Building Bridges from Breeding to Biometry and Biostatistics

Brian S. Yandell
Professor of Horticulture & Statistics
Chair of Statistics

April 2012

www.stat.wisc.edu/~yandell

Real knowledge is to know the extent of one's ignorance.

Confucius (on a bench in Seattle)

how did I get here?

- Biostatistics, School of Public Health, UC-Berkeley 1981
  - RA/TA with EL Scott, J Neyman, CL Chiang, S Selvin
  - PhD 1981
    - non-parametric inference for hazard rates (Kjell A Doksum)
  - Annals of Statistics (1983) 50 citations to date
- research evolution
  - early career focus on survival analysis
  - shift to non-parametric regression (1984-99)
  - shift to statistical genomics (1991--)
- joined Biometry Program at UW-Madison in 1982
  - attracted by chance to blend statistics, computing and biology
  - valued balance of mathematical theory against practice
  - enjoyed developing methodology driven by collaboration
- Chair of Statistics 2011---

outline

1. What are stat training options?
2. How to find that gene?
3. Are hotspots real?
4. Which came first? (causal models)

what are stat training options?

Undergraduate major in stat, bioinfo: hands on training

Minor in stat: set of courses

MS in biometry: research training in stat methods

- Companion to PhD in biosci fields

MS in stat/biostat: deeper methods training

- Skills in consulting across disciplines
- Realistic comprehensive exam (triage, write for researcher)

PhD in stat/biostat: develop new methods

- Develop methods from collaboration with biologist

Non-traditional training: shorter time frame

- Graduate certificate: set of course on methods
  - bioinformatics (now), big data analytics (coming)
- Prof MS in big data science under development

why train more statisticians?

- 200K new jobs in stat by 2018
- Big data explosion
  - Lagging analytics expertise in every field
  - Increasing demand for graduates…
- White House Big Data Initiative: $200M
  - Build capacity: algorithms, machines, people
- Madison Advanced Research Cyber Infrastructure
  - Campus-level coordination
  - Substantial $$/yr requested
  - Statistics will be major player

Statistical Genomics at UW-Madison

Cecile Ane, Statistics and Botany

Karl Broman, BMI and Genetics

Sunduz Keles, Statistics and BMI

Bret Larget, Statistics and Botany

Christina Kendziorski, BMI

Michael Newton, BMI and Statistics

Sebastien Roch, Math

Sushmita Roy, BMI and WID

Grace Wahba, Statistics

Sijian Wang, BMI and Statistics

Brian Yandell, Statistics and Horticulture, Chair of Statistics

Yingqi Zhao, BMI

Mark Craven, BMI and Computer Science, Director of CIBM

Colin Dewey, BMI and Computer Science

Michael Ferris, Computer Science and IsyE, Director of Optimization Theme of WID

Michael Gleischer, Computer Science (Human-Computer Interface)

Miron Livny, Computer Science, Director of CHTC

Julie Mitchell, Math, Biochem, Biophys, Dir BACTER Inst Comp Bio

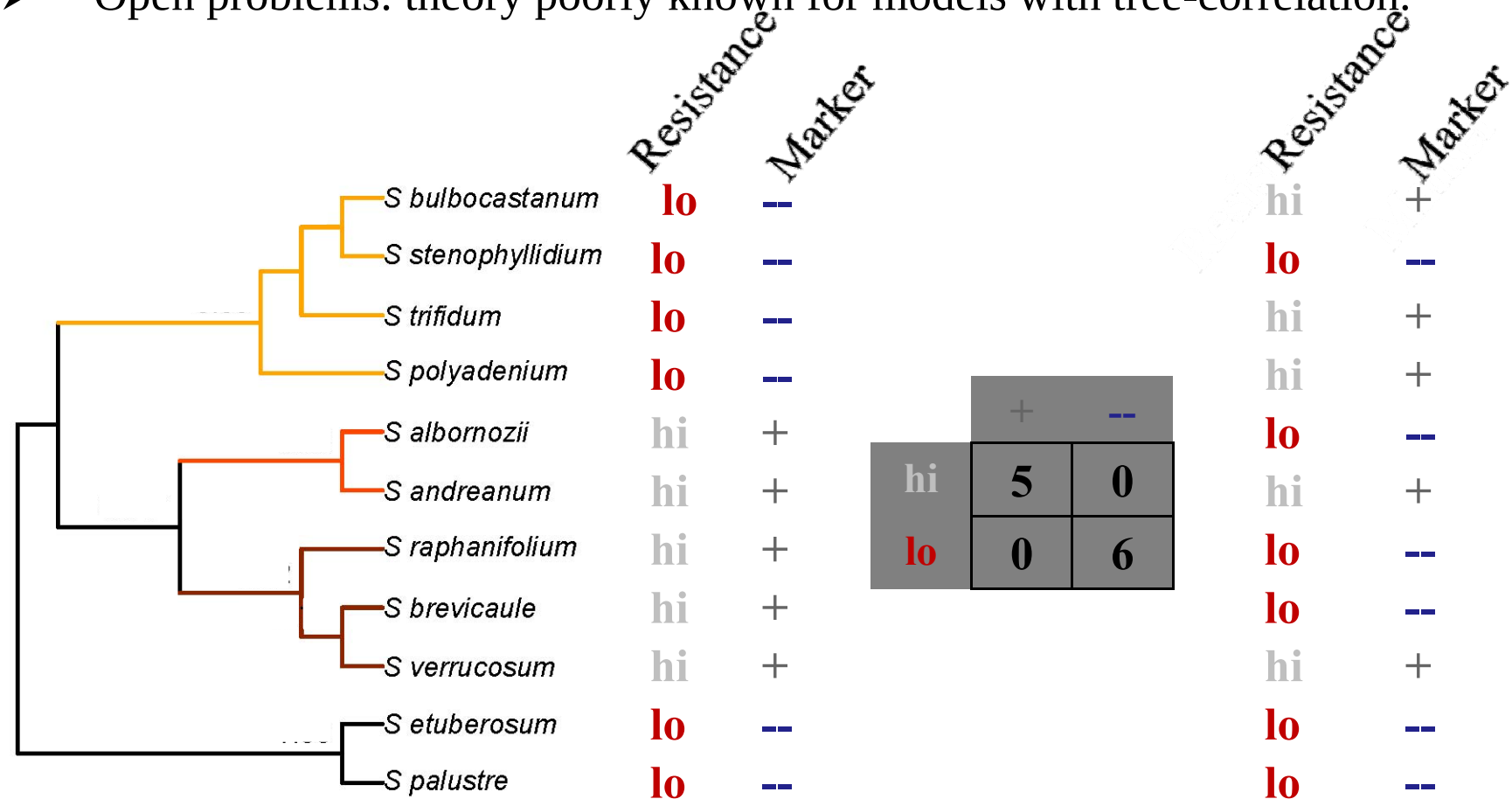
Dan Negrut, Computer Aided Engineering, Nvidia Fellow

Umberto Tachinardi, Assoc Dean and Chief Research Information Officer, SMPH

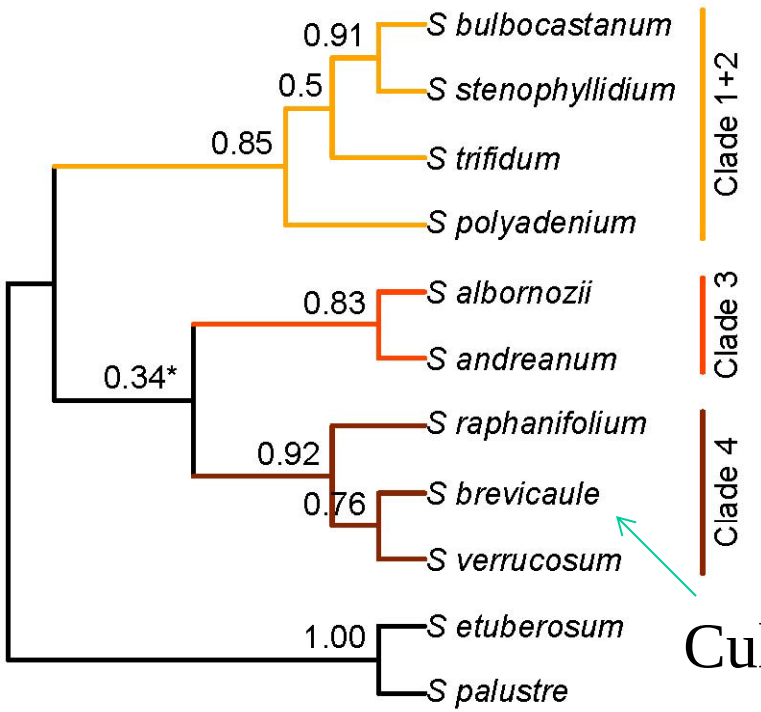
Statistical Phylogenetics

Bret Larget & Cecile Ane

- Phylogenetic trees used to model correlation due to shared ancestry
- Develop appropriate methods to detect correlation between traits & markers (or other covariates)
- Open problems: theory poorly known for models with tree-correlation.



Phylogenetic analysis of molecular sequences



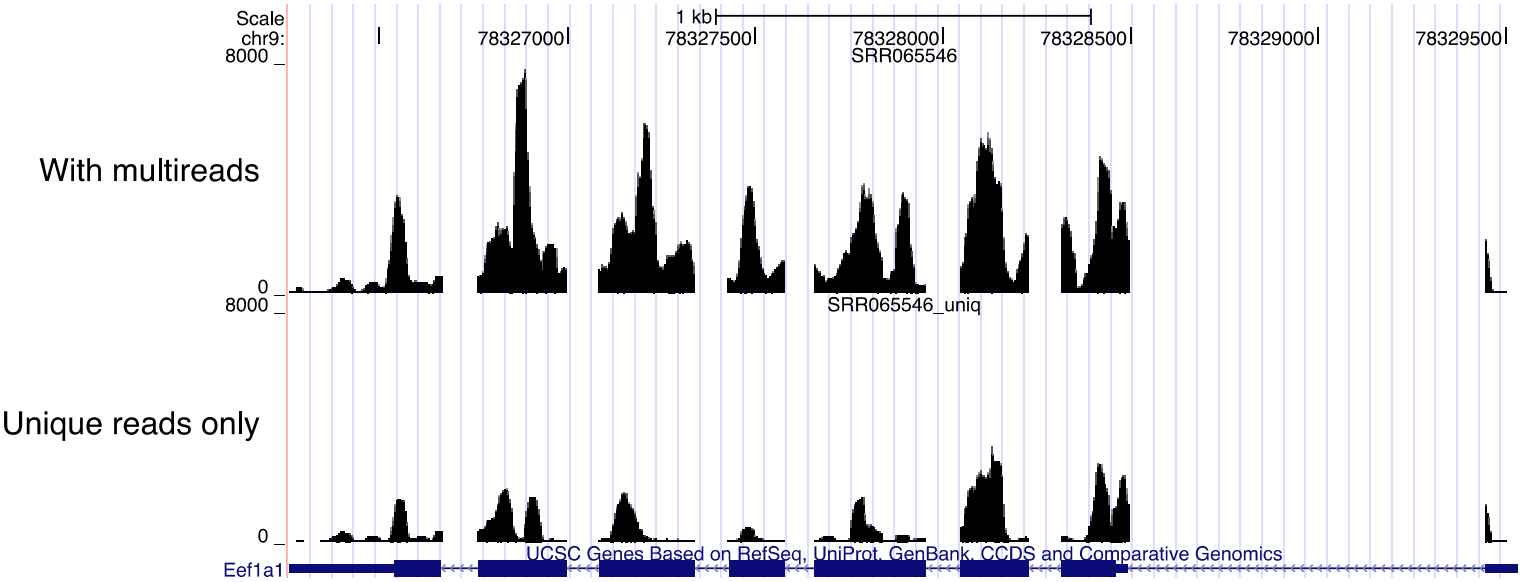
- Extremely large data sets: address computational challenges.
- Methods to deal with thousands of loci, e.g. from Next Gen. sequencing
- Resolve conflict between multiple loci
- Detect hybridization or horizontal gene transfers

Cultivated potato

Phylogeny of wild potatoes : extensive discordance among gene trees

joint work with David Spooner (Horticulture)

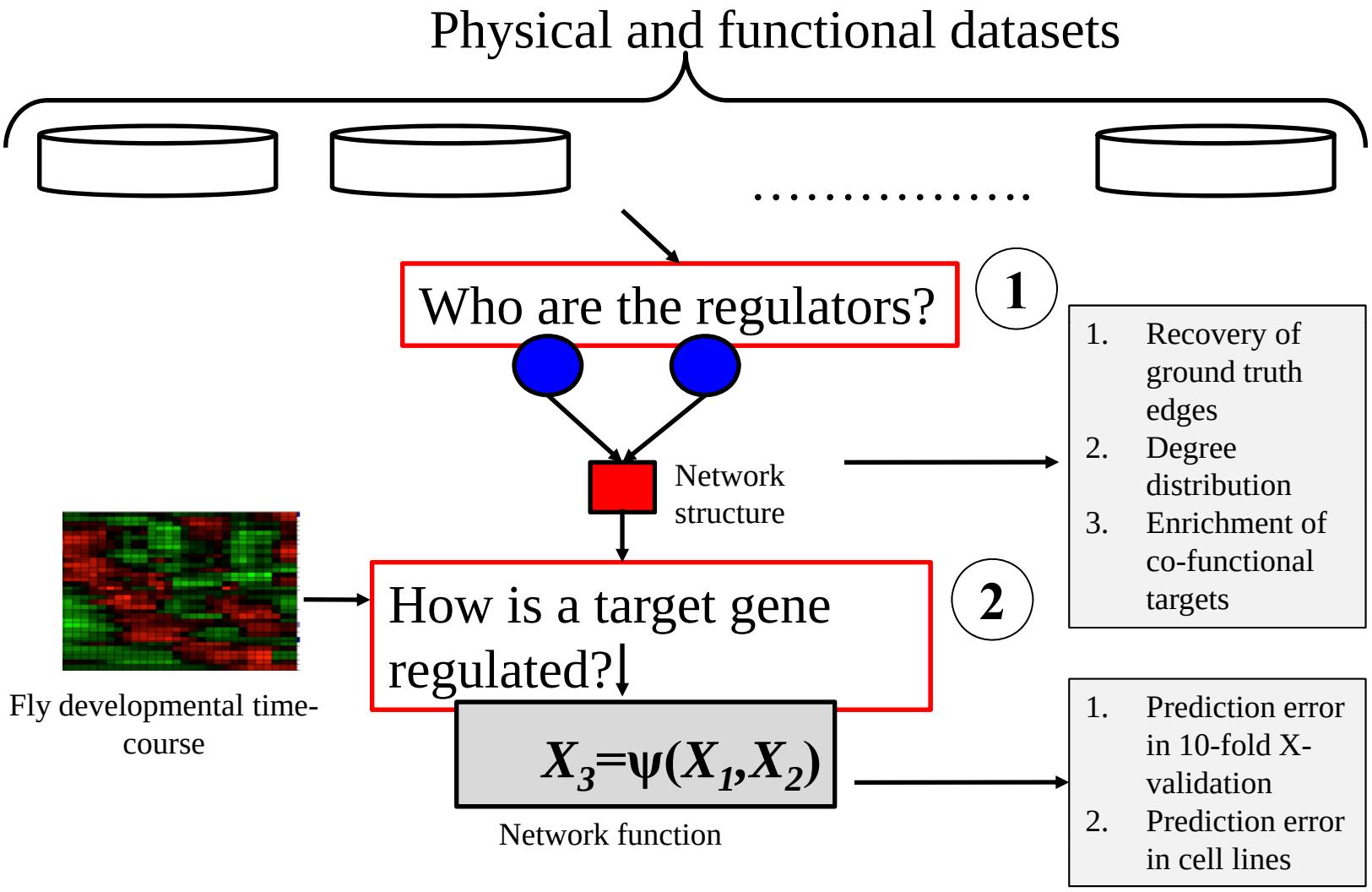
Estimating gene expression levels from RNA-Seq: handling ambiguous reads

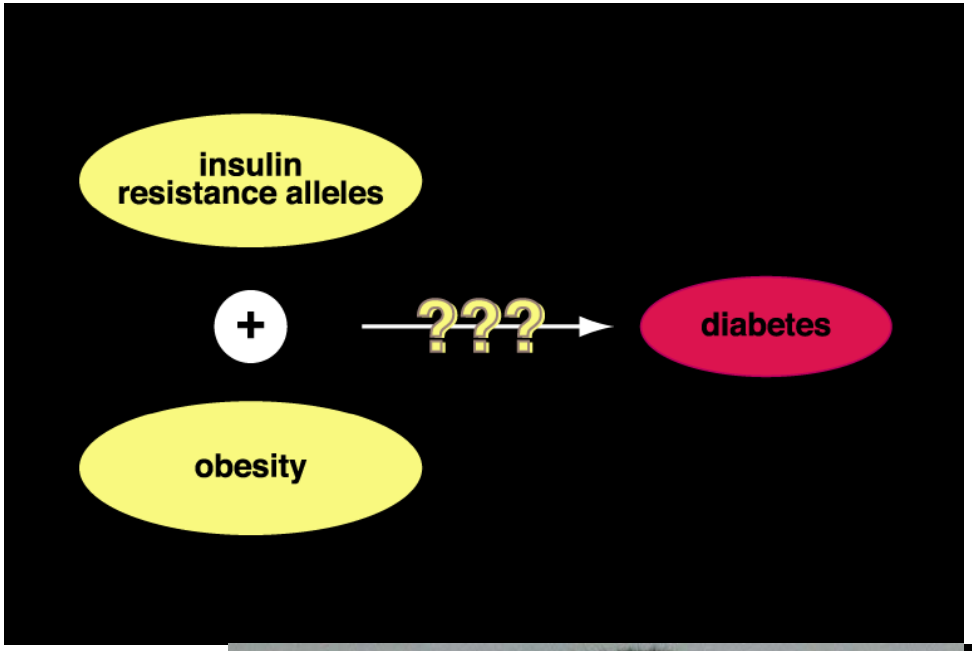


RSEM extracts more signal from the data through a statistical model of the RNA-Seq process

B. Li, V. Ruotti, R. Stewart, J. Thomson, and C. Dewey (2010) **RNA-Seq gene expression estimation with read mapping uncertainty**. *Bioinformatics* 26(4): 493-500.

Learning the regulatory network for fly



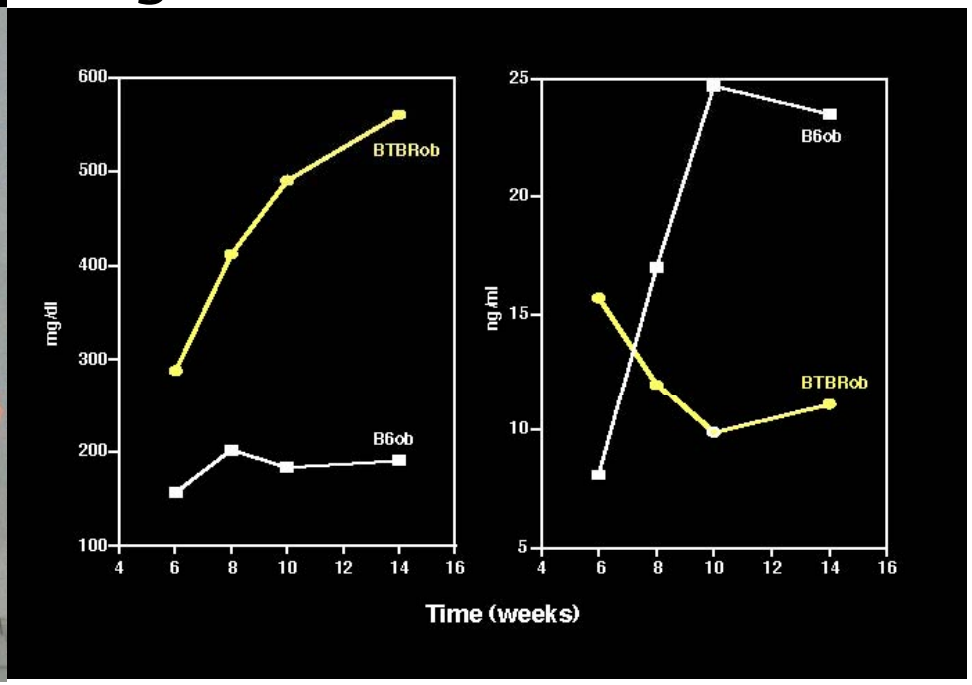


BTBR mouse is insulin resistant

B6 is not

make both obese…

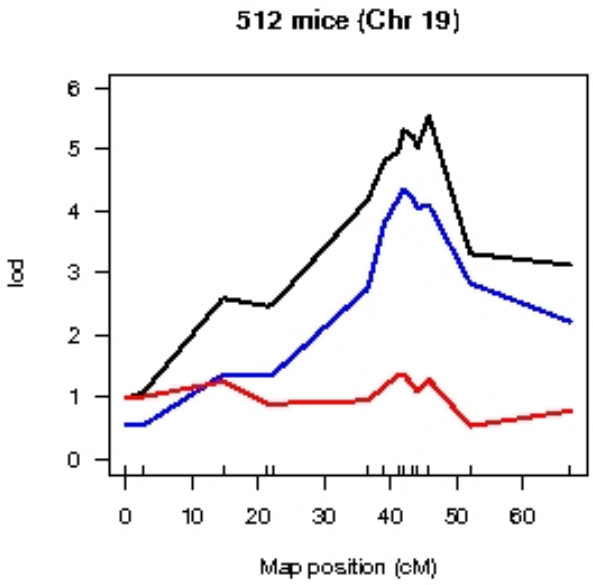
glucose insulin

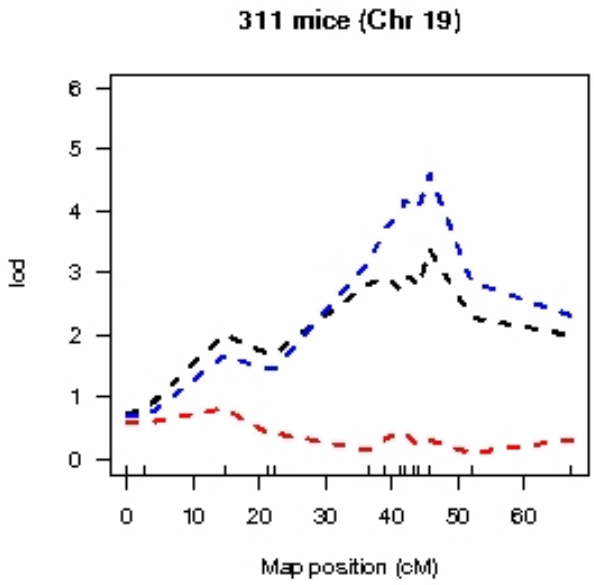


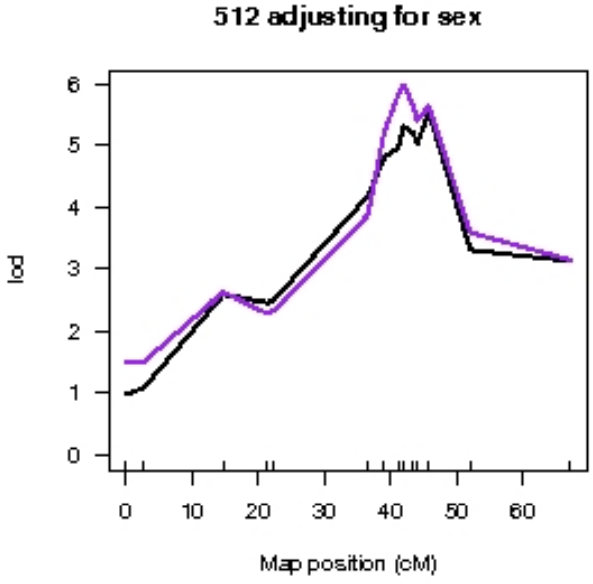
How to find that gene?

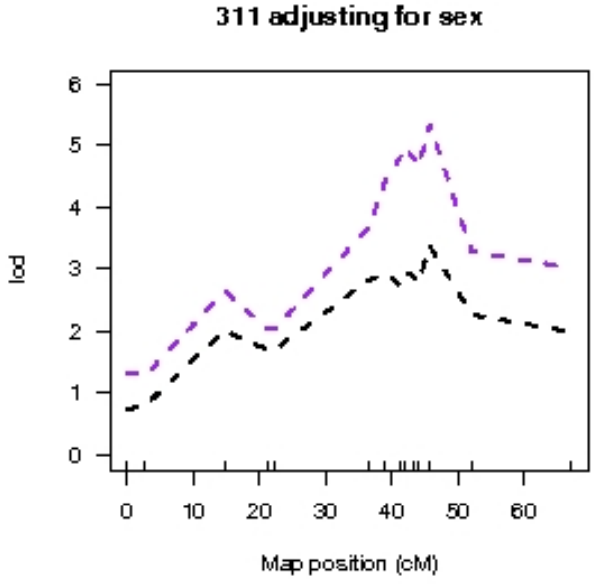
log10(ins10)
Chr 19
black=all
blue=male
red=female
purple=sex-adjusted

solid=512 mice
dashed=311 mice









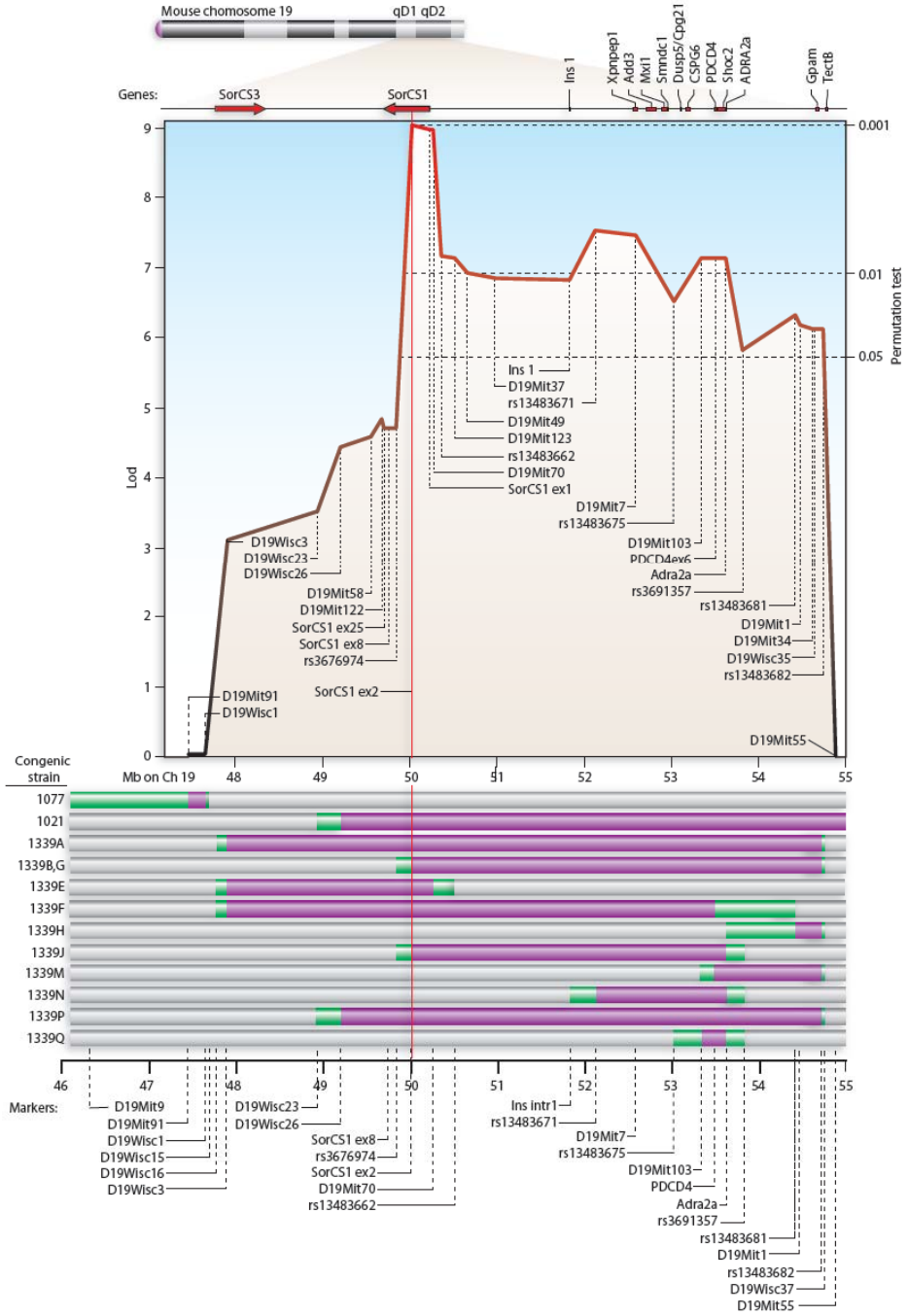
Sorcs1 study
in mice:

11 sub-congenic strains

marker regression
meta-analysis

within-strain
permutations

Nature Genetics 2006
Clee, Yandell *et al.*



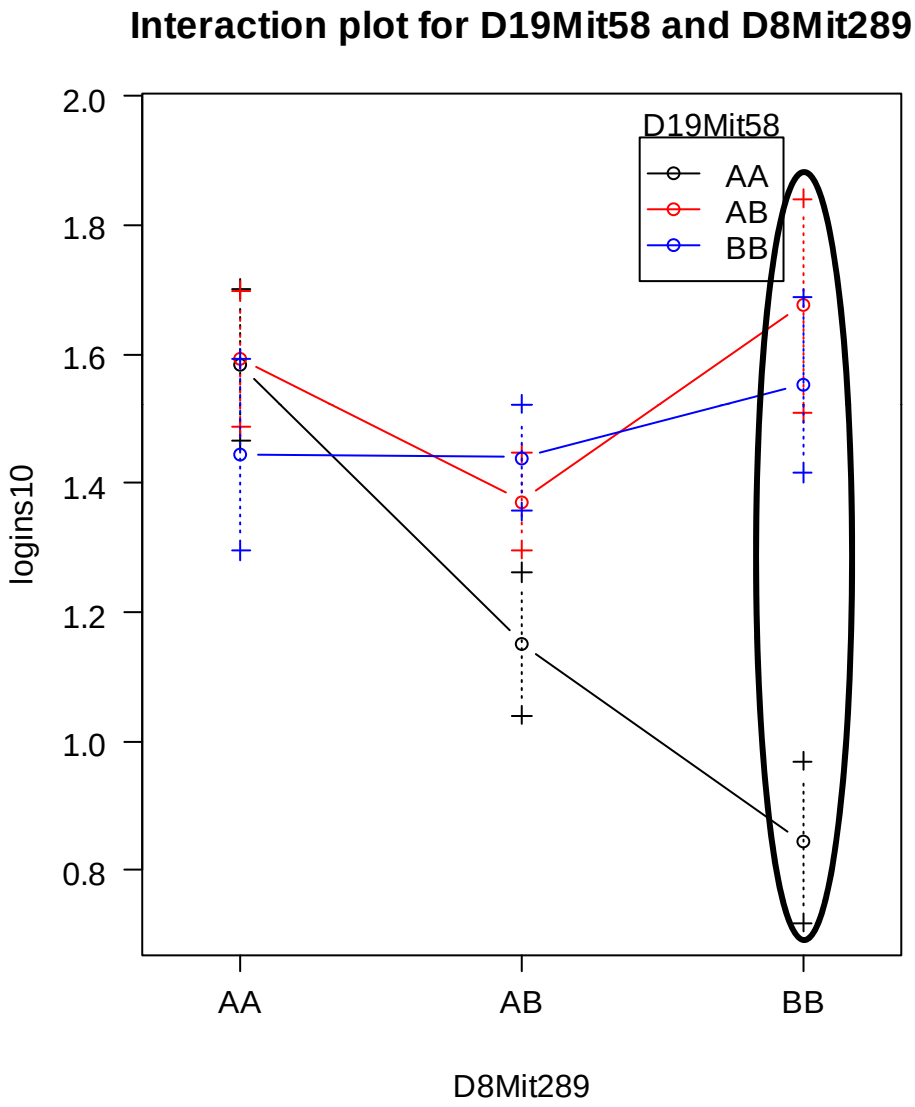
we were lucky!

BTBR background needed to see SORCS1

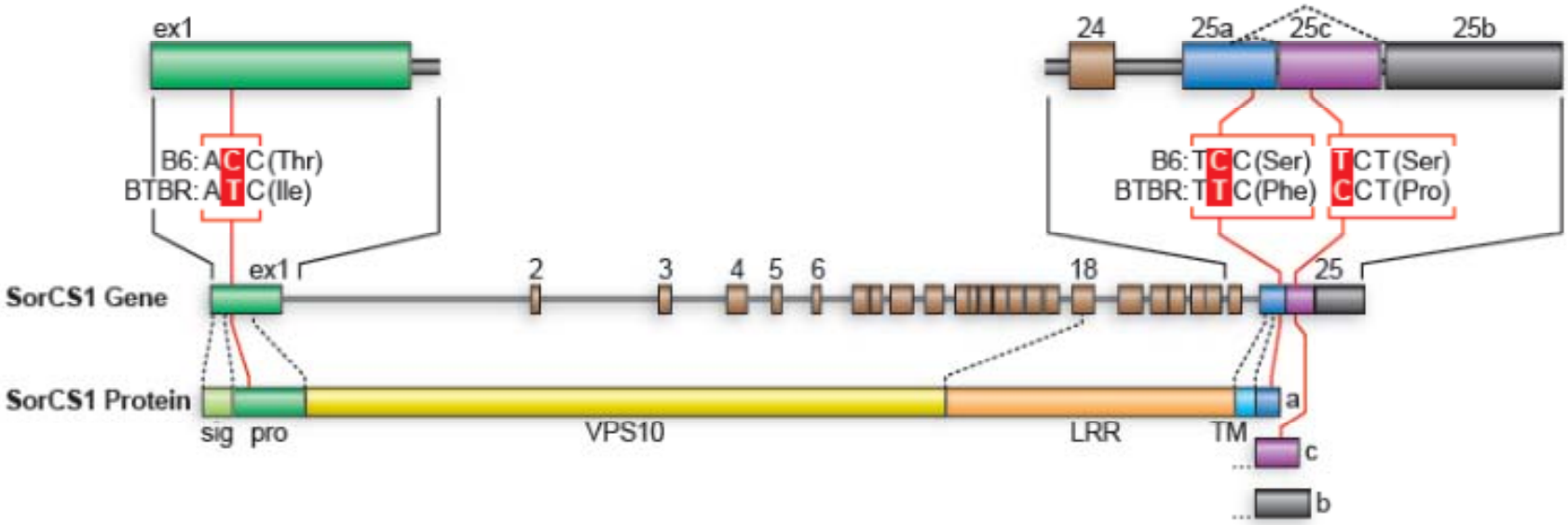
epistatic interaction of chr 19 and 8

…

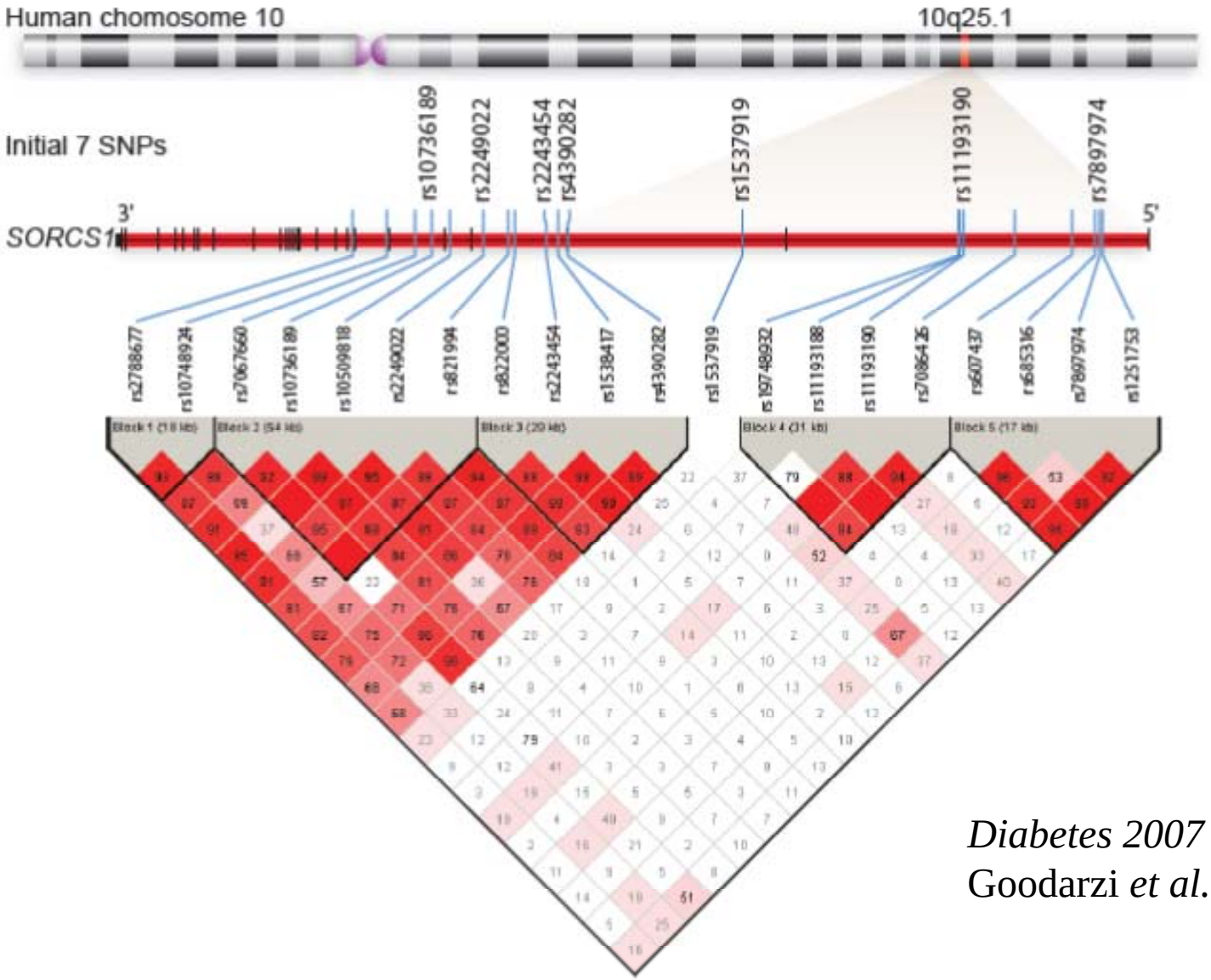
discovered much later



Sorcs1 gene & SNPs

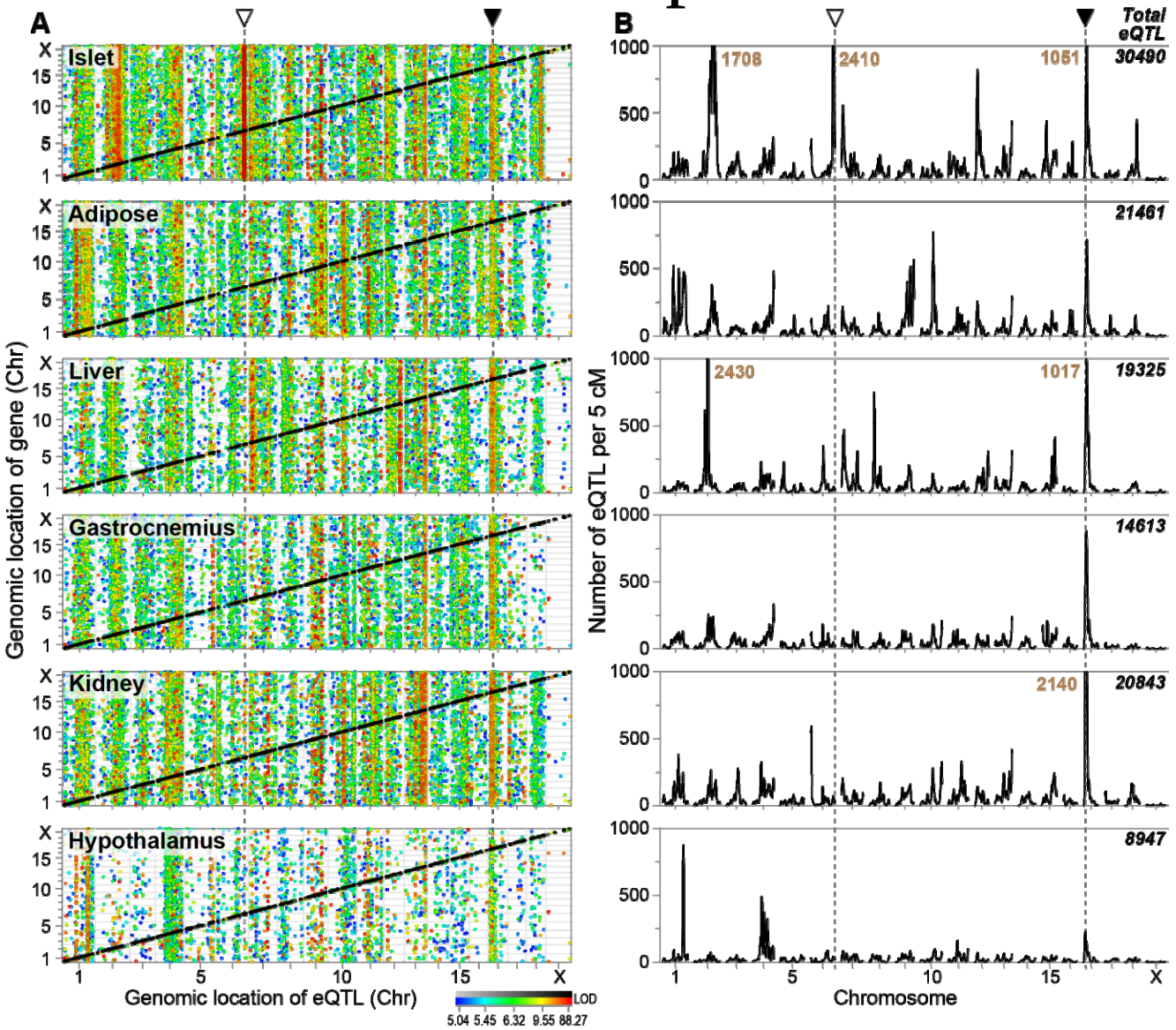


Sorcs1 study in humans



Diabetes 2007
Goodarzi *et al.*

Are these hotspots real?

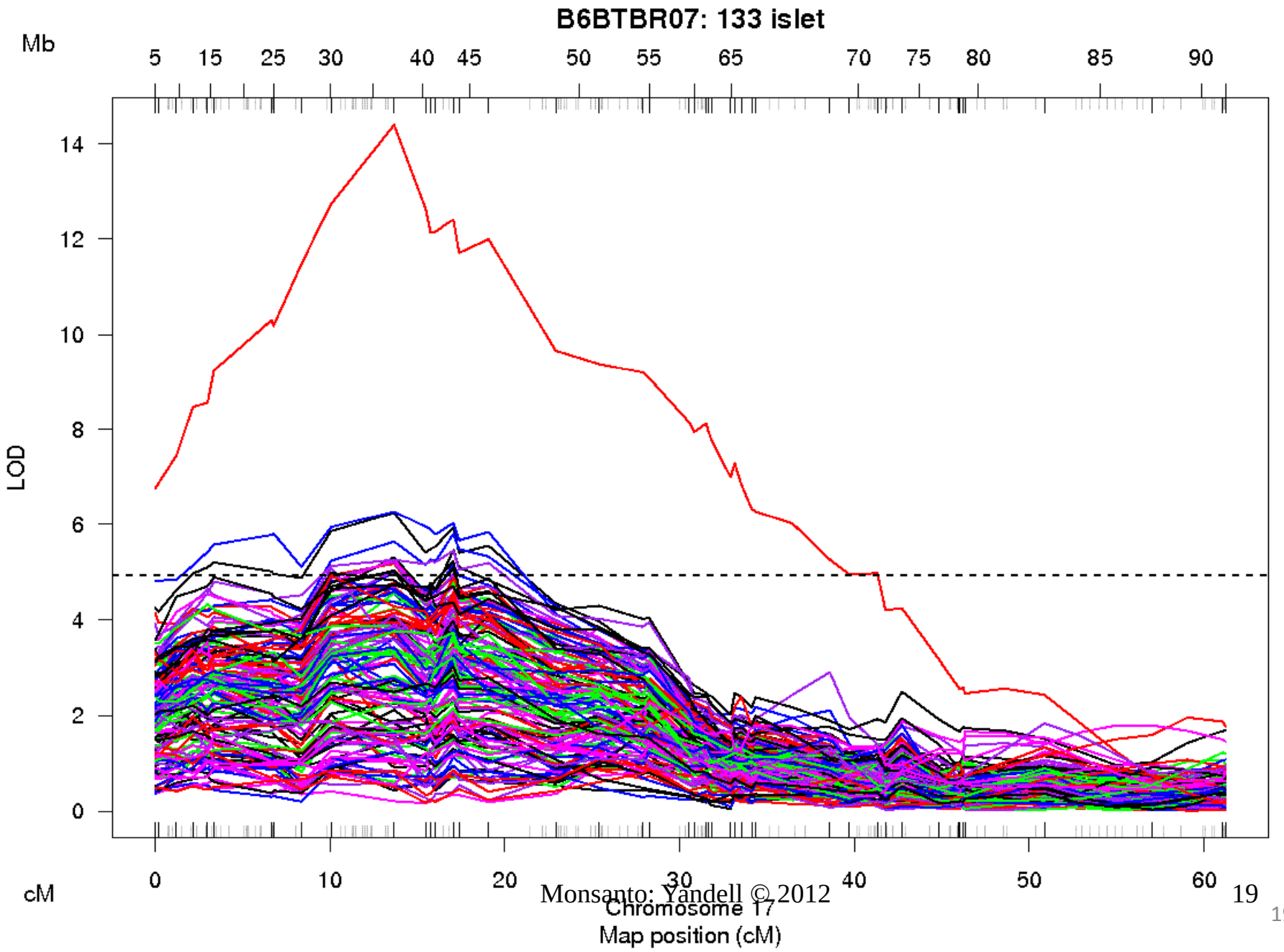


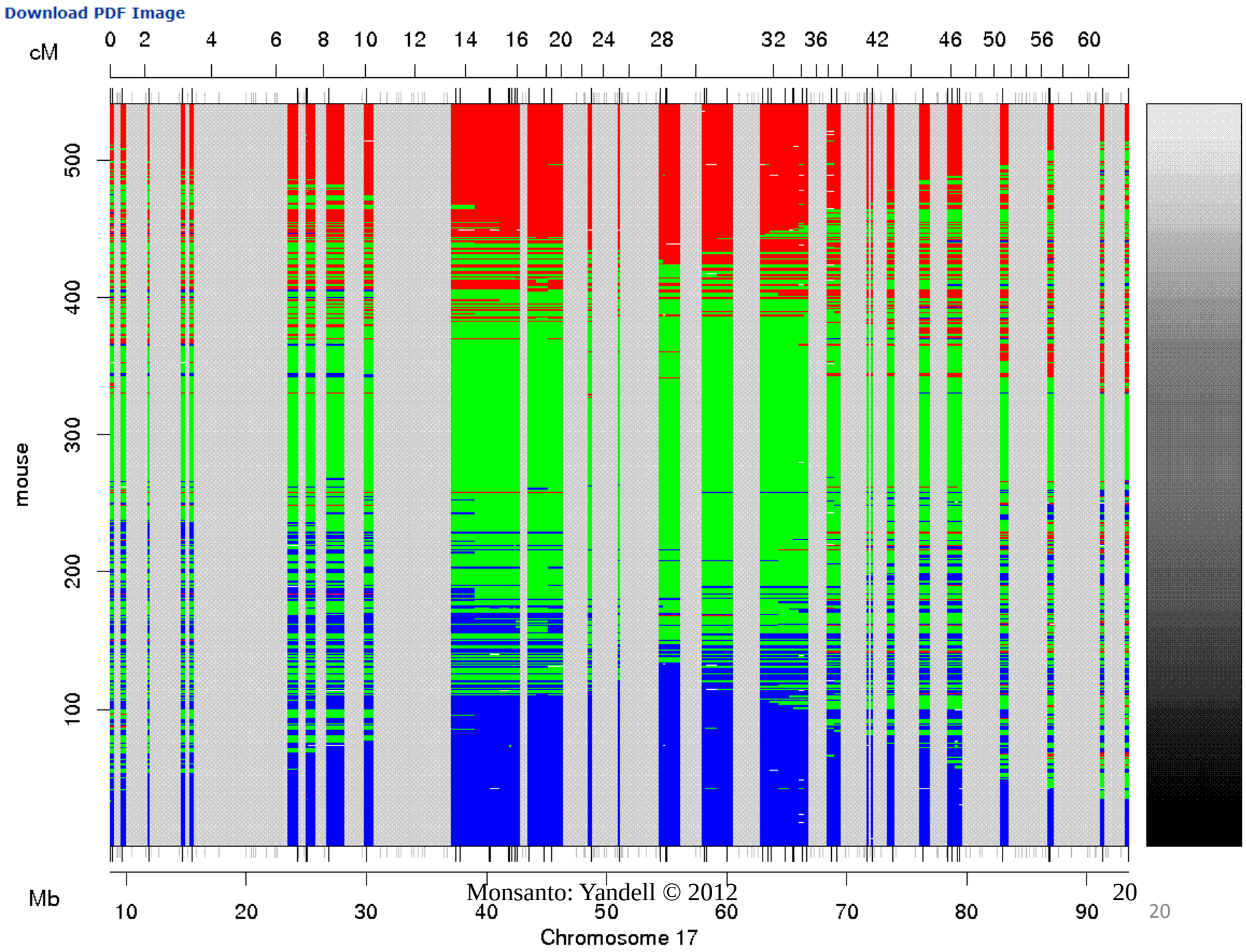
experimental context

- B6 x BTBR obese mouse cross
  - model for diabetes and obesity
  - 500+ mice from intercross (F2)
  - collaboration with Rosetta/Merck
- genotypes (1M values)
  - 5K SNP Affymetrix mouse chip (2K segregating SNPs)
  - care in curating genotypes! (map version, errors, …)
- phenotypes (120M values)
  - clinical phenotypes (200 / mouse)
  - gene expression traits (40K / mouse / 6 tissues)
  - other molecular traits (proteomic, miRNA, metabolomic)

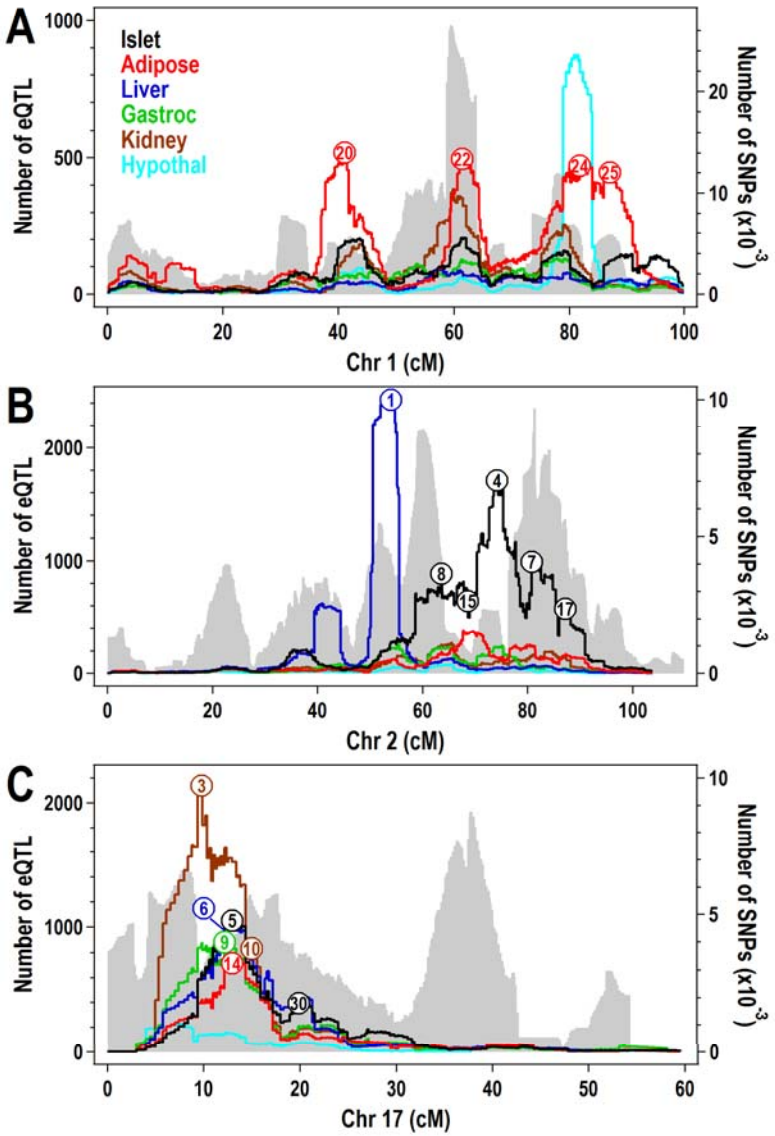
R script executed for 5 seconds.

Download PDF Image





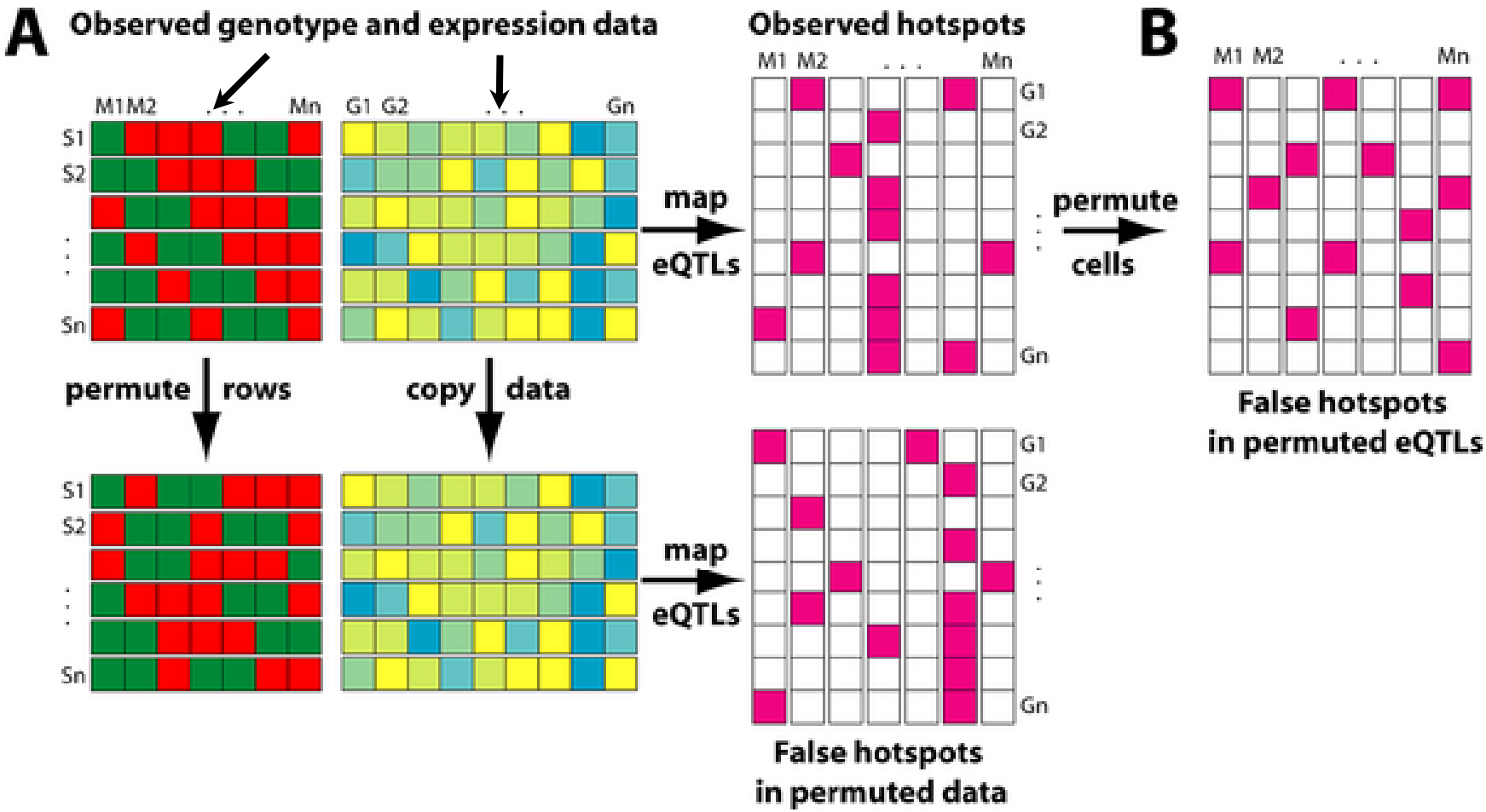
Tissue-specific hotspots with eQTL and SNP architecture



Are these hotspots real?

permutation across traits

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



hotspot permutation test

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

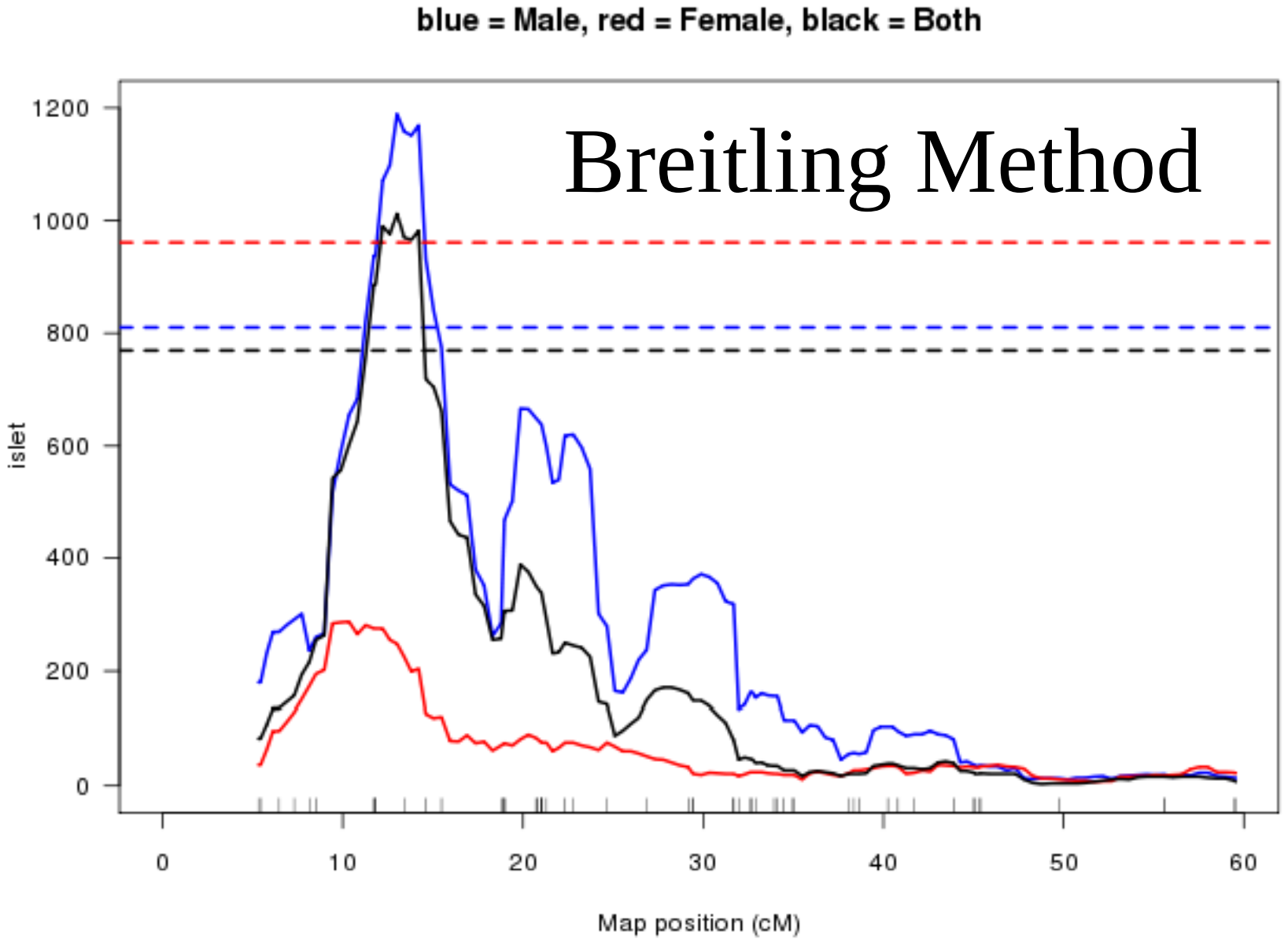
for original dataset and each permuted set:

- set single trait LOD threshold *T*
  - could use Churchill-Doerge (1994) permutations
- count number of traits with LOD above *T*
  - do this at every marker (or pseudomarker)
  - smooth counts with 5cM window

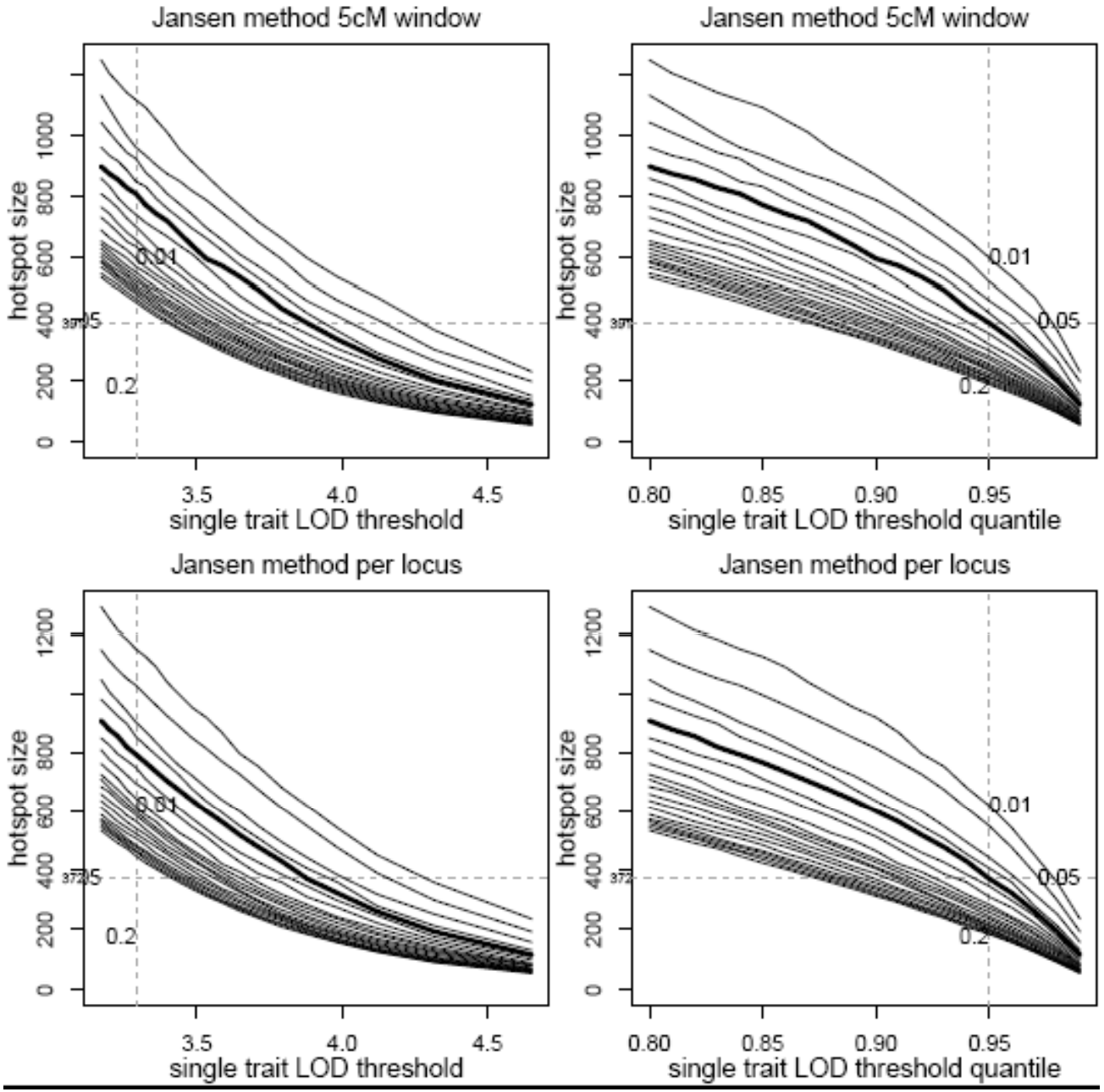
find 5% count threshold *N*(*T*)

- at most 5% of permuted sets above *N*(*T*)

conclude original counts above *N*(*T*) are real


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)

Brietling et al (2008) hotspot size thresholds from permutations



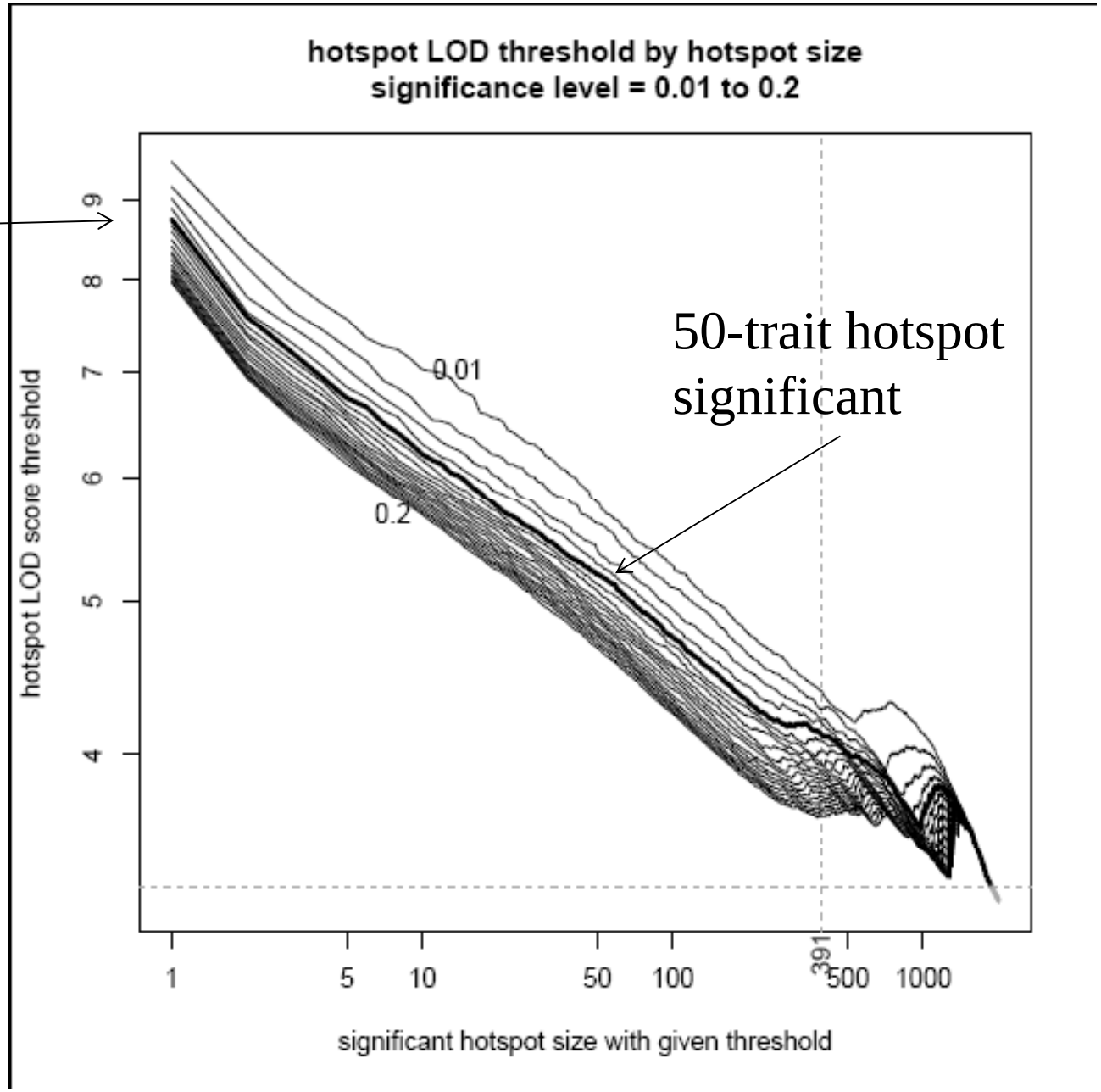
rethinking the approach

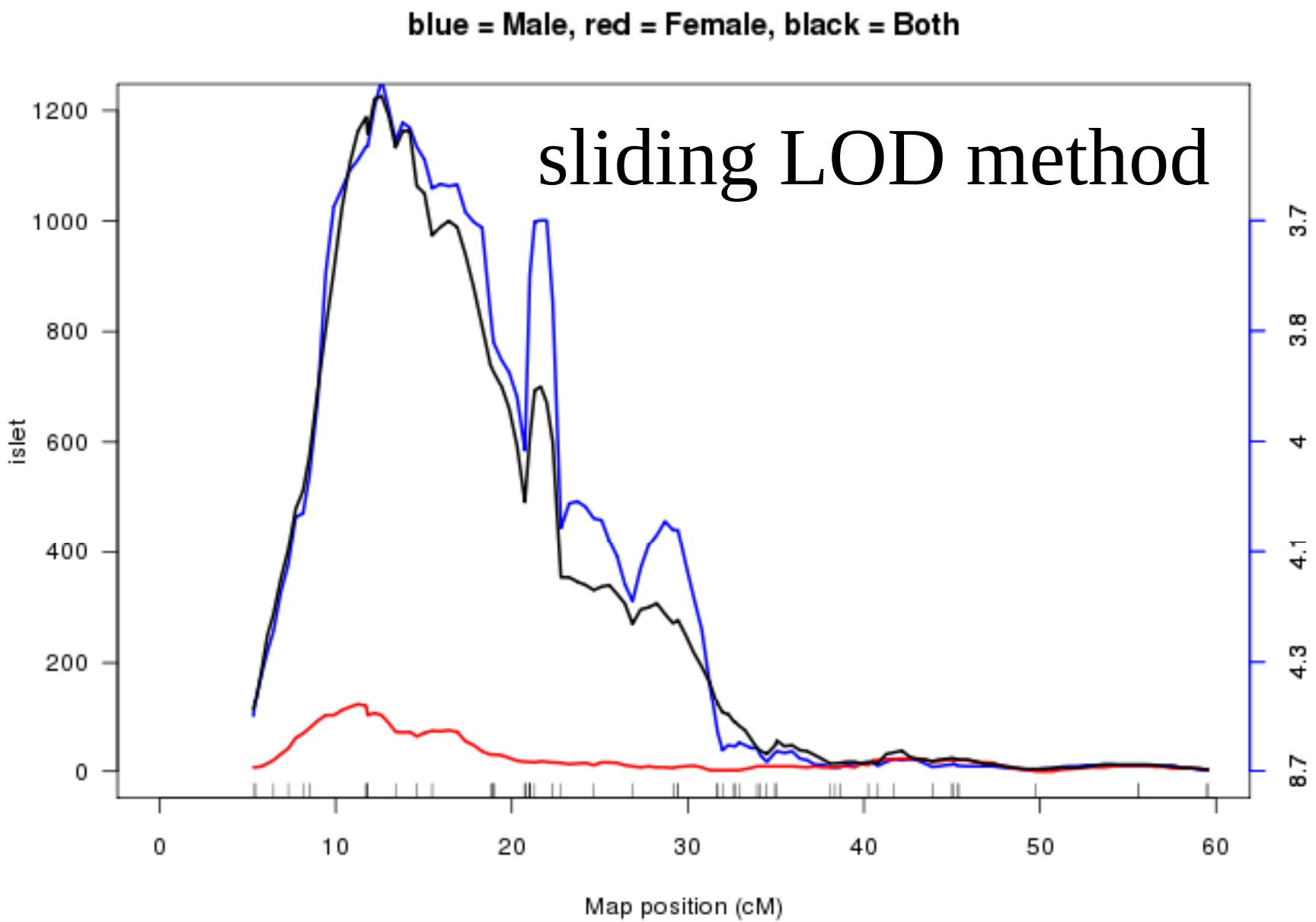
- For a hotspots of size *N*, what threshold *T*(*N*) is just large enough to declare 5% significance?
- *N = 1* (single trait)
  - What threshold *T*(1) is needed to declare any single peak significant?
  - valid across all traits and whole genome

Chaibub Neto E, Keller MP, Broman AF, Attie AD, Jansen RC, Broman KW, Yandell BS, Quantile-based permutation thresholds for QTL hotspots. *Genetics* (tent. accepted).

single trait significant

Chaibub Neto sliding LOD thresholds




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

Scaling up calculations

Genetics paper: 10B linear models to fit

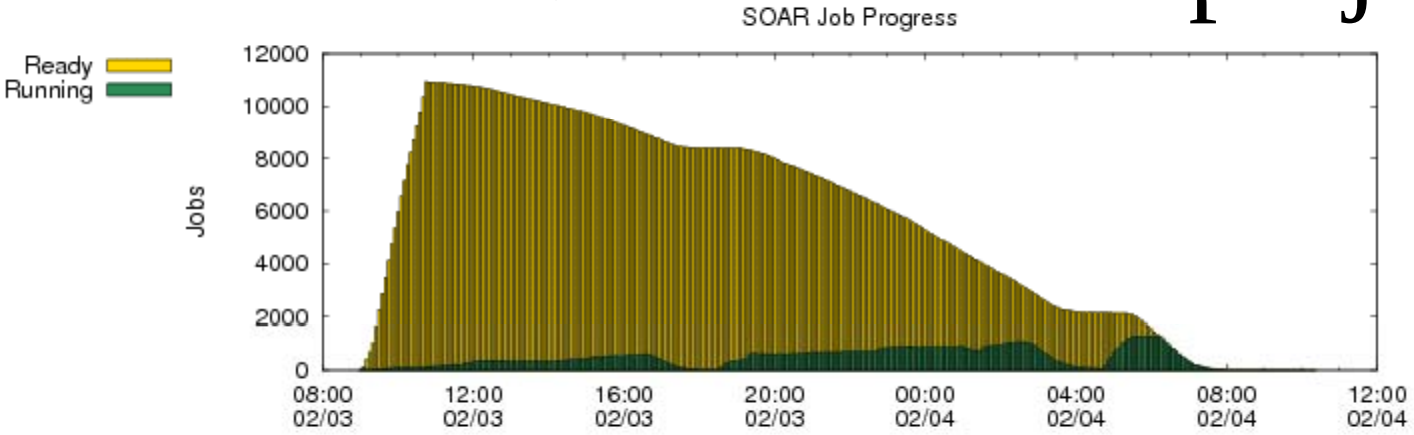
mouse study: 1000 x 10B linear models!

parallelize computations on OpenScienceGrid

www.chtc.wisc.edu

500 individuals
30,000 traits * 6 tissues
2000 markers
1000 permutations

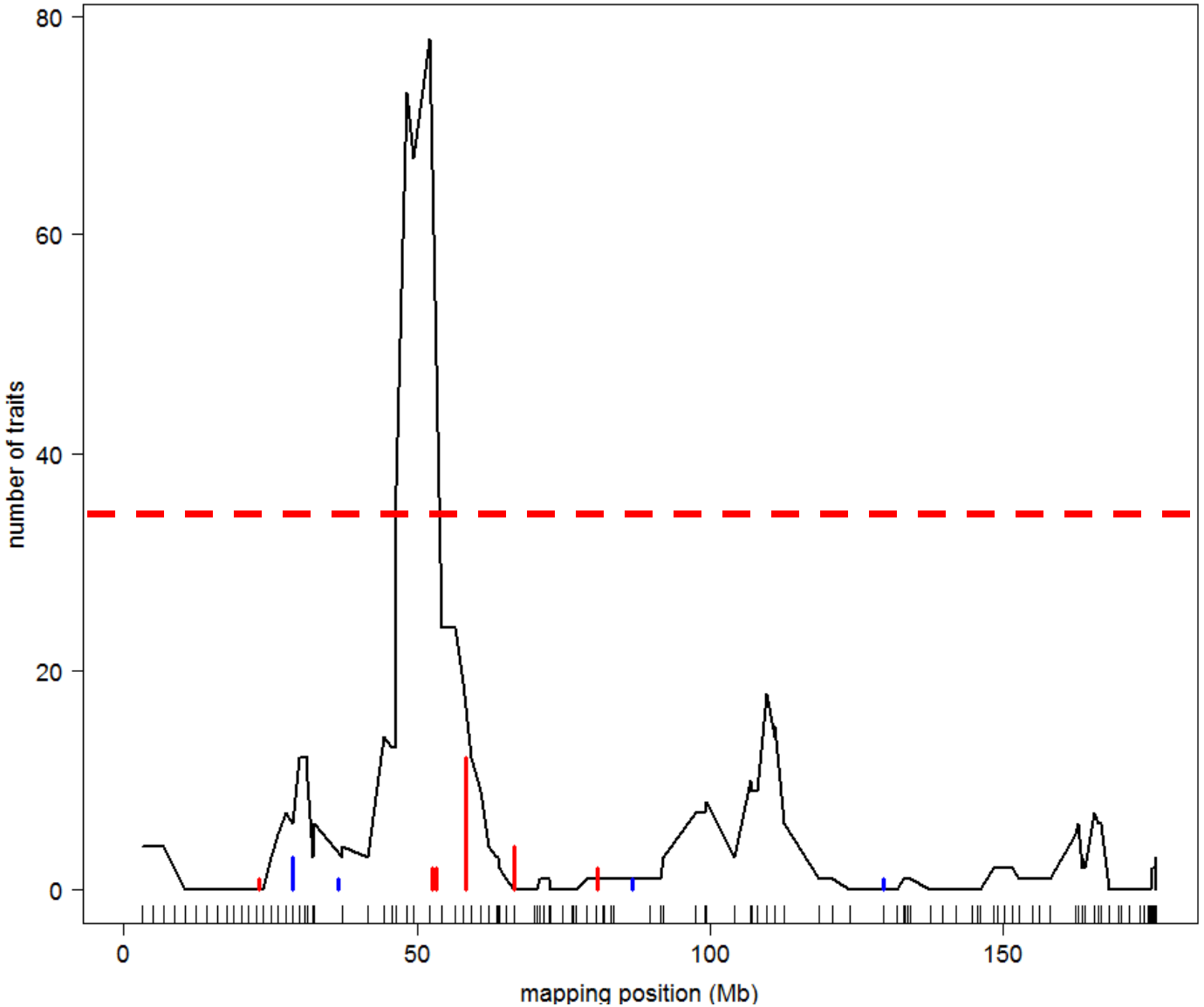
CHTC use: one “small” project



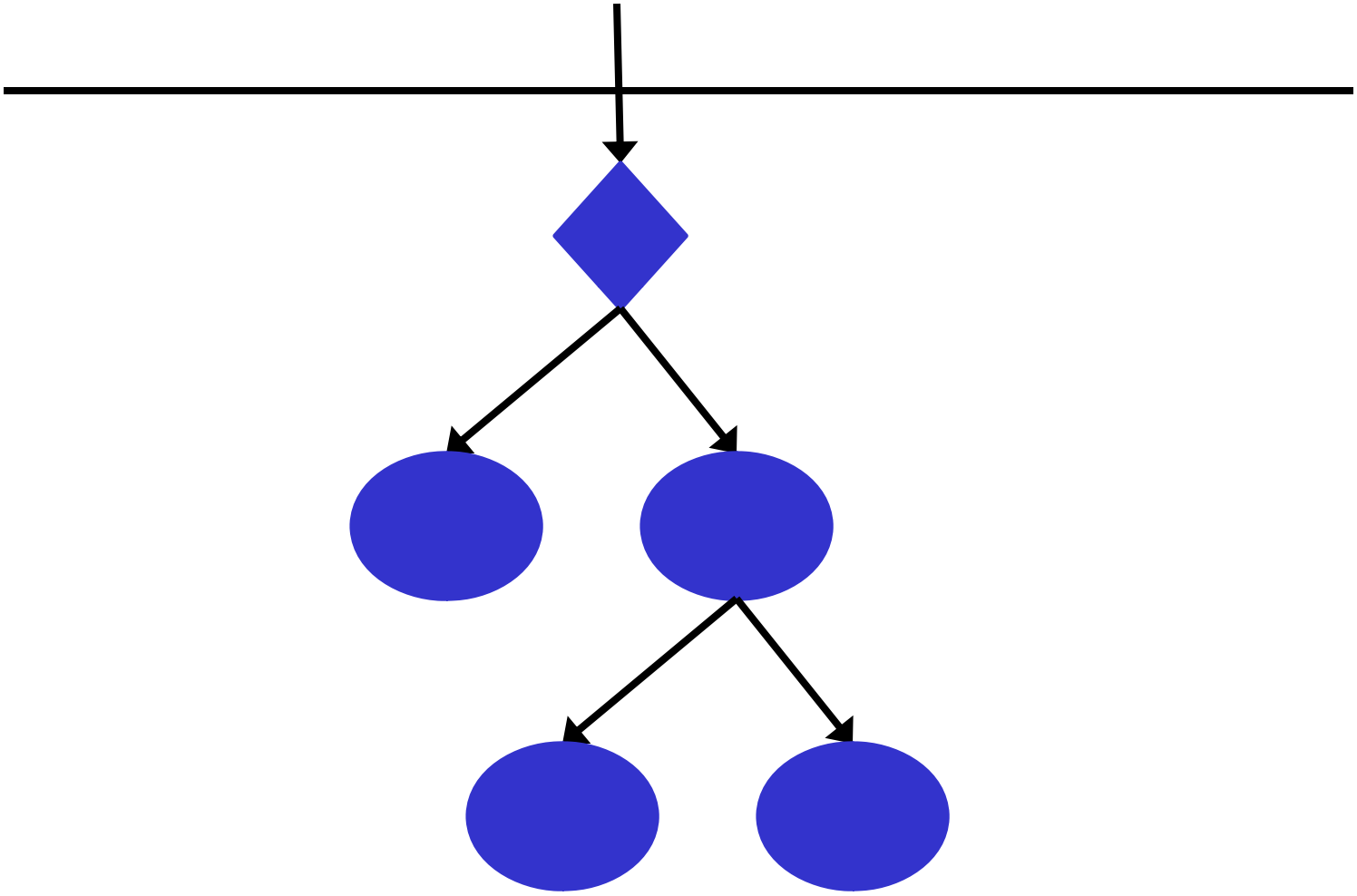
Open Science Grid Glidein Usage (4 feb 2012)

	group	hours	percent
1	BMRB	10710.3	73.49%
2	Biochem_Attie	3660.2	25.11%
3	Statistics_Wahba	178.5	1.22%

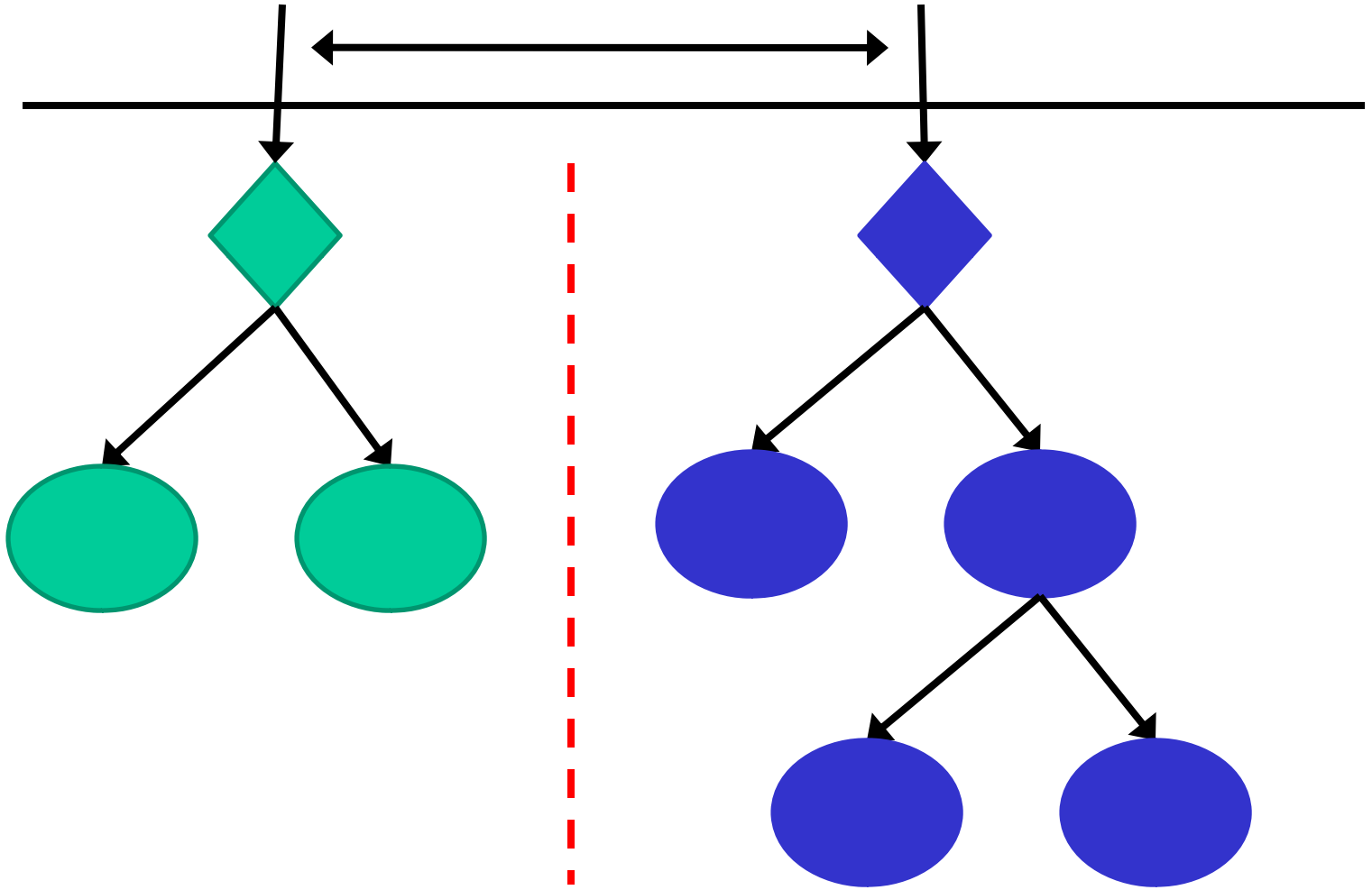
which came first? (causal models)



one causal driver



two linked causal drivers
pathways independent given drivers



causal architecture references

BIC: Schadt et al. (2005) *Nature Genet*

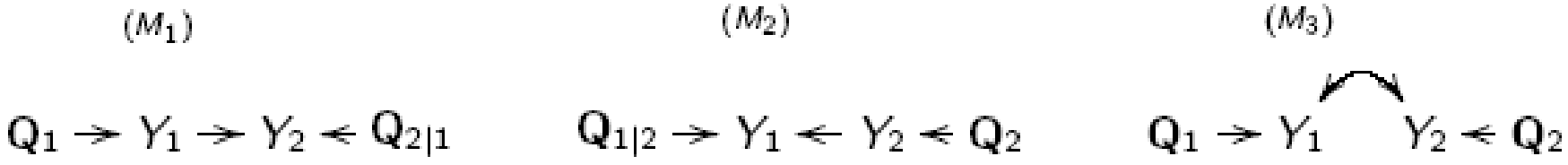
CIT: Millstein et al. (2009) *BMC Genet*

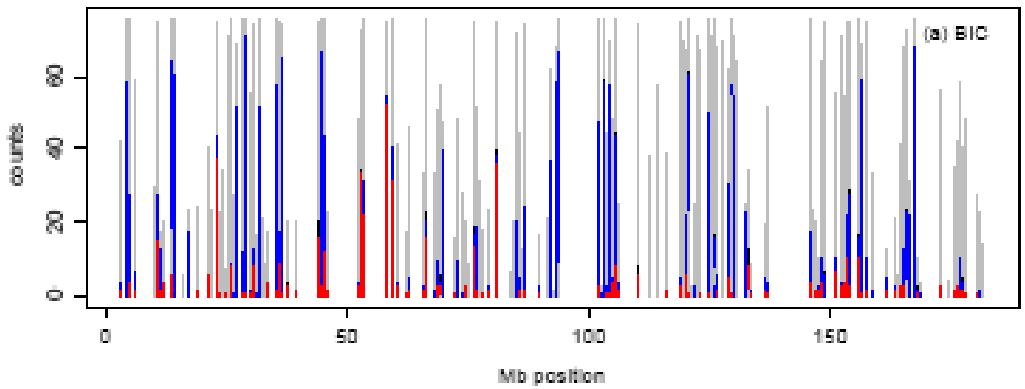
Aten et al. Horvath (2008) *BMC Sys Bio*

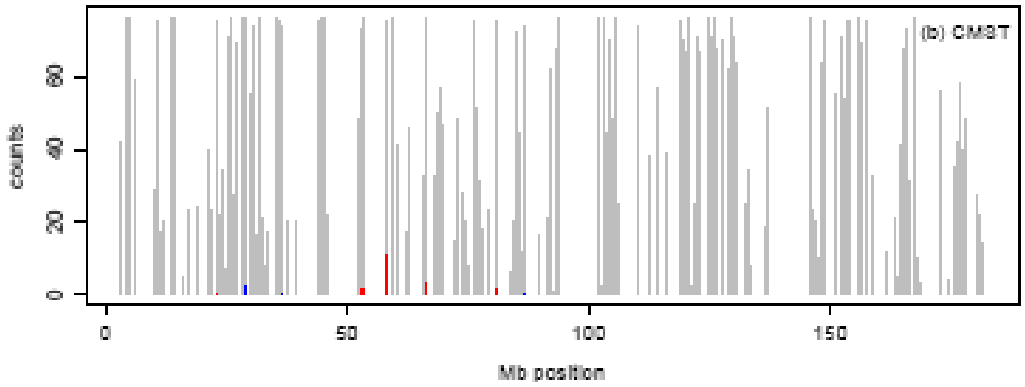
CMST: Chaibub Neto et al. (2012) *Genetics* (in review)

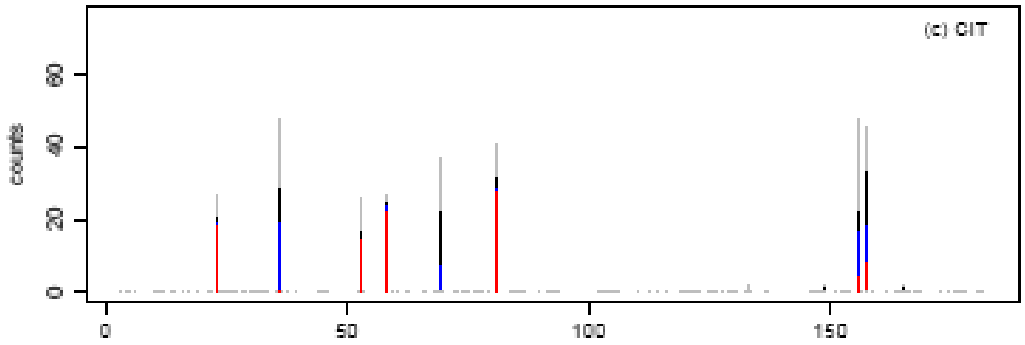
data: Ghazalpour et al. (2008) *PLoS Genetics*

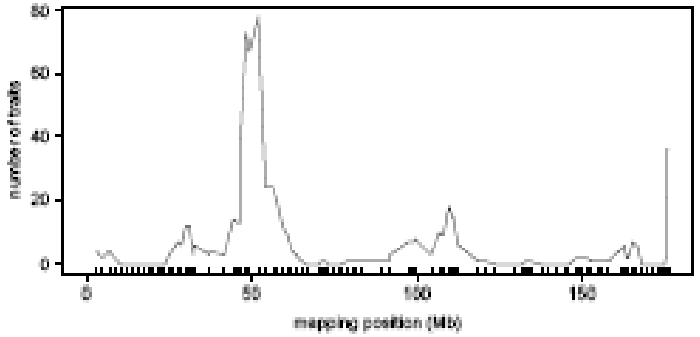
Extends Vuong's model selection tests to the comparison of 3, possibly **misspecified**, models.

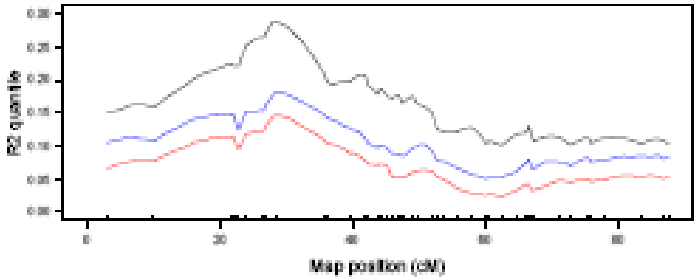










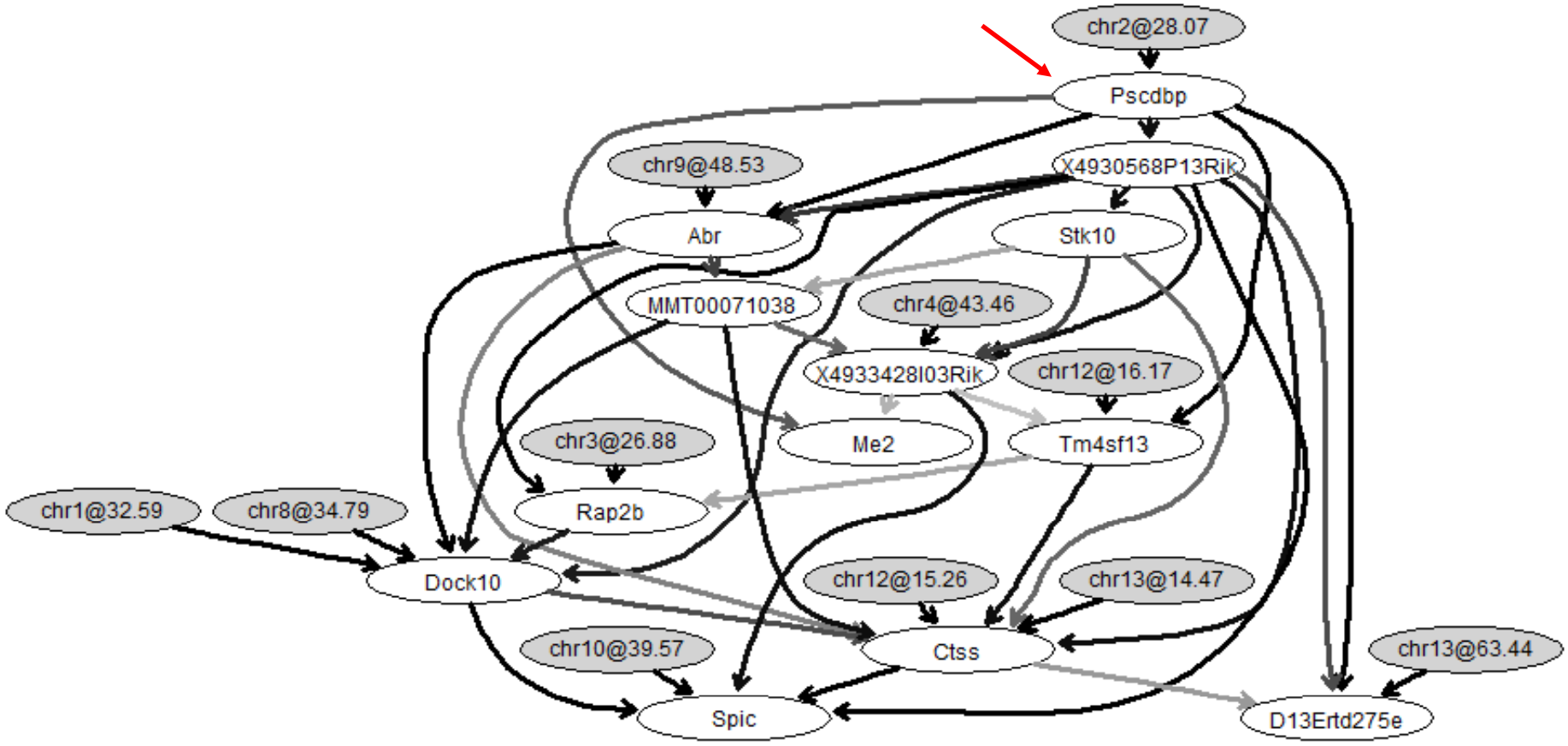


Analysis restricted to 78 traits composing a hotspot around 54.2Mb.

This collection of traits enriches for "immune system process".

Pscdbp, the local trait at 58.4Mb, is a transcription factor.

BxH ApoE-/- causal network for transcription factor Pscdbp



causal phenotype networks

- goal: mimic biochemical pathways with directed (causal) networks
- problem: association (correlation) does not imply causation
- resolution: bring in driving causes
  - genotypes (at conception)
  - processes earlier in time

QTL-driven directed graphs

given genetic architecture (QTLs), what causal network structure is supported by data?

R/qdg available at www.github.org/byandell

references

Chaibub Neto, Ferrara, Attie, Yandell (2008) Inferring causal phenotype networks from segregating populations. *Genetics 179*: 1089-1100. [doi:genetics.107.085167]

Ferrara et al. Attie (2008) Genetic networks of liver metabolism revealed by integration of metabolic and transcriptomic profiling. *PLoS Genet 4*: e1000034. [doi:10.1371/journal.pgen.1000034]

causal graphical models in systems genetics

What if genetic architecture and causal network are unknown? jointly infer both using iteration

Chaibub Neto, Keller, Attie, Yandell (2010) Causal Graphical Models in Systems Genetics: a unified framework for joint inference of causal network and genetic architecture for correlated phenotypes. *Ann Appl Statist 4*: 320-339. [doi:10.1214/09-AOAS288]

R/qtlnet available from www.github.org/byandell

Related references

Schadt et al. Lusis (2005 *Nat Genet*); Li et al. Churchill (2006 *Genetics*); Chen Emmert-Streib Storey(2007 *Genome Bio*); Liu de la Fuente Hoeschele (2008 *Genetics*); Winrow et al. Turek (2009 *PLoS ONE*); Hageman et al. Churchill (2011 *Genetics*)

Basic idea of QTLnet

iterate between finding QTL and network

genetic architecture given causal network

- trait y depends on parents pa(y) in network
- QTL for y found conditional on pa(y)
  - Parents pa(y) are interacting covariates for QTL scan

causal network given genetic architecture

- build (adjust) causal network given QTL
- each direction change may alter neighbor edges

scaling up to larger networks

reduce complexity of graphs

- use prior knowledge to constrain valid edges
- restrict number of causal edges into each node

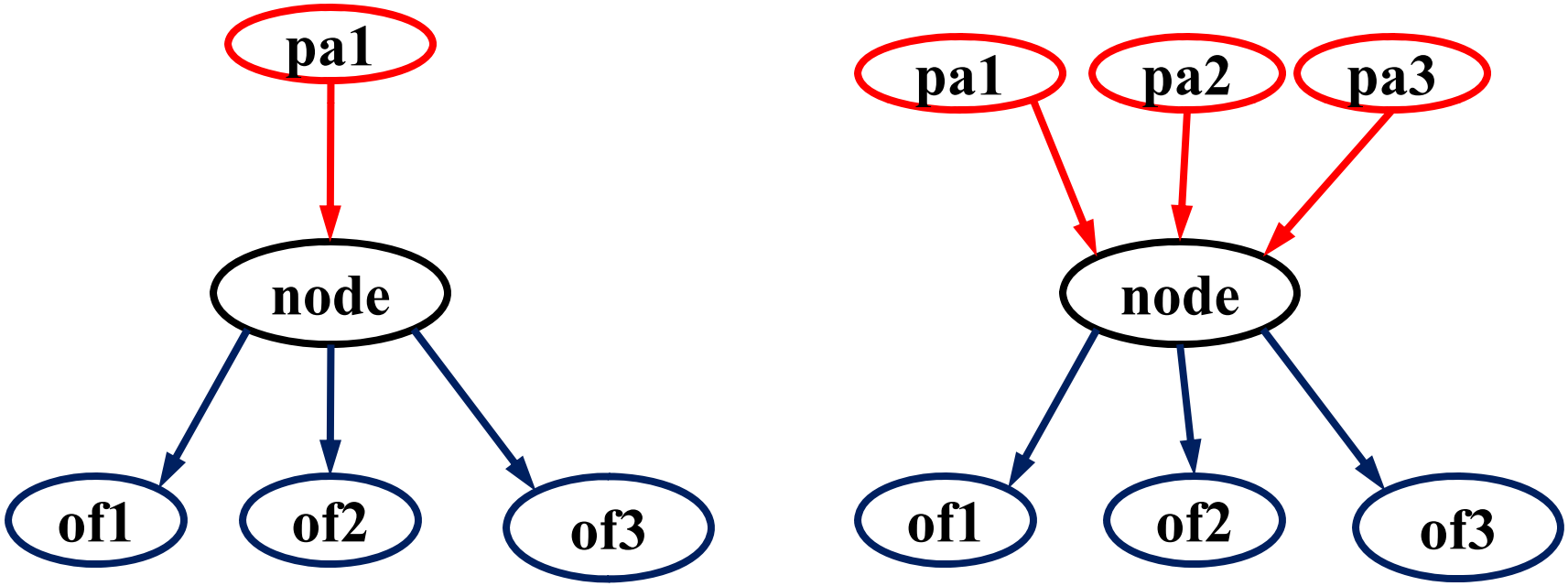
make task parallel: run on many machines

- pre-compute conditional probabilities
- run multiple parallel Markov chains

rethink approach

- LASSO, sparse PLS, other optimization methods

graph complexity with node parents



parallel phases for larger projects

Phase 1: identify parents

Phase 2: compute BICs

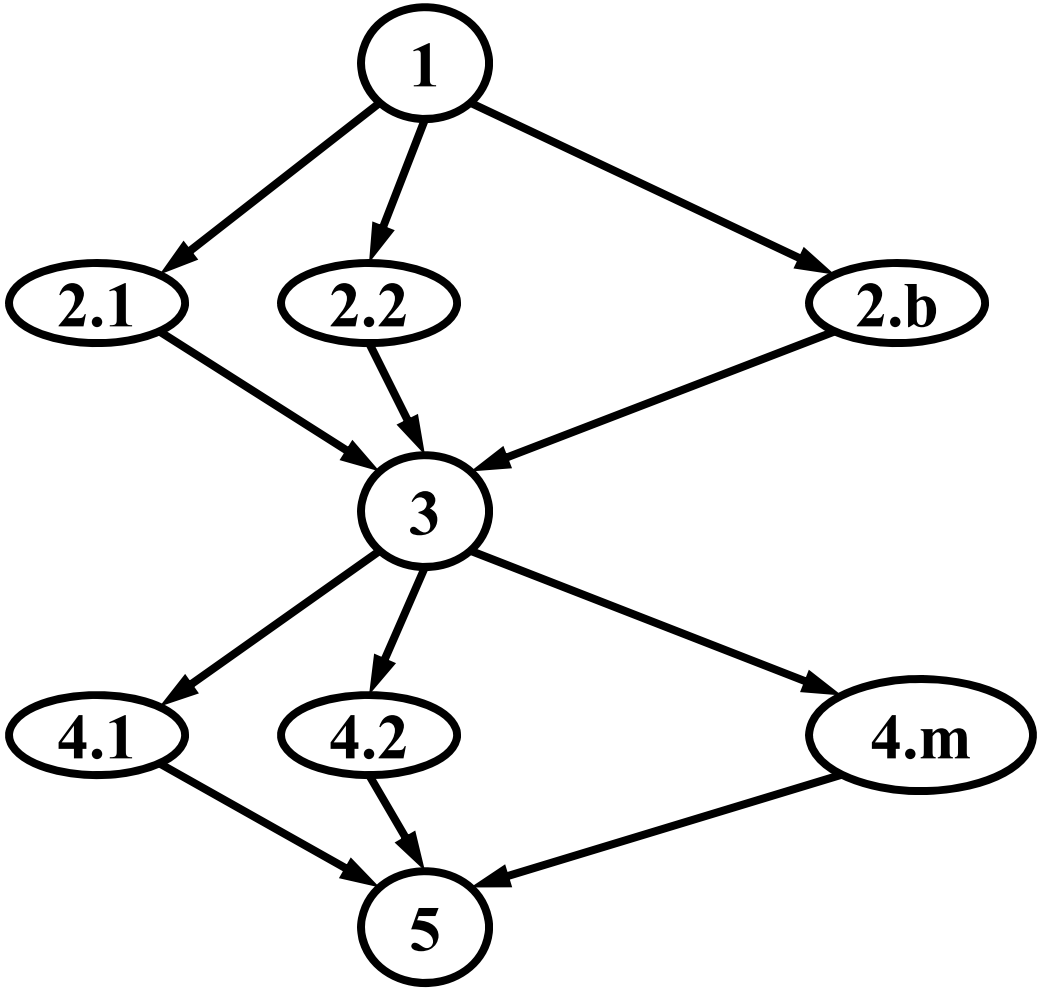
BIC = LOD – penalty

all possible parents to all nodes

Phase 3: store BICs

Phase 4: run Markov chains

Phase 5: combine results



parallel implementation

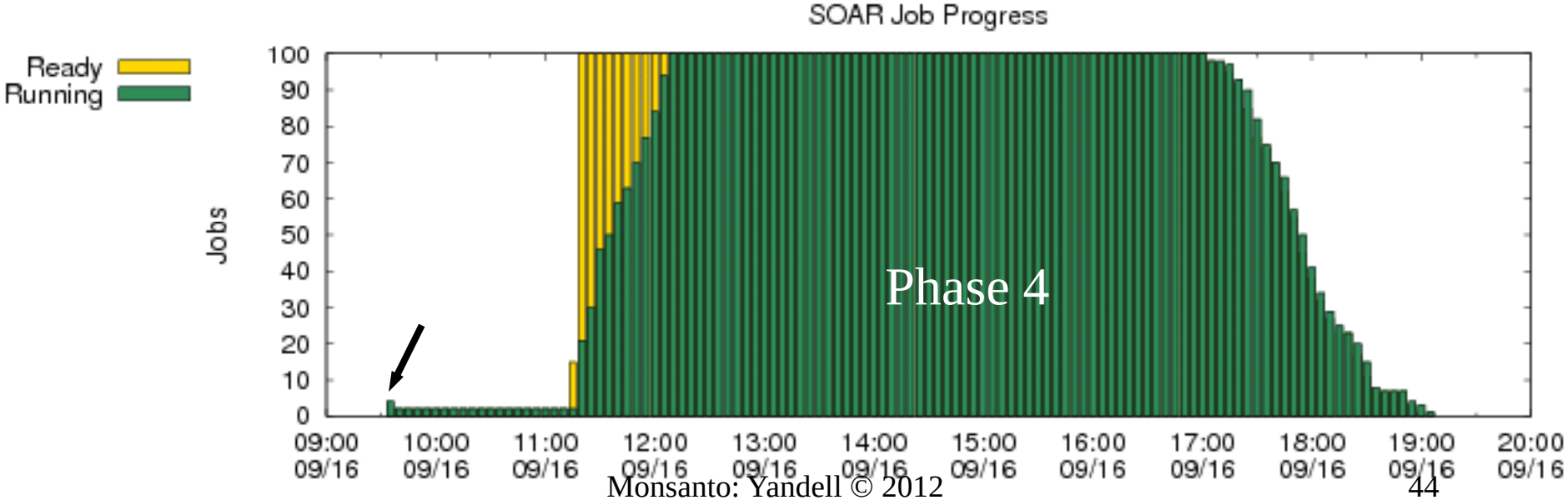
R/qtlnet available at www.github.org/byandell

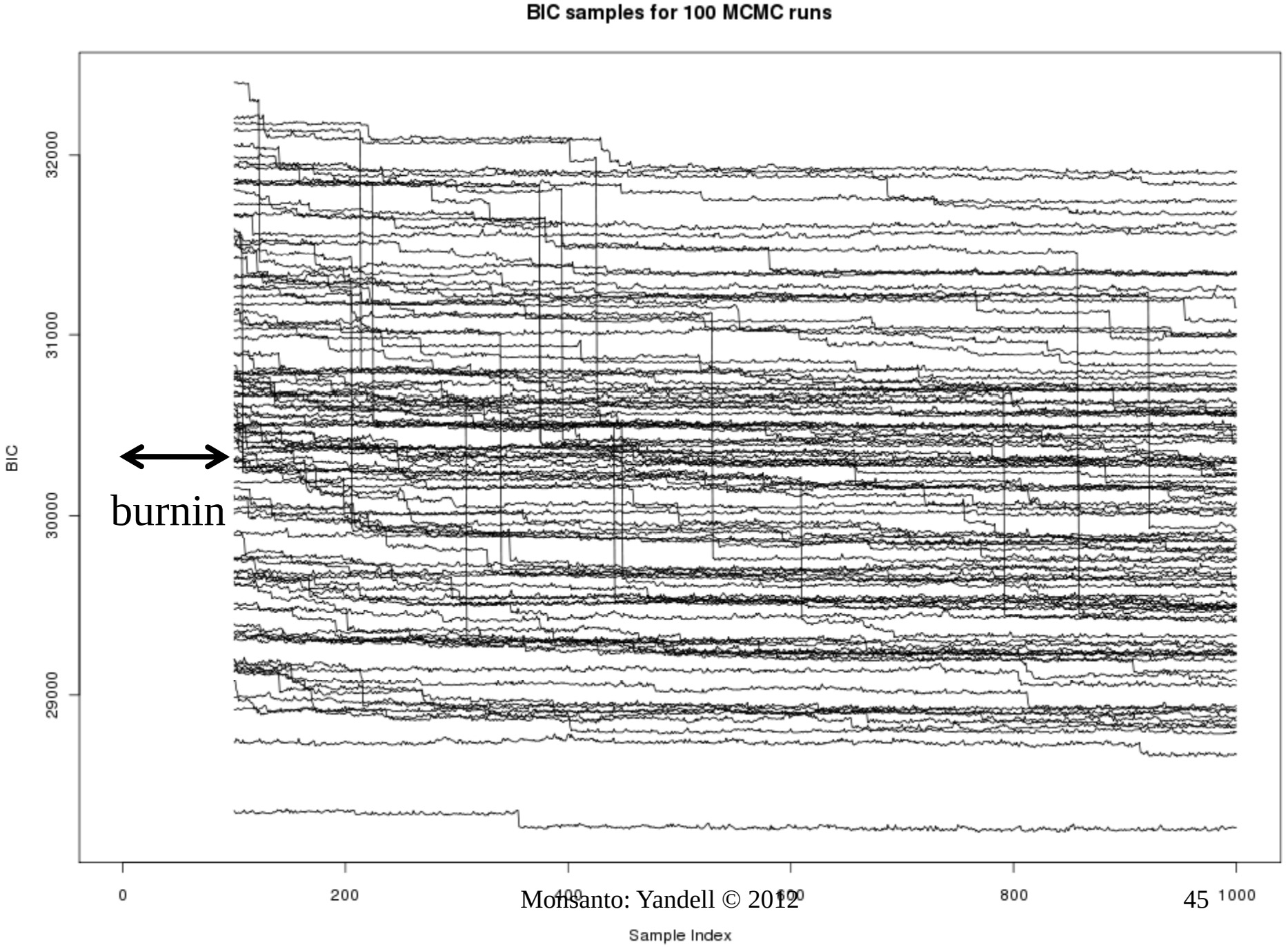
Condor cluster: chtc.cs.wisc.edu

System Of Automated Runs (SOAR)

~2000 cores in pool shared by many scientists

automated run of new jobs placed in project

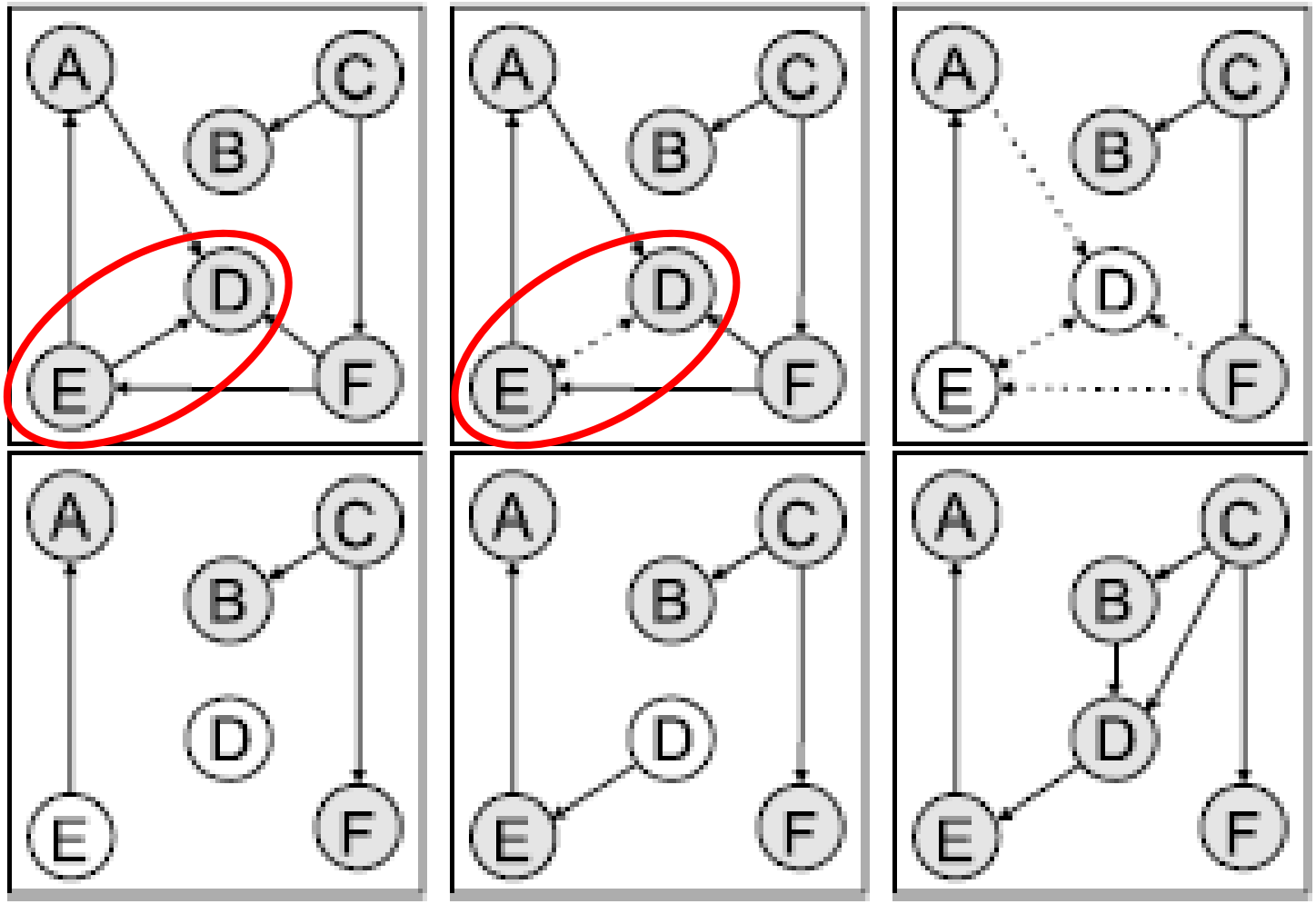


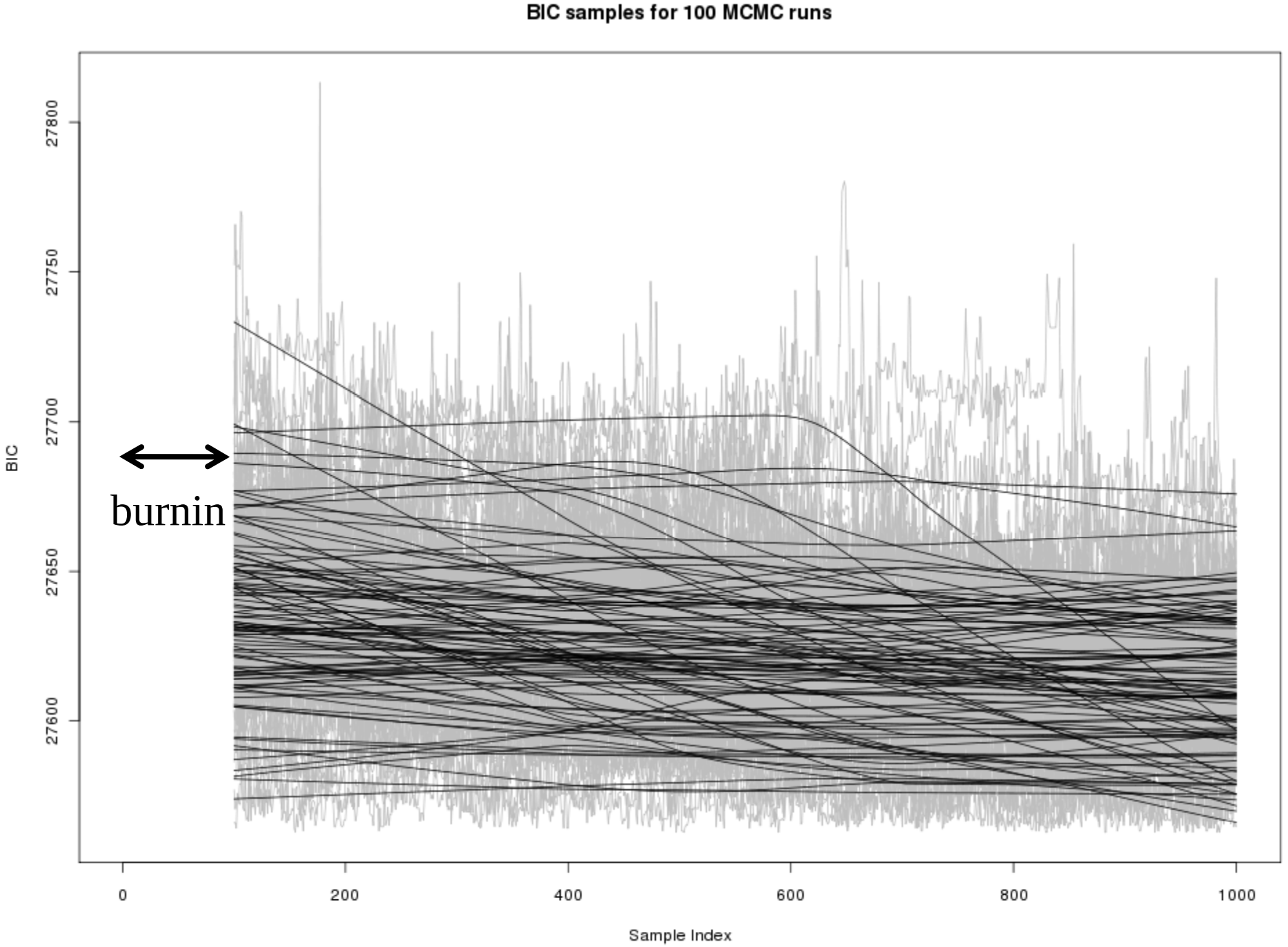

BIC samples for 100 MCMC runs
BIC
32000
31000
30000
29000
burnin
0
200
400
600
800
1000
Sample Index

neighborhood edge reversal

select edge
drop edge
identify parents

orphan nodes
reverse edge
find new parents




BIC samples for 100 MCMC runs
BIC
27800
27750
27700
27650
27600
burnin
0
200
400
600
800
1000
Sample Index

how to use functional information?

- functional grouping from prior studies
  - may or may not indicate direction
  - gene ontology (GO), KEGG
  - knockout (KO) panels
  - protein-protein interaction (PPI) database
  - transcription factor (TF) database
- methods using only this information
- priors for QTL-driven causal networks
  - more weight to local (*cis*) QTLs?

modeling biological knowledge

infer graph G from biological knowledge B

$\Pr(G \mid B, W) = \exp(- W * |B - G|) / \text{constant}$

B = prob of edge given TF, PPI, KO database

derived using previous experiments, papers, etc.

G = 0-1 matrix for graph with directed edges

W = inferred weight of biological knowledge

W=0: no influence; W large: assumed correct

Werhli and Husmeier (2007) *J Bioinfo Comput Biol*

combining eQTL and bio knowledge

probability for graph G and bio-weights W

given phenotypes Y, genotypes X, bio info B

$$\Pr(G, W \mid Y, Q, B) = \Pr(Y|G,Q)\Pr(G|B,W)\Pr(W|B)$$

$\Pr(Y|G,Q)$ is genetic architecture (QTLs)

using parent nodes of each trait as covariates

$\Pr(G|B,W)$ is relation of graph to biological info

see previous slides

put priors on QTL based on proximity, biological info

Moon JY, Chaibub Neto E, Deng X, Yandell BS (2011) Growing graphical models to infer causal phenotype networks. In *Probabilistic Graphical Models Dedicated to Applications in Genetics.* Sinoquet C, Mourad R, eds. (in review)