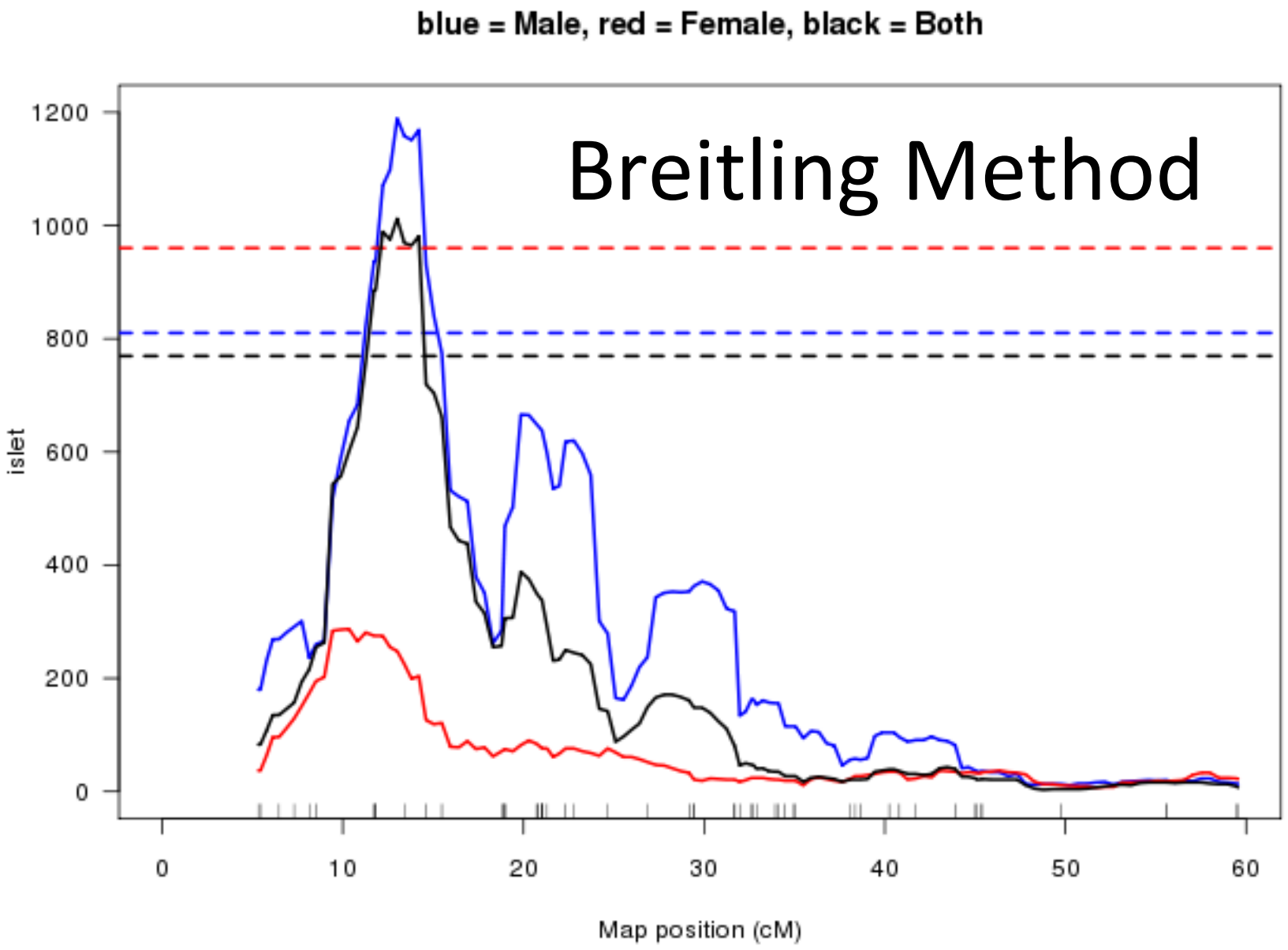


Open Science Grid Glidein Usage (4 feb 2012)

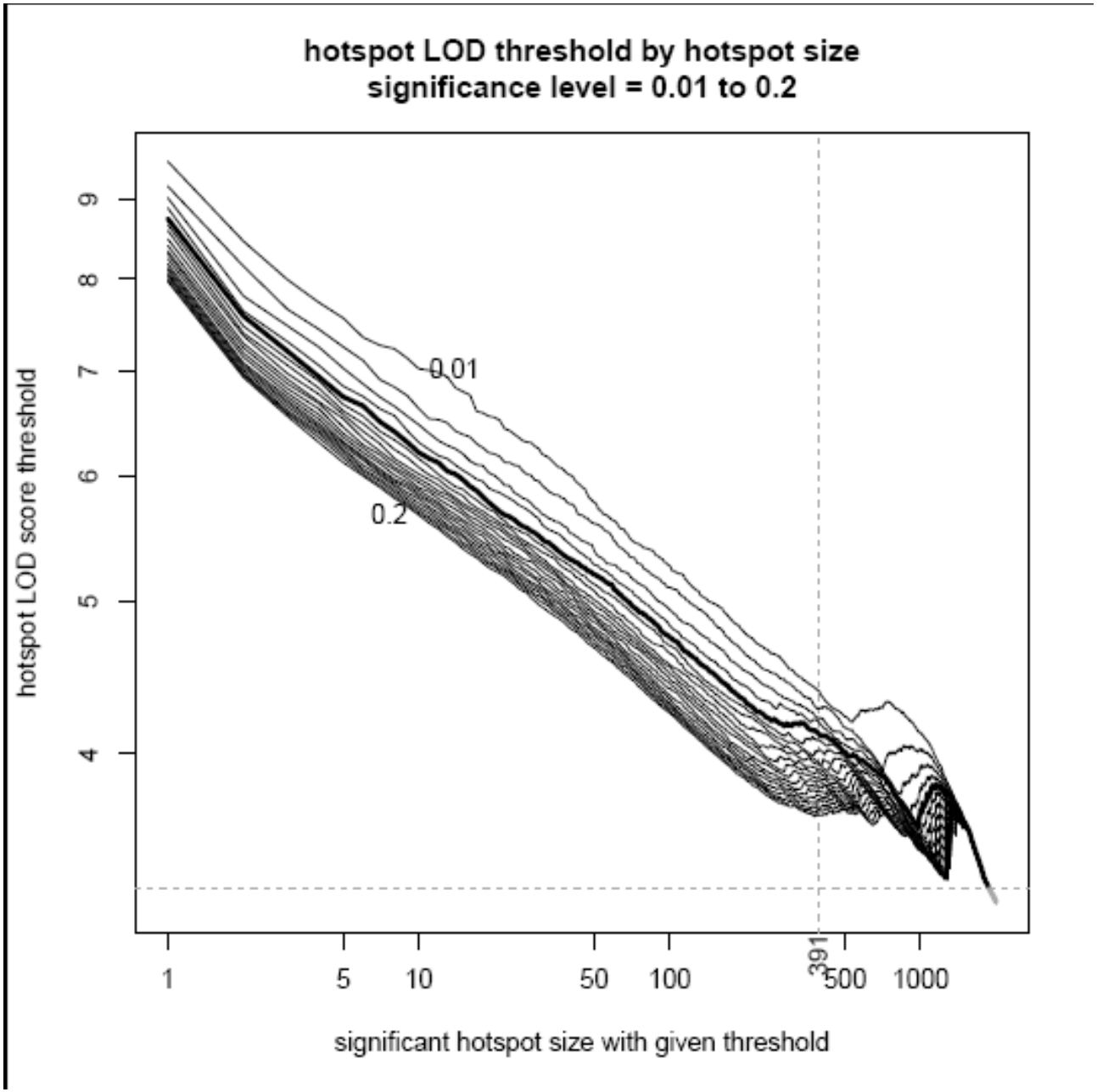
	group	hours	percent
1	BMRB	10710.37	3.49%
2	Biochem_Attie	3660.22	5.11%
3	Statistics_Wahba	178.5	1.22%

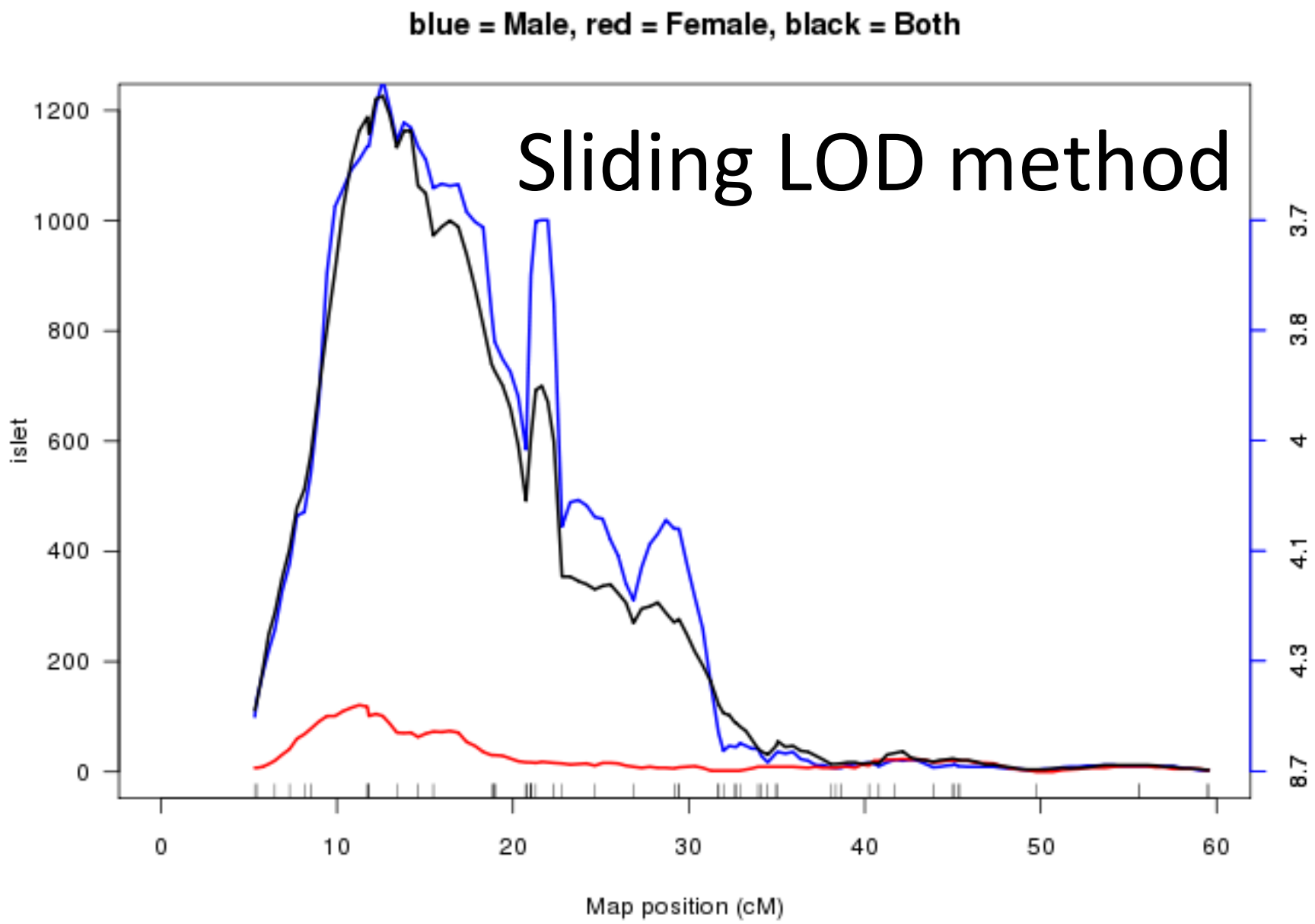
Brietling et al (2008) hotspot size thresholds from permutations




blue = Male, red = Female, black = Both
Breitling Method
1200
1000
800
600
400
200
0
islet
0
10
20
30
40
50
60
Map position (cM)

Chaibub Neto sliding LOD thresholds




blue = Male, red = Female, black = Both
Sliding LOD method
islet
1200
1000
800
600
400
200
0
3.7
3.8
4
4.1
4.3
8.7
0
10
20
30
40
50
60
Map position (cM)

What's next?

- Further assess properties (power of test)
- Drill into identified hotspots
  - Find correlated subsets of traits
  - Look for local causal agents (*cis* traits)
  - Build causal networks (another talk …)
- Validate findings for narrow hotspot
- Incorporate as tool in pipeline
  - Increase access for discipline researchers
  - Increase visibility of method

References

- Chaibub Neto E, Keller MP, Broman AF, Attie AD, Jansen RC, Broman KW, Yandell BS, Quantile-based permutation thresholds for QTL hotspots. *Genetics* (in review).
- Breitling R, Li Y, Tesson BM, Fu J, Wu C, Wiltshire T, Gerrits A, Bystrykh LV, de Haan G, Su AI, Jansen RC (2008) Genetical Genomics: Spotlight on QTL Hotspots. *PLoS Genetics 4*: e1000232.
- Churchill GA, Doerge RW (1994) Empirical threshold values for quantitative trait mapping. *Genetics 138*: 963-971.