Causal Network Models for Correlated Quantitative Traits

Brian S. Yandell

UW-Madison

September 2013

www.stat.wisc.edu/~yandell/statgen

outline

- how are correlation and causation connected?
- hotspots: do many traits really map to the same locus?
- causal pairs: how to find causal drivers for hotspots?
- causal networks: how to infer signal cascades?
- how to scale up to larger problems

••• correlation & causation •••

“The old view of cause and effect … could only fail; things are not in our experience either independent or causative. All classes of phenomena are linked together, and the problem in each case is how close is the degree of association.”

Karl Pearson (1911)

The Grammar of Science

“The ideal … is the study of the direct influence of one condition on another …[when] all other possible causes of variation are eliminated.... The degree of correlation between two variables … [includes] all connecting paths of influence…. [Path coefficients combine] knowledge of … correlation among the variables in a system with … causal relations.

Sewall Wright (1921)

Correlation and causation. *J Agric Res*

"Causality is not mystical or metaphysical. It can be understood in terms of simple processes, and it can be expressed in a friendly mathematical language, ready for computer analysis."

Judea Pearl (2000)

Causality: Models, Reasoning and Inference

problems and controversies

- Correlation does not imply causation.
  - Common knowledge in field of statistics.
- Steady state (static) measures may not reflect dynamic processes.
  - Przytycka and Kim (2010) *BMC Biol*
  - Blair, Kleibenstein, Churchill (2012) *PLoS Comp Bio*
- Population-based estimates may not reflect within-individual processes.

randomization and causation

- RA Fisher (1926) *Design of Experiments*
- control other known factors
- randomize assignment of treatment
  - no causal effect of individuals on treatment
  - no common cause of treatment and outcome
  - reduce chance correlation with unknown factors
- conclude (subsequent) outcome differences are caused by (due to) treatment

correlation and causation

- temporal aspect: cause before reaction
  - genotype (usually) drives phenotype
  - phenotypes in time series
  - *but* time order is not enough
- axioms of causality
  - transitive: if A → B, B → C, then A → C
  - local (Markov): events have proximate causes
  - asymmetric: if A → B, then not B → A
- Shipley (2000) *Cause and Correlation in Biology*

causation casts probability shadows

- causal relationship
  - $Y_1 \rightarrow Y_2 \rightarrow Y_3$
- conditional probability
  - $\Pr(Y_1) * \Pr(Y_2 \mid Y_1) * \Pr(Y_3 \mid Y_2)$
- linear model
  - $Y_1 = \mu_1 + \mathrm{e}$
  - $Y_2 = \mu_2 + \beta_1 \bullet Y_1 + \mathrm{e}$
  - adding in QTLs: $Q_1 \rightarrow Y_1 \rightarrow Y_2 \leftarrow Q_2$
  - $Y_1 = \mu_1 + \theta_1 \bullet Q_1 + \mathrm{e}$
  - $Y_2 = \mu_2 + \beta_1 \bullet Y_1 + \theta_2 \bullet Q_2 + \mathrm{e}$

organizing correlated traits

- functional grouping from prior studies
  - GO, KEGG; KO panels; TF and PPI databases
- co-expression modules (Horvath's WGCNA)
- eQTL hotspots (here briefly)
- traits used as covariates for other traits
  - does one trait essentially explain QTL of another?
- causal networks (here and Horvath talk)
  - modules of highly correlated traits

strategy from hotspot to causality

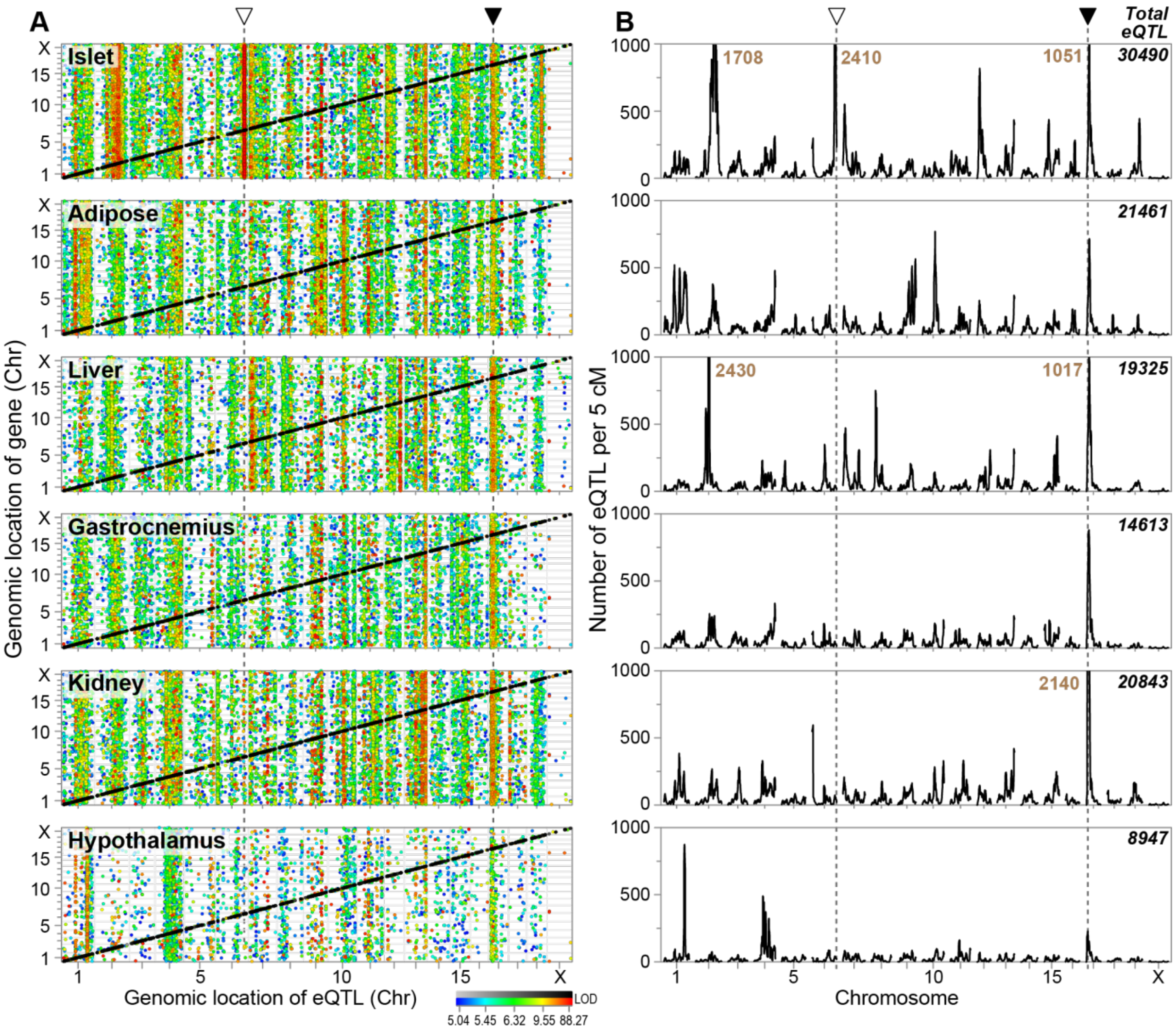
- detect "real" hotspots
  - hotspot = locus where many traits map
  - use permutation test to assess
- find causal architecture for each hotspot
  - causal model selection tests for pairs of traits
  - do local traits (at hotspot) drive other traits?
- build causal network for small set of traits
  - *cis* (local) trait ideally is top of signal cascade

••• hotspots •••

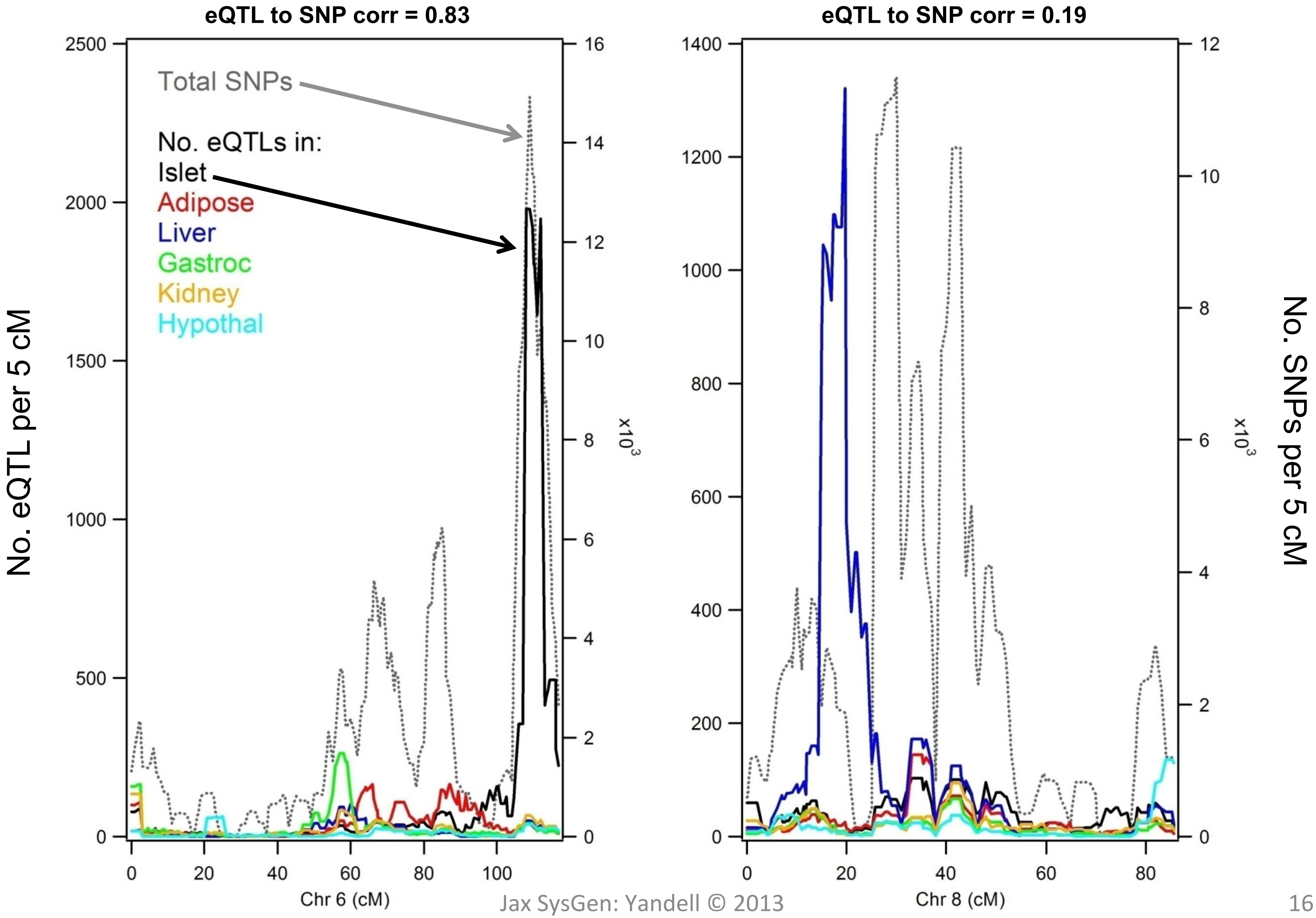
hotspots of correlated traits

- multiple correlated traits map to same locus
  - is this a real hotspot, or an artifact of correlation?
  - use QTL permutation across traits
- references
  - 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. [doi:10.1371/journal.pgen.1000232]
  - 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).

genetic architecture of gene expression in 6 tissues



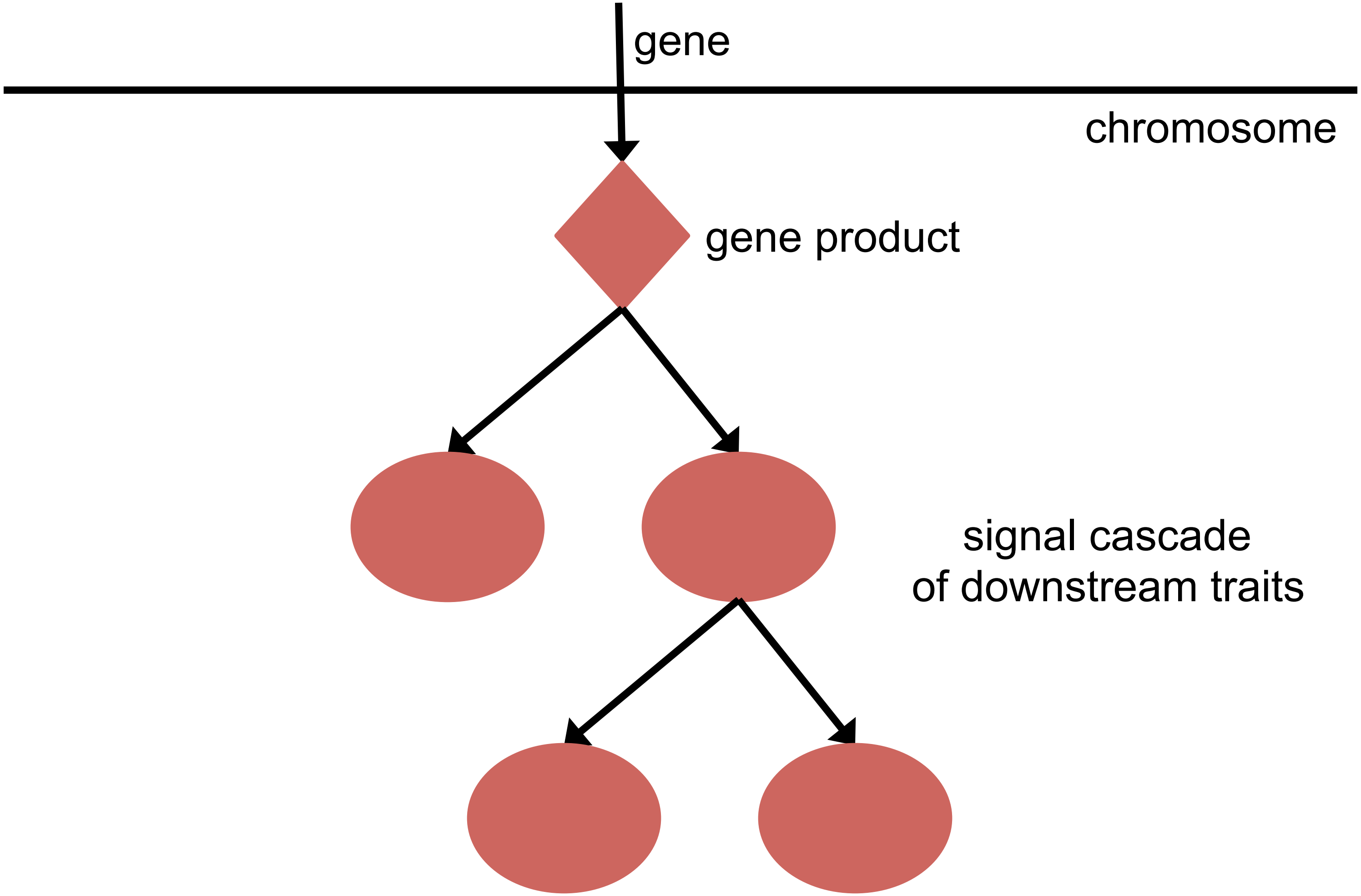
eQTL vs SNP architecture



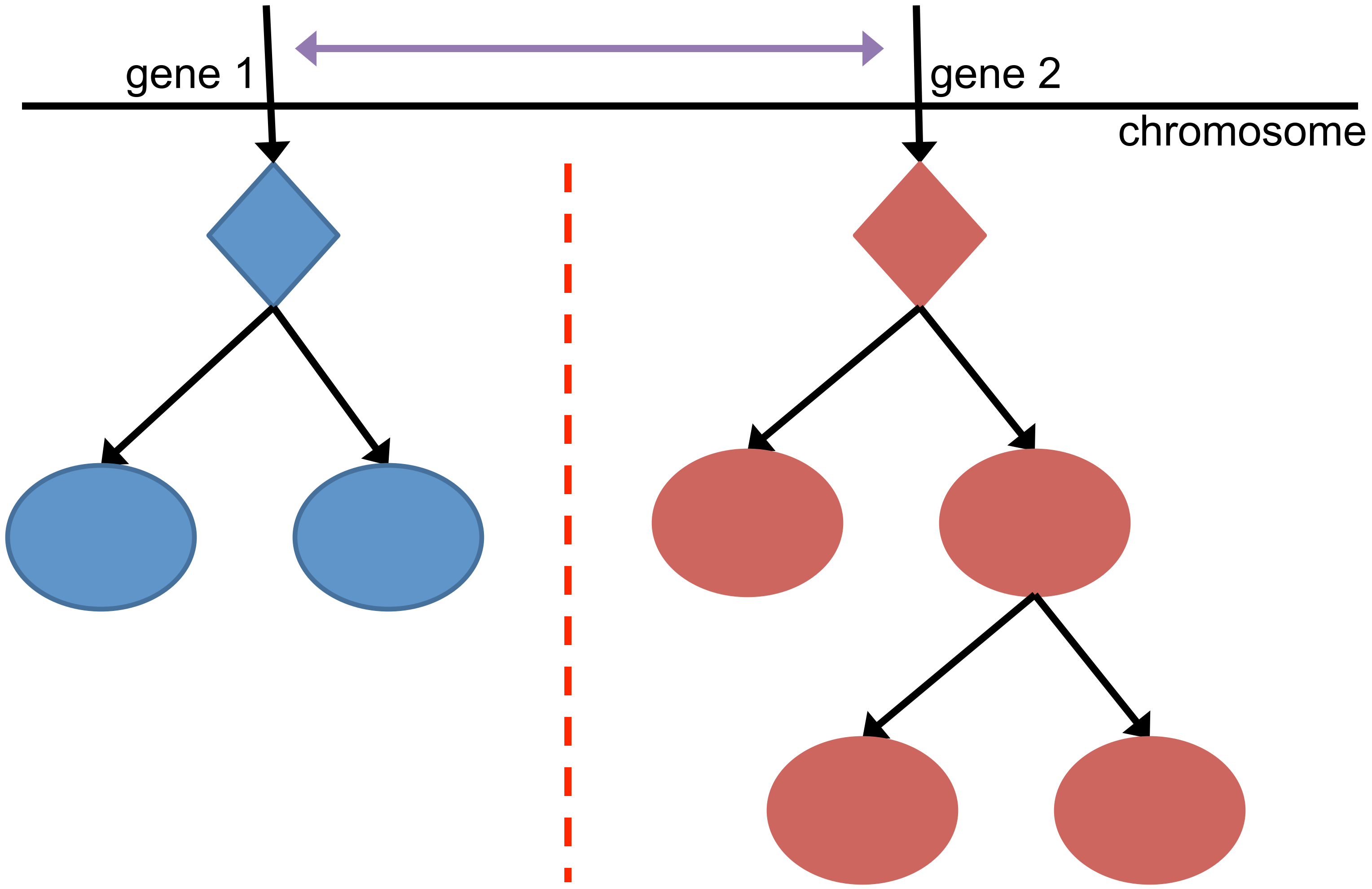
correlated traits in a hotspot why are traits correlated?

- Environmental
  - hotspot is spurious
- Genetics and causal networks
  - One causal driver at locus
    - Traits organized in causal cascade
  - Multiple causal drivers at locus
    - Several closely linked driving genes
    - Correlation due to close linkage
    - Separate networks are not causally related

one causal driver



two linked causal drivers pathways independent given drivers

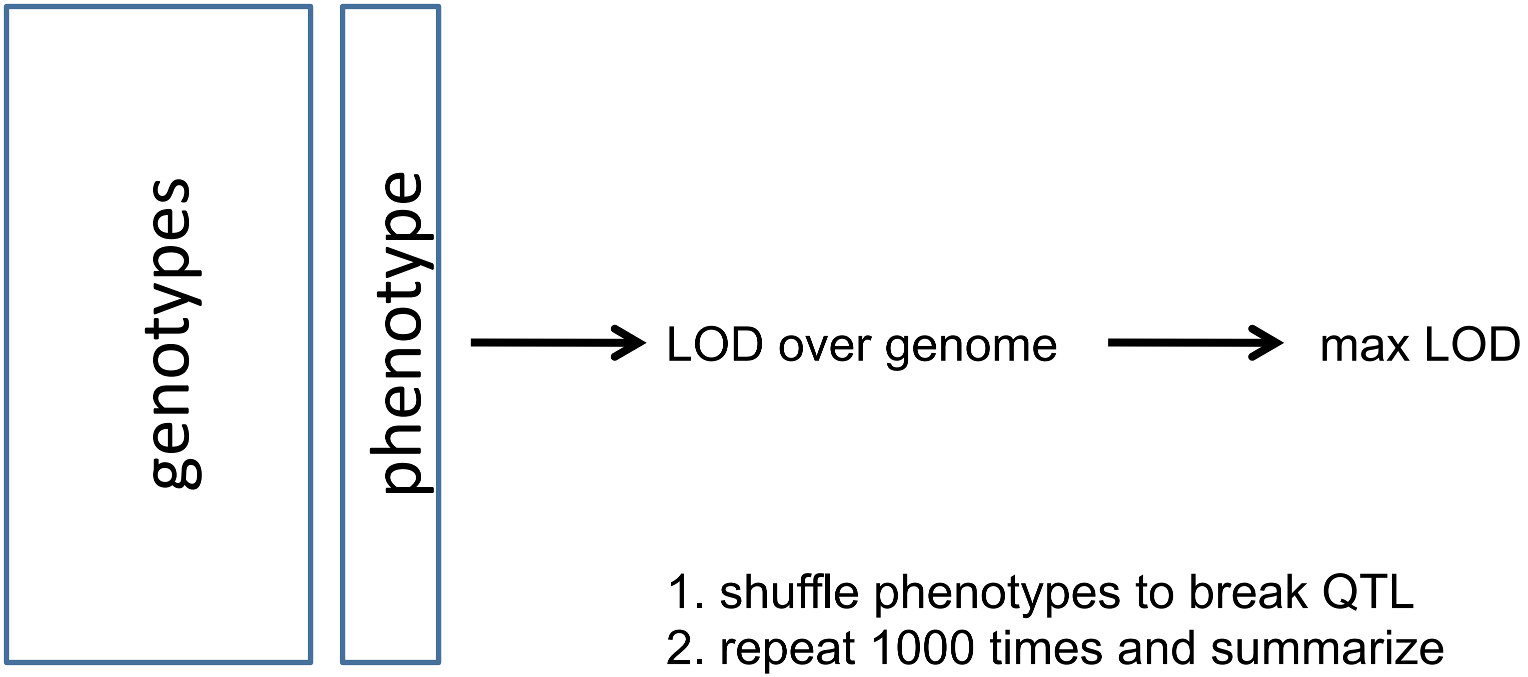


hotspot permutation test

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

- for original dataset and each permuted set:
  - set single trait LOD threshold *T*
    - use Churchill-Doerge (1994) permutations
  - count number of traits (*N*) with LOD above *T*
    - count for every locus (marker or pseudomarker)
    - smooth counts if markers are dense
- find count with at most 5% of permuted sets above (critical value) as count threshold
- conclude original counts above threshold are real

Single trait permutation schema



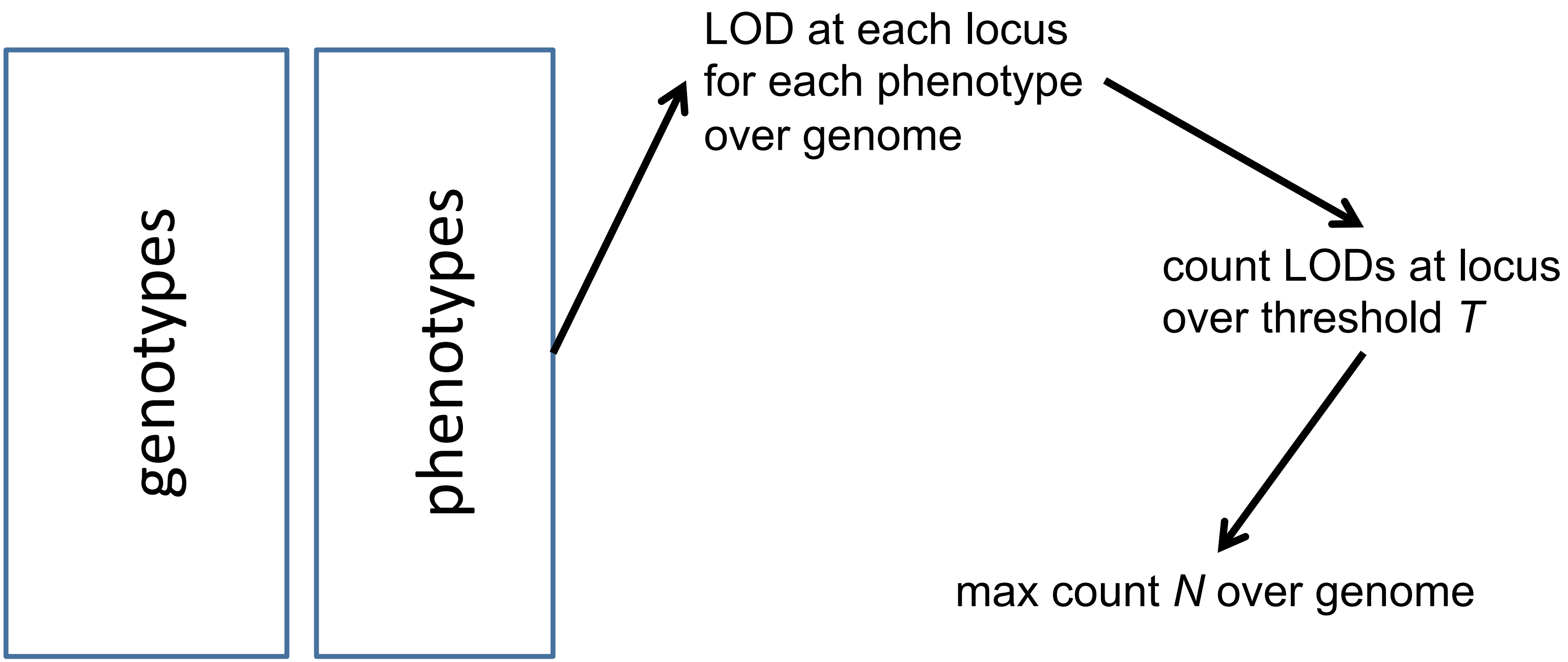
genotypes

phenotype

LOD over genome → max LOD

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

Hotspot permutation schema



genotypes

phenotypes

LOD at each locus
for each phenotype
over genome

count LODs at locus
over threshold *T*

max count *N* over genome

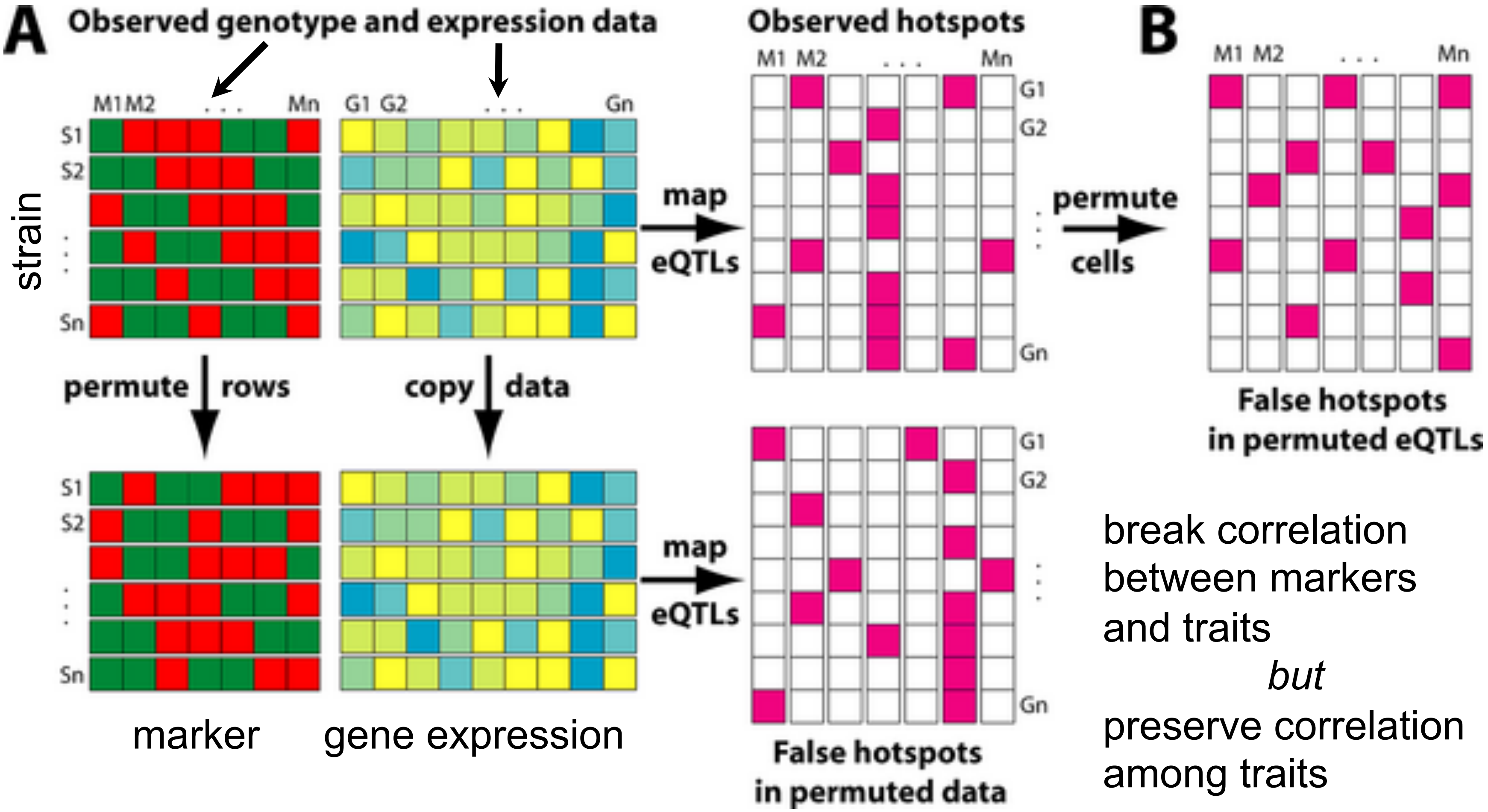
1. shuffle phenotypes by row to break QTL, keep correlation
2. repeat 1000 times and summarize

permutation across traits

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

right way

wrong way



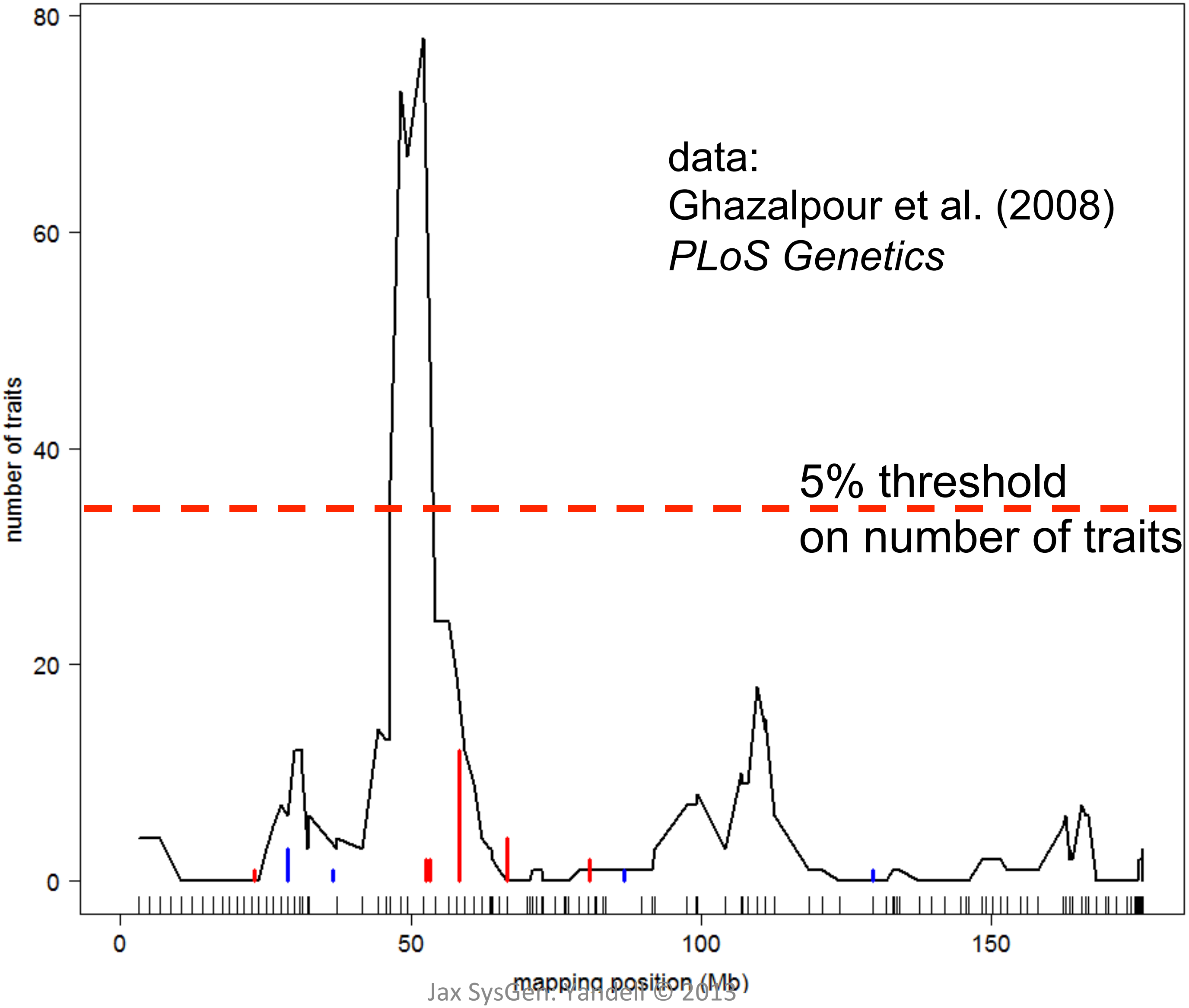
strain

marker

gene expression

break correlation
between markers
and traits
but
preserve correlation
among traits

BxH ApoE-/- chr 2: hotspot

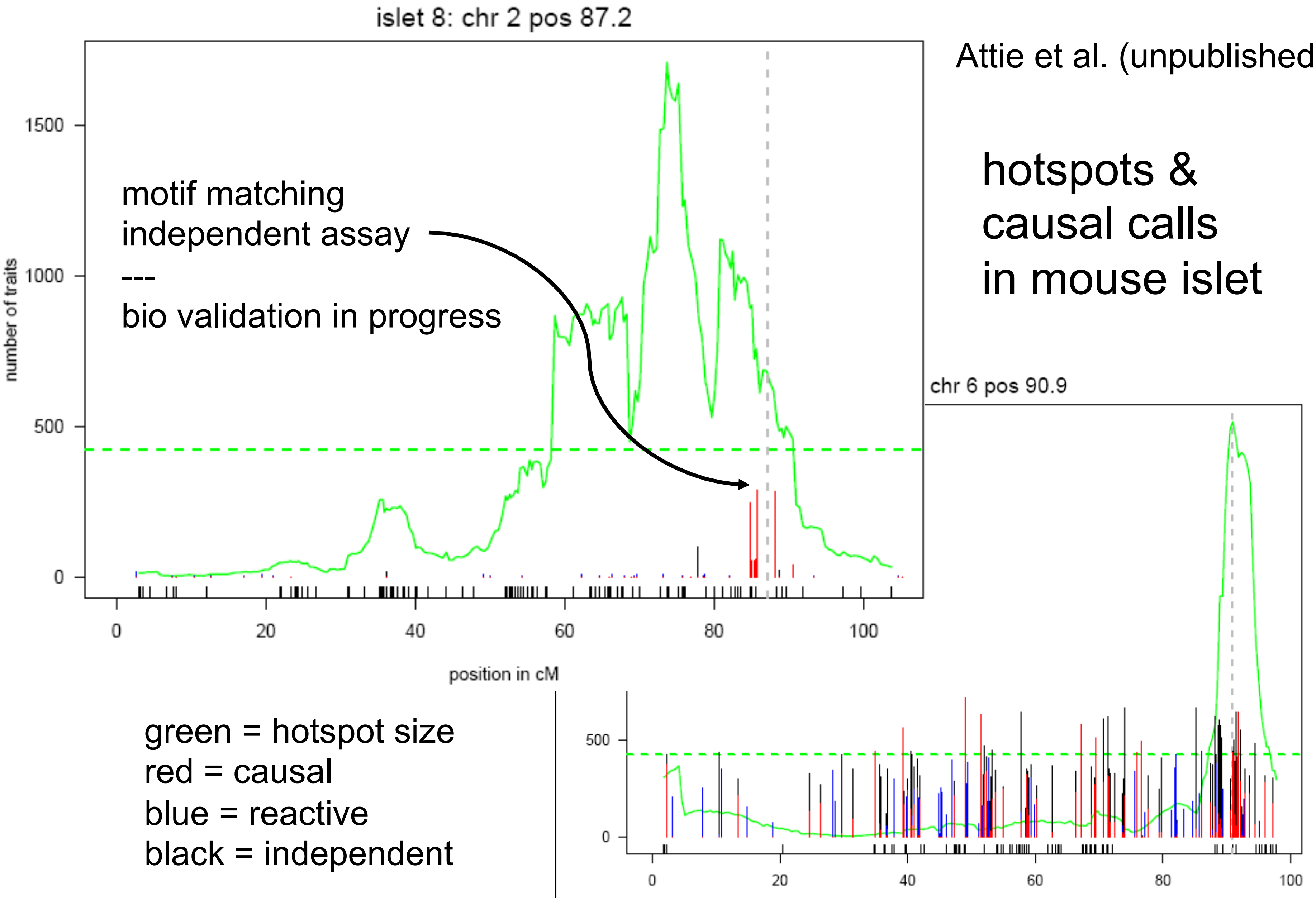


data:
Ghazalpour et al. (2008)
PLoS Genetics

5% threshold
on number of traits

hotspots & causal calls in mouse islet

Attie et al. (unpublished)



motif matching
independent assay
---
bio validation in progress

green = hotspot size
red = causal
blue = reactive
black = independent

quality vs. quantity in hotspots
(Chaibub Neto et al. 2012 *Genetics*)

- detect single trait with very large LOD
  - control FWER across genome *and* all traits
- find small hotspots with very significant traits
  - all traits have large LODs at same locus
  - maybe one strongly disrupted signal pathway?
- use sliding LOD threshold across hotspot sizes
  - small LOD threshold (~4) for large hotspots
  - large LOD threshold (~8) for small hotspots

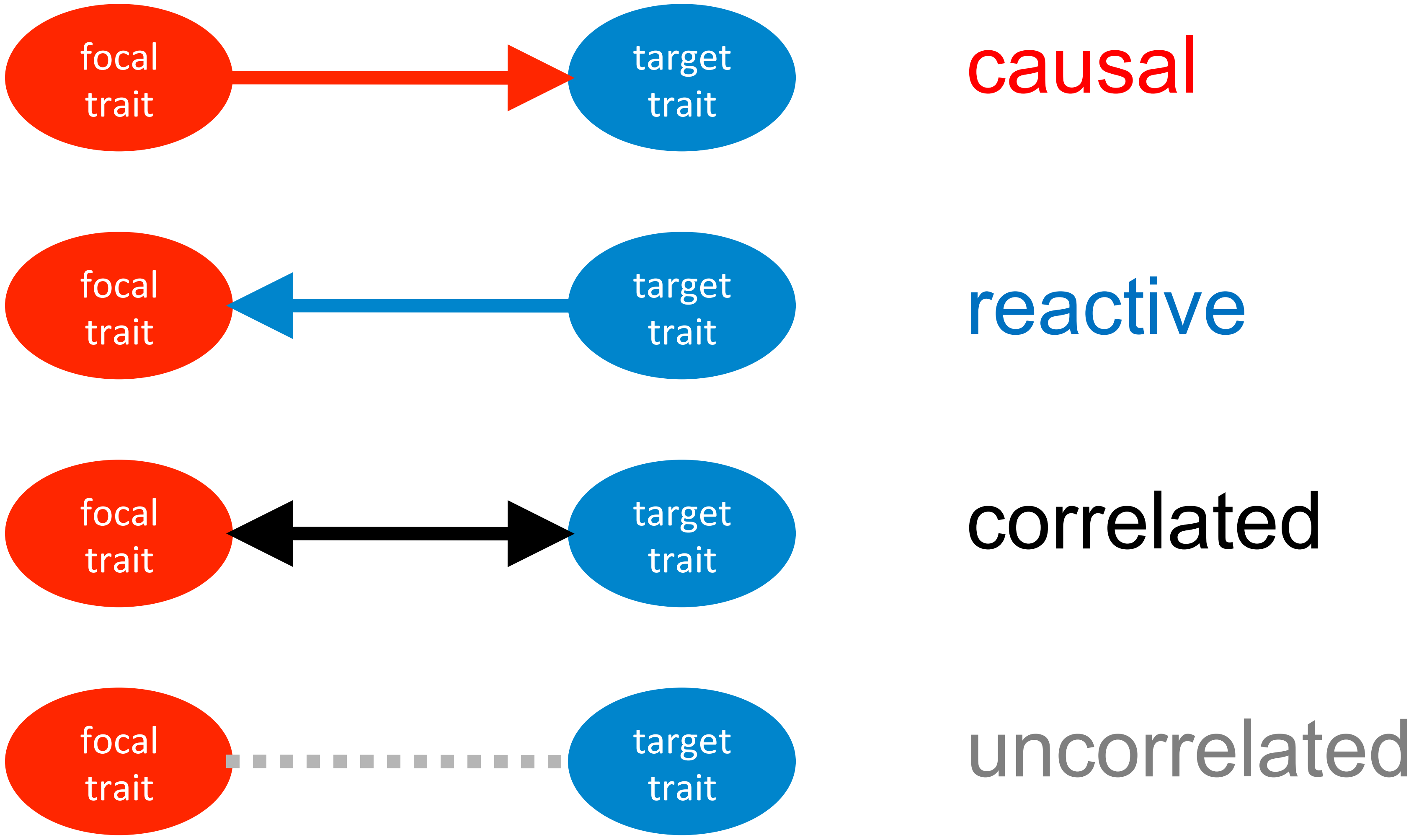
••• causal pairs •••

causal architecture

- focus on one hotspot
- identify all traits physically near hotspot
  - local traits (called *cis* if it also maps to hotspot)
- what traits are up/downstream of local trait?
  - focal trait causal to downstream target traits
  - record count at Mb position of focal gene
  - red = downstream, blue = upstream

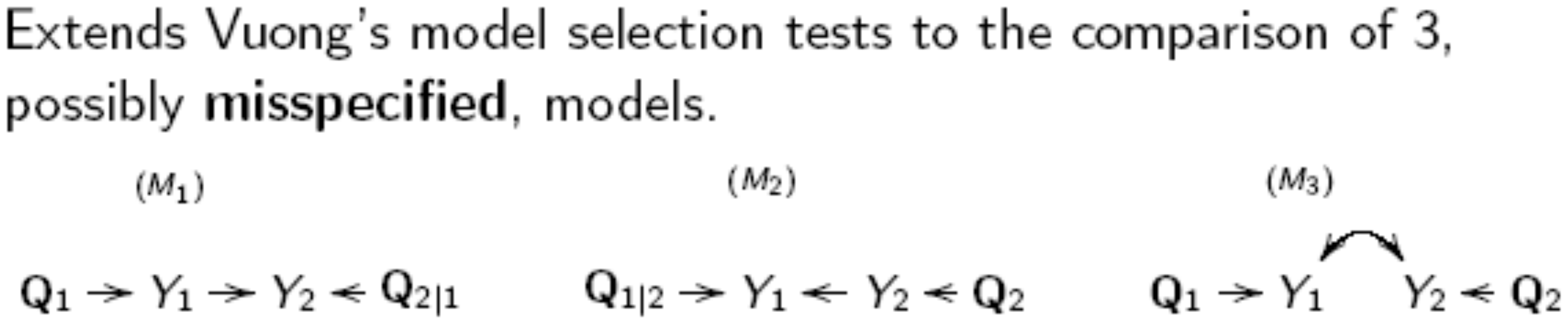
causal model selection choices

in context of larger, unknown network



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*



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

(M_1)

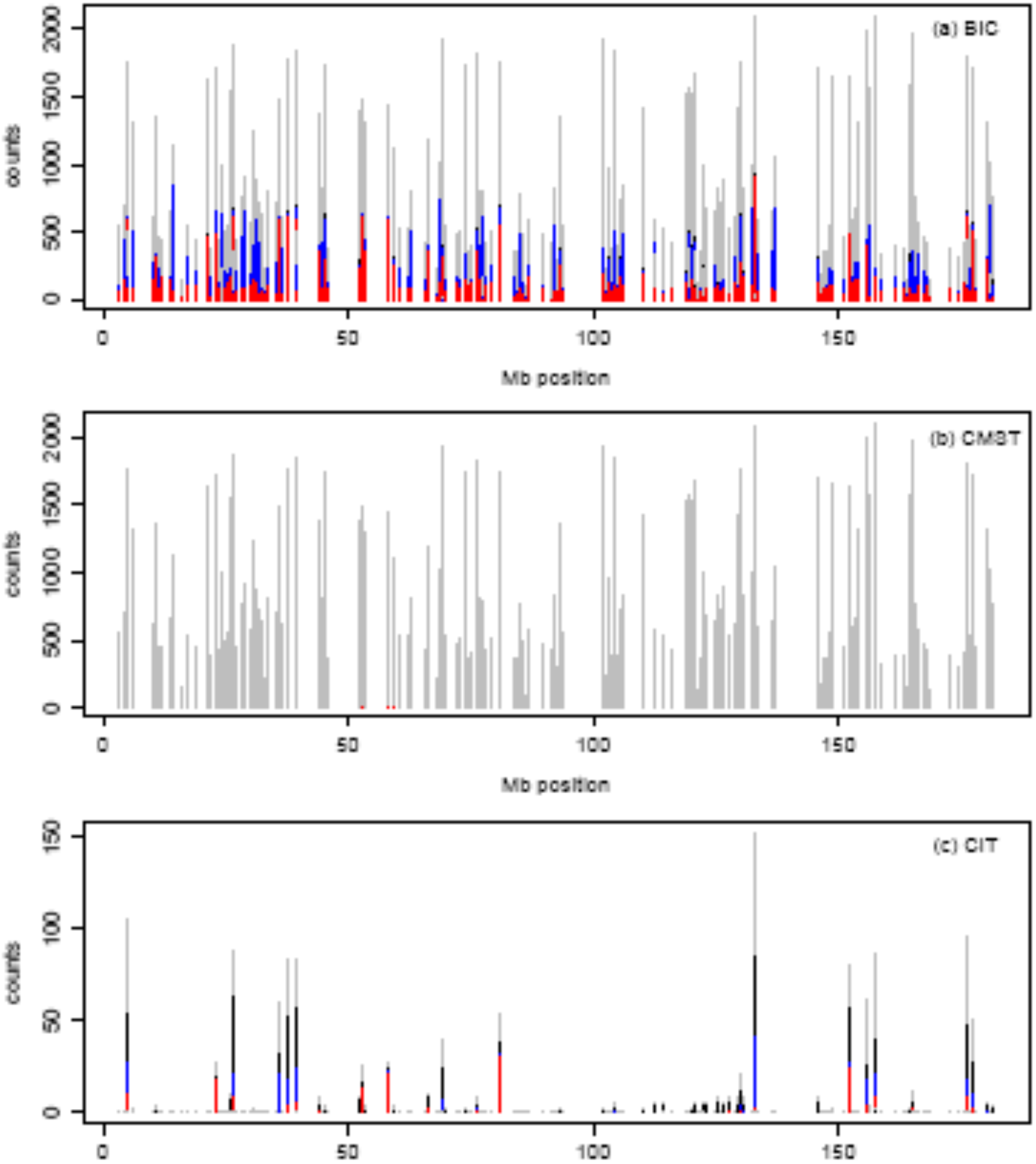
$$\mathbf{Q}_1 \to Y_1 \to Y_2 \leftarrow \mathbf{Q}_{2|1}$$

(M_2)

$$\mathbf{Q}_{1|2} \to Y_1 \leftarrow Y_2 \leftarrow \mathbf{Q}_2$$

(M_3)

$$\begin{tikzcd}\mathbf{Q}_1 \arrow[r] & Y_1 \arrow[r, phantom] \arrow[r, leftrightarrow, bend left] & Y_2 & \mathbf{Q}_2 \arrow[l]\end{tikzcd}$$

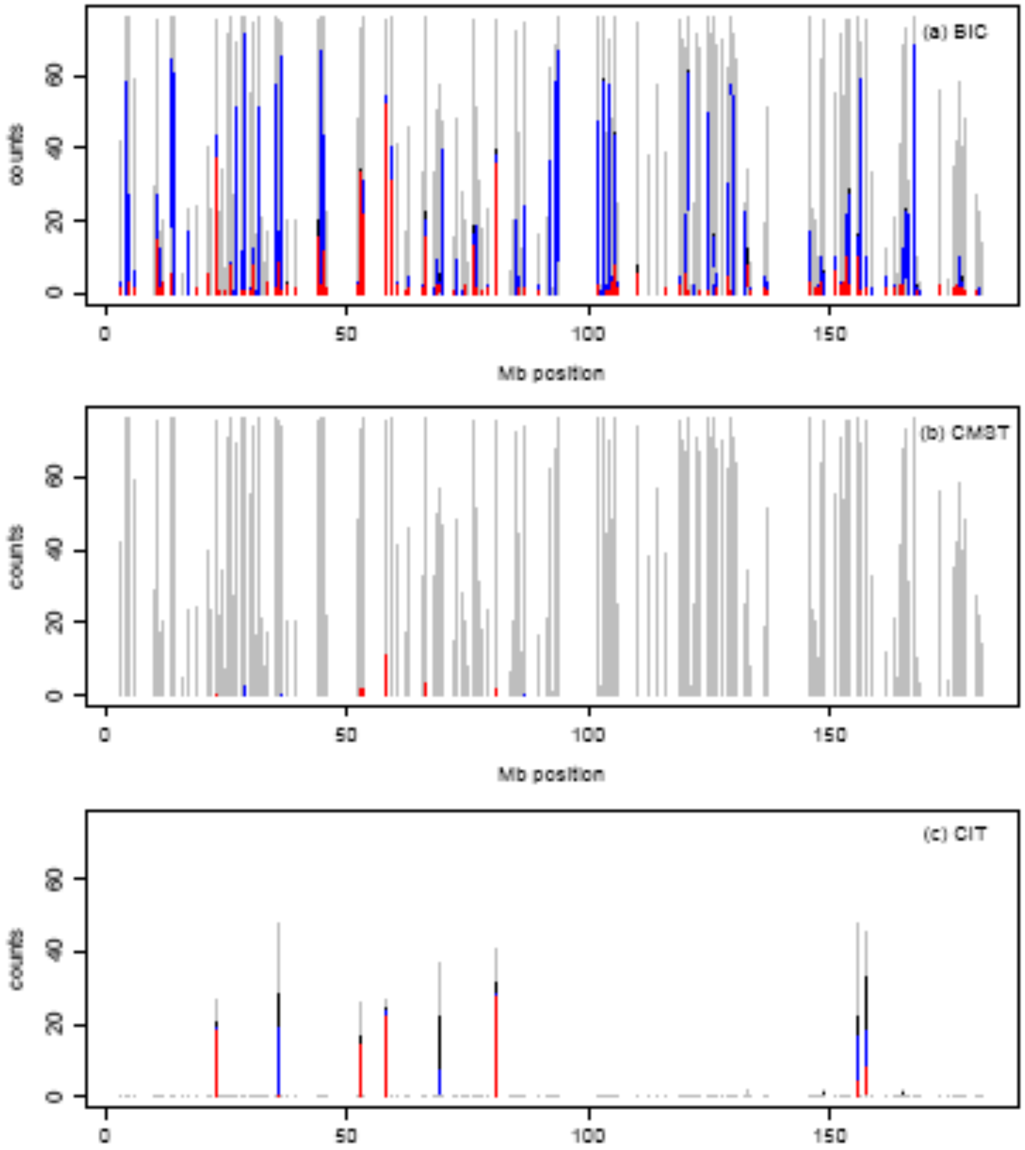


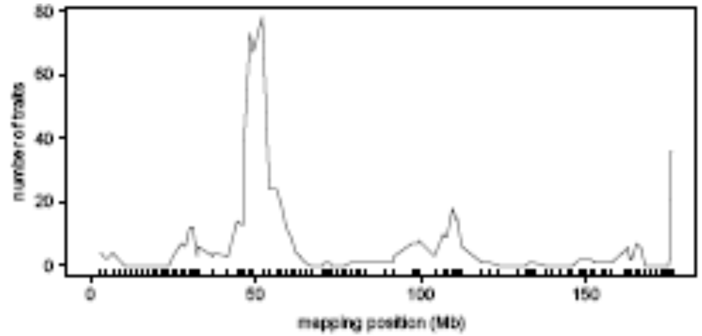
BxH ApoE-/- study
Ghazalpour et al. (2008)
PLoS Genetics

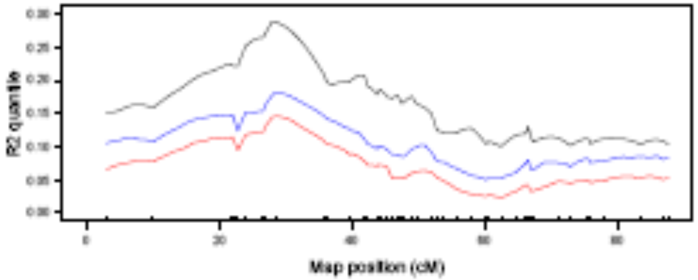
Liver expression data in a mice intercross.

3,421 transcripts and 1,065 markers.

261 transcripts physically located on chr 2.







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

This collection of traits enriches for “immune system process”.

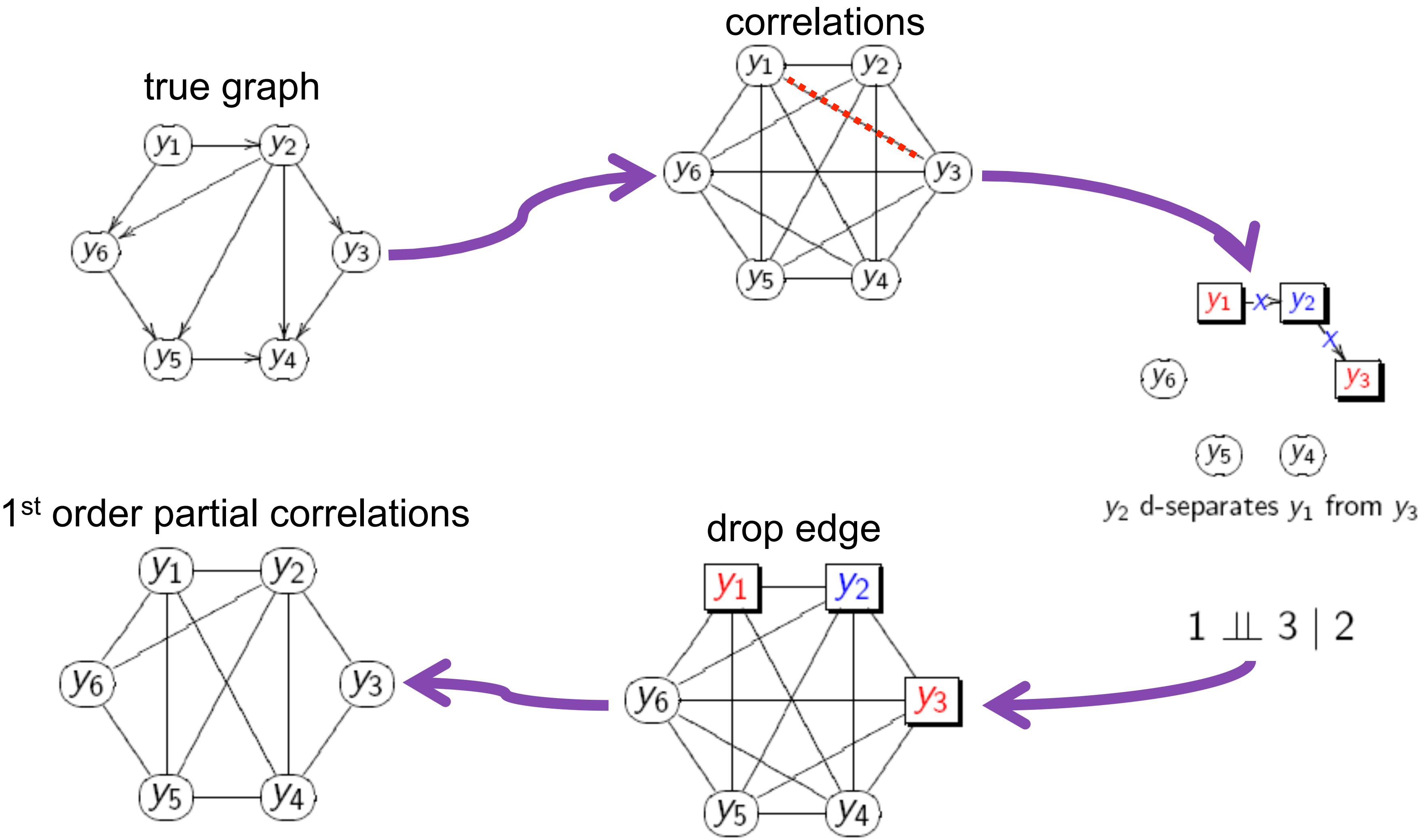
Pscdbp, the local trait at 58.4Mb, is a transcription factor.

••• causal networks •••

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]

partial correlation (PC) skeleton



true graph

correlations

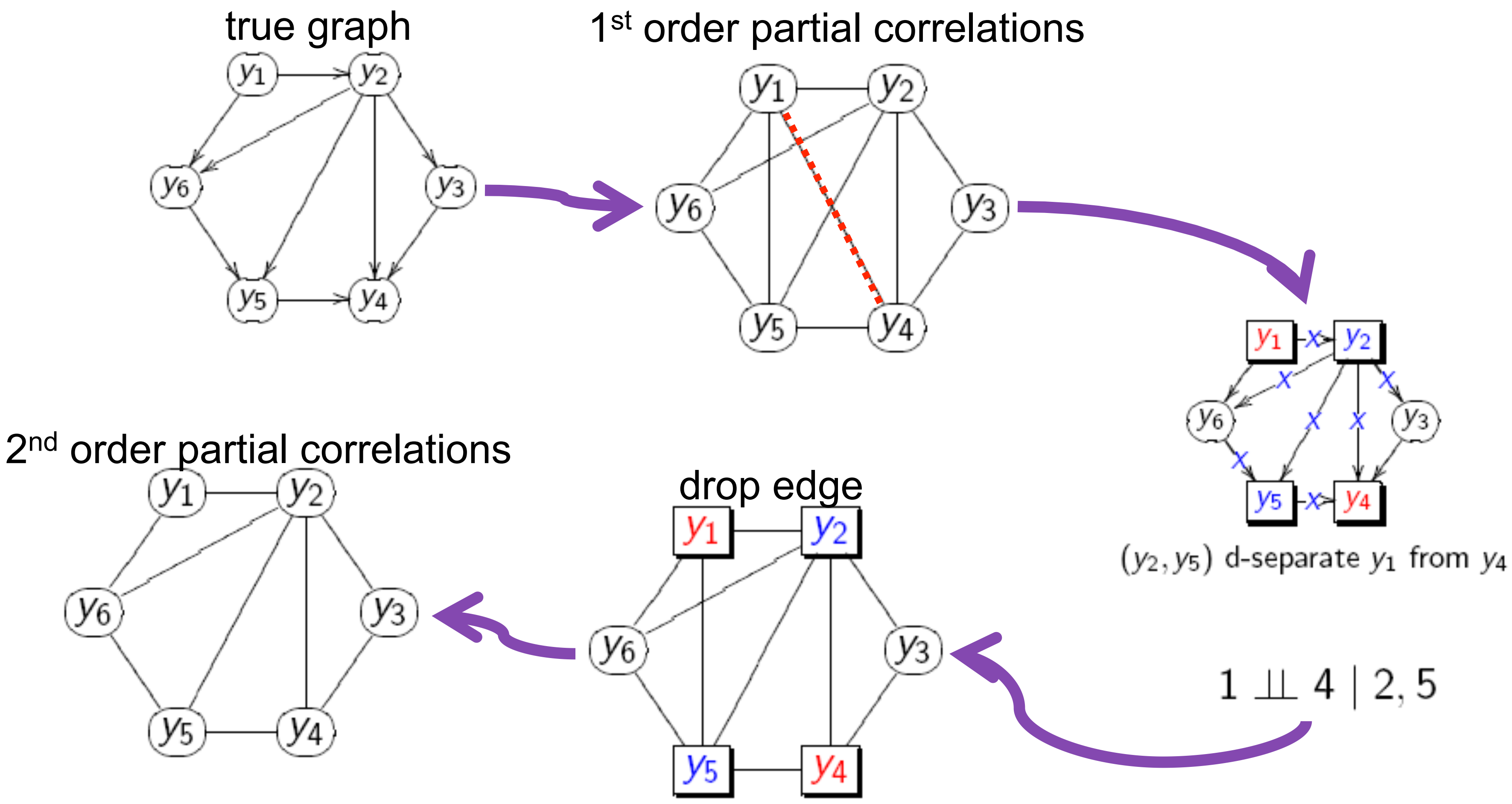
y_2 d-separates y_1 from y_3

$1 \perp\!\!\!\perp 3 \mid 2$

drop edge

1st order partial correlations

partial correlation (PC) skeleton



true graph

1st order partial correlations

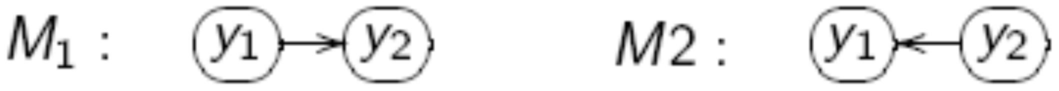
(y_2, y_5) d-separate y_1 from y_4

$1 \perp\!\!\!\perp 4 \mid 2, 5$

drop edge

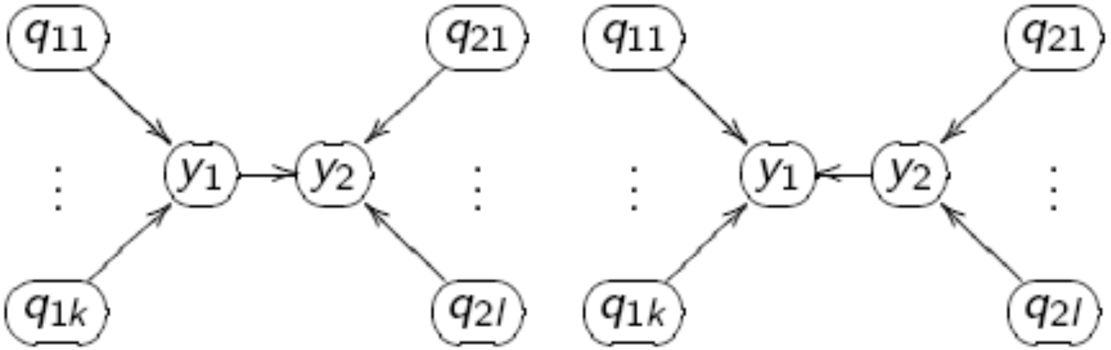
2nd order partial correlations

edge direction: which is causal?



the above models are likelihood equivalent,

$$f(y_1)f(y_2 \mid y_1) = f(y_1, y_2) = f(y_2)f(y_1 \mid y_2)$$

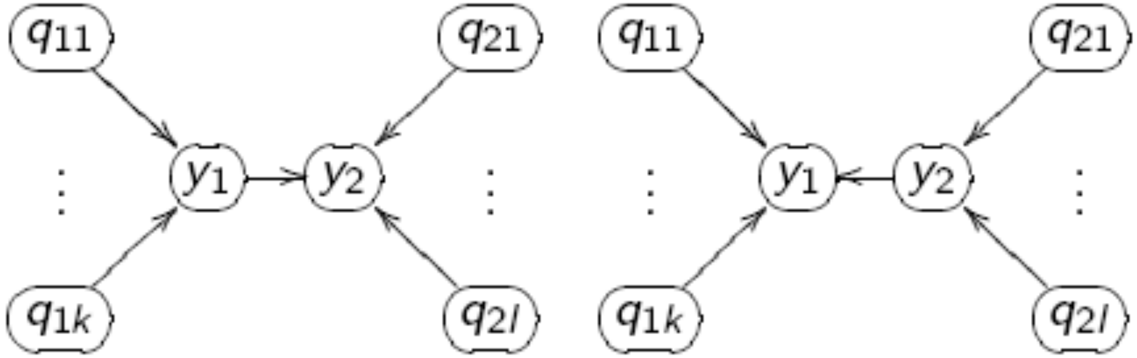


not likelihood equivalent due to QTL

$$f(\mathbf{q}_1)f(y_1 \mid \mathbf{q}_1)f(y_2 \mid y_1, \mathbf{q}_2)f(\mathbf{q}_2)$$
$$\neq$$
$$f(\mathbf{q}_2)f(y_2 \mid \mathbf{q}_2)f(y_1 \mid y_2, \mathbf{q}_1)f(\mathbf{q}_1)$$

test edge direction using LOD score

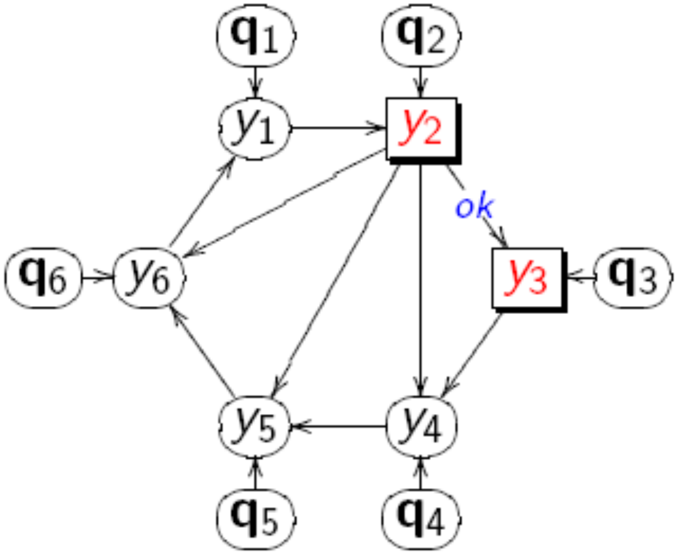
$$LOD = \log_{10}\left\{\frac{\prod_{i=1}^{n} f(y_{1i} \mid \mathbf{q}_{1i}) f(y_{2i} \mid y_{1i},\, \mathbf{q}_{2i})}{\prod_{i=1}^{n} f(y_{2i} \mid \mathbf{q}_{2i}) f(y_{1i} \mid y_{2i},\, \mathbf{q}_{1i})}\right\}$$

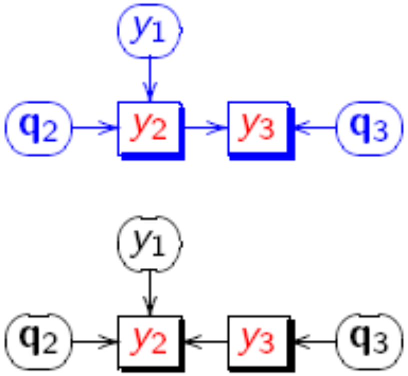


not likelihood equivalent because

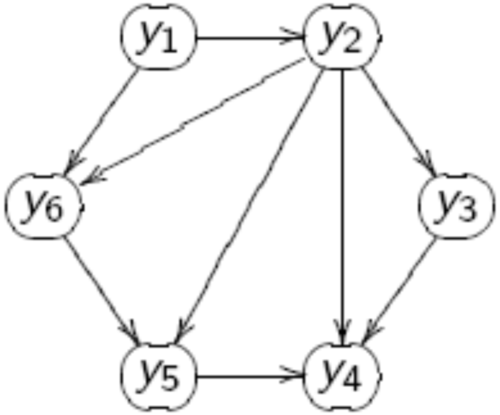
$$f(\mathbf{q}_1)f(y_1 \mid \mathbf{q}_1)f(y_2 \mid y_1, \mathbf{q}_2)f(\mathbf{q}_2)$$
$$\neq$$
$$f(\mathbf{q}_2)f(y_2 \mid \mathbf{q}_2)f(y_1 \mid y_2, \mathbf{q}_1)f(\mathbf{q}_1)$$

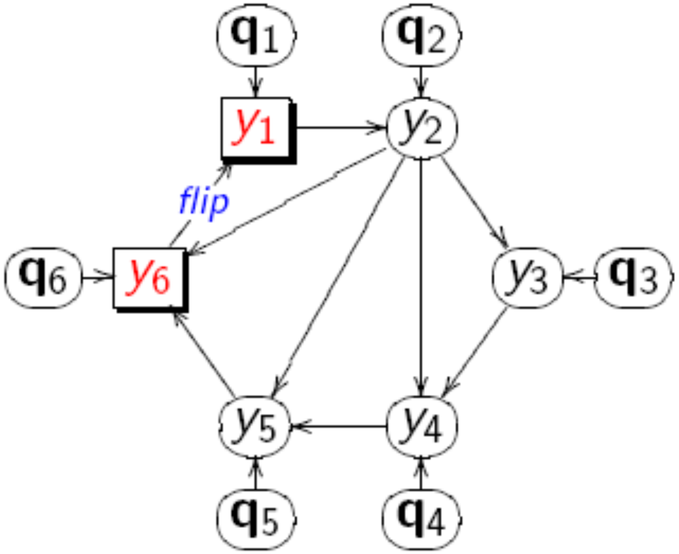
reverse edges using QTLs

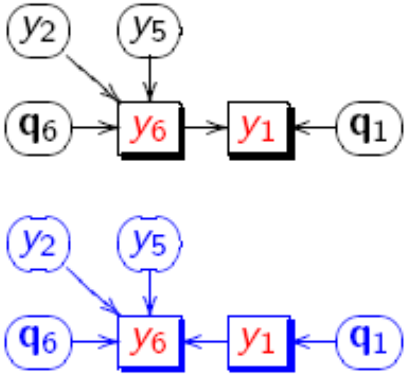


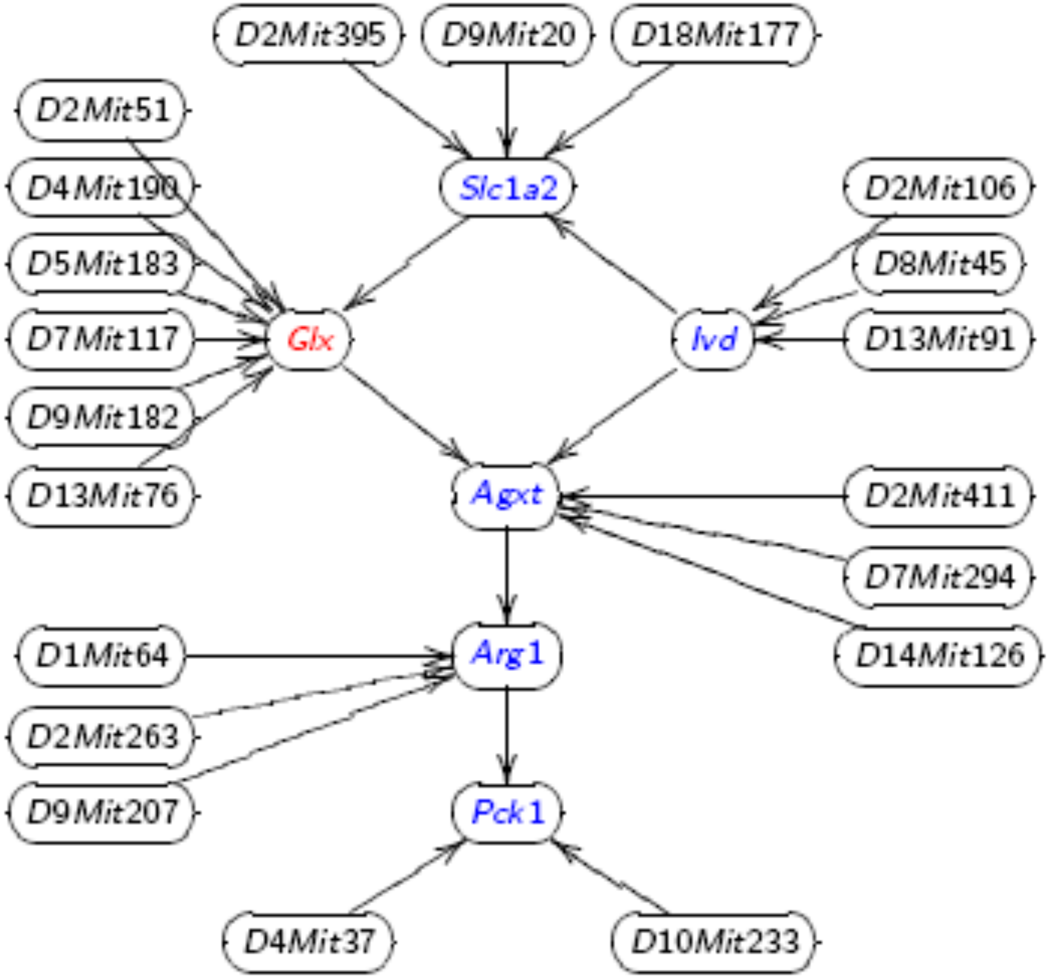


true graph









- We constructed a network from metabolites and transcripts involved in liver metabolism.
- We validated this network with *in vitro experiments* (Ferrara et al 2008). Four out of six predictions were confirmed.

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

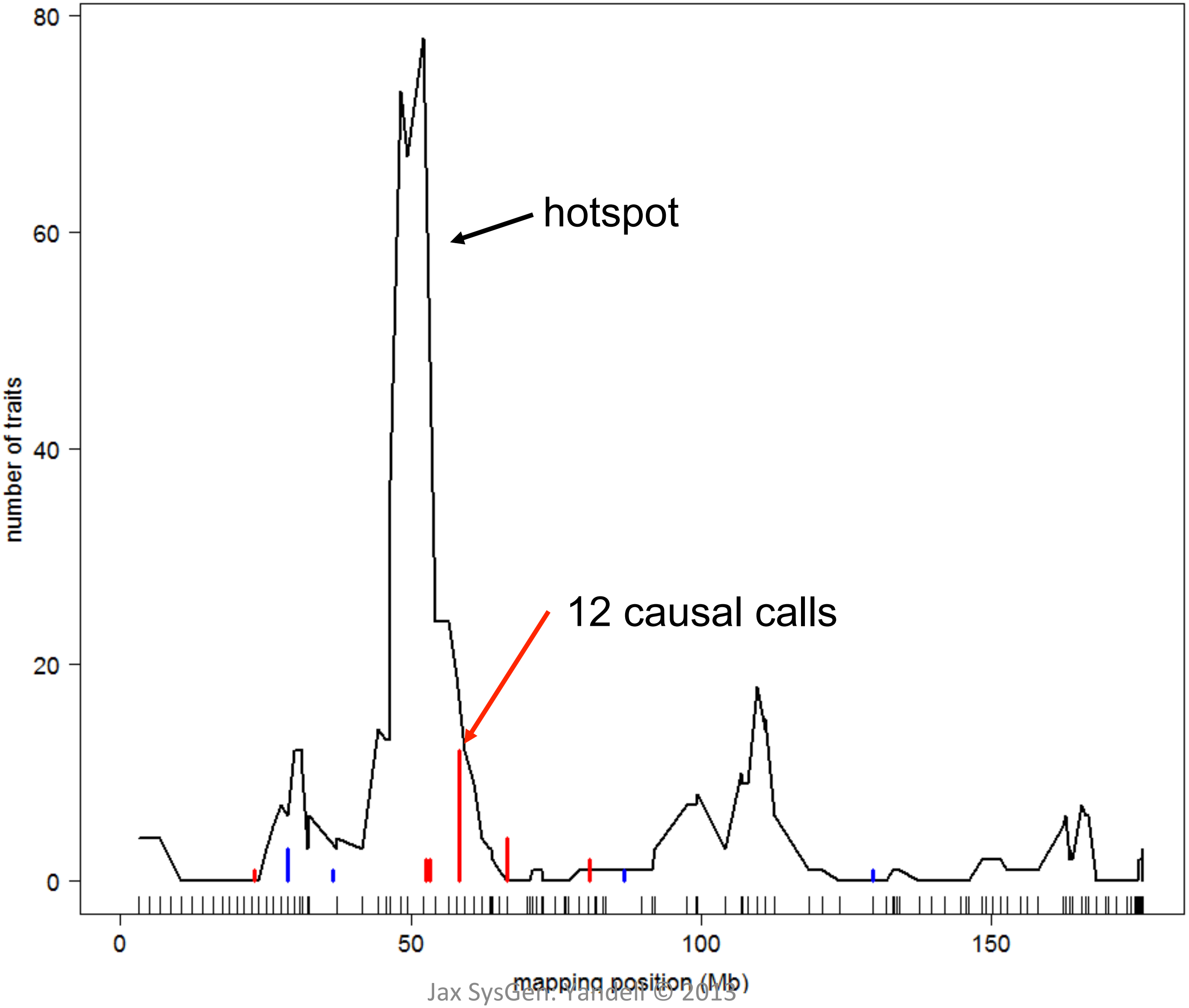
missing data method: MCMC

- known phenotypes *Y*, genotypes *Q*
- unknown graph *G*
- want to study Pr(Y | G, Q)
- break down in terms of individual edges
  - Pr($Y|G,Q$) = sum of Pr(Y_i | pa(Y_i), Q)
- sample new values for individual edges
  - given current value of all other edges
- repeat many times and average results

MCMC steps for QTLnet

- propose new causal network *G*
  - with simple changes to current network:
  - change edge direction
  - add or drop edge
- find any new genetic architectures *Q*
  - update phenotypes when parents pa(y) change in new *G*
- compute likelihood for new network and QTL
  - Pr(Y | G, Q)
- accept or reject new network and QTL
  - usual Metropolis-Hastings idea

BxH ApoE-/- chr 2: causal architecture



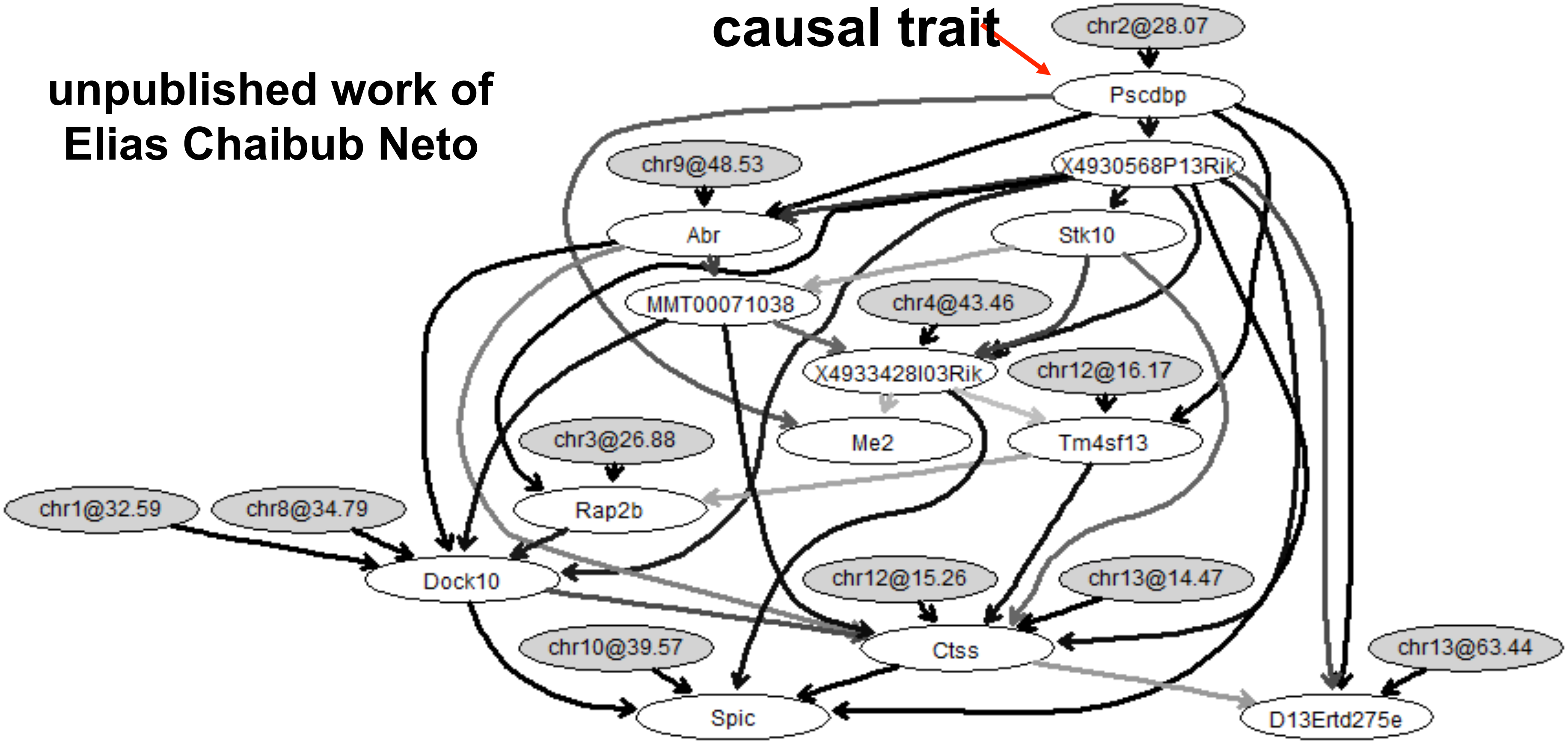
hotspot

12 causal calls

BxH ApoE-/- causal network for transcription factor Pscdbp

unpublished work of Elias Chaibub Neto

causal trait

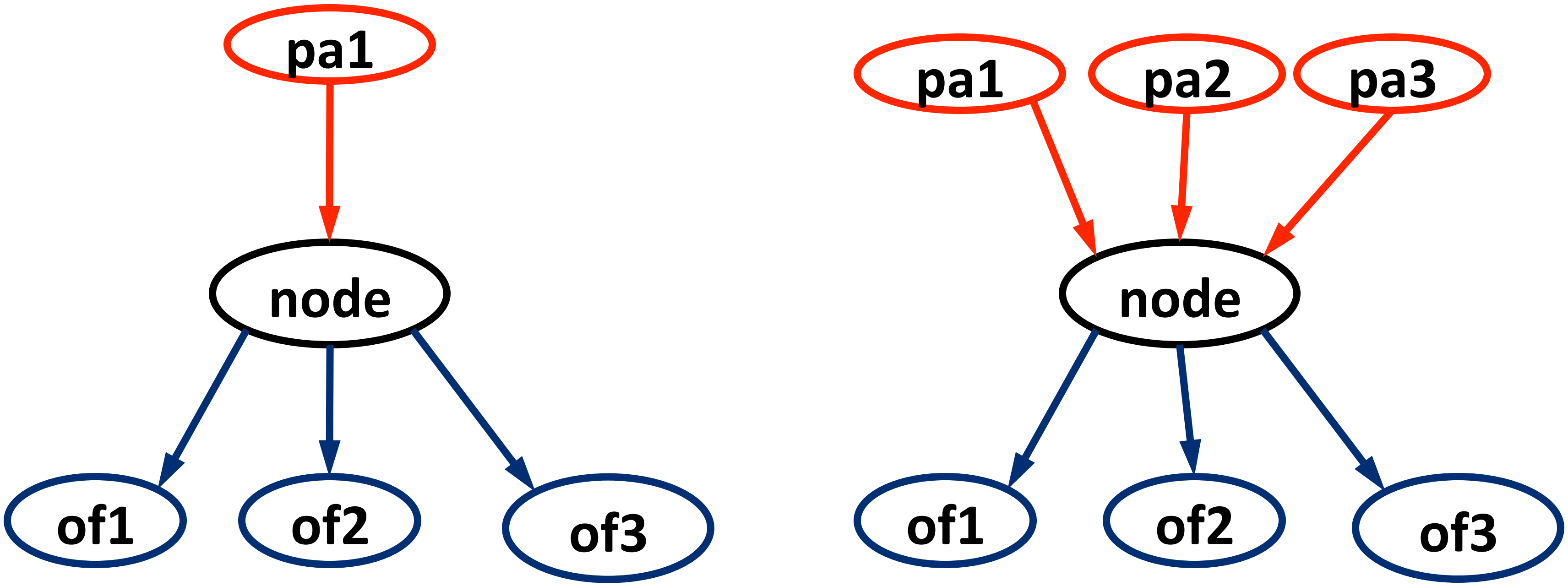


••• scaling up •••

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



how many node parents?

- how many edges per node? (fan-in)
  - – few parents directly affect one node
  - – many offspring affected by one node

BIC computations by maximum number of parents

#	3	4	5	6	all
10	1,300	2,560	3,820	4,660	5,120
20	23,200	100,720	333,280	875,920	10.5M
30	122,700	835,230	4.40M	18.6M	16.1B
40	396,800	3.69M	26.7M	157M	22.0T
50	982,500	11.6M	107M	806M	28.1Q

BIC computation

- each trait (node) has a linear model
  - Y ~ QTL + pa(Y) + other covariates
- BIC = LOD – penalty
  - BIC balances data fit to model complexity
  - penalty increases with number of parents
- limit complexity by allowing only 3-4 parents

parallel phases for larger projects

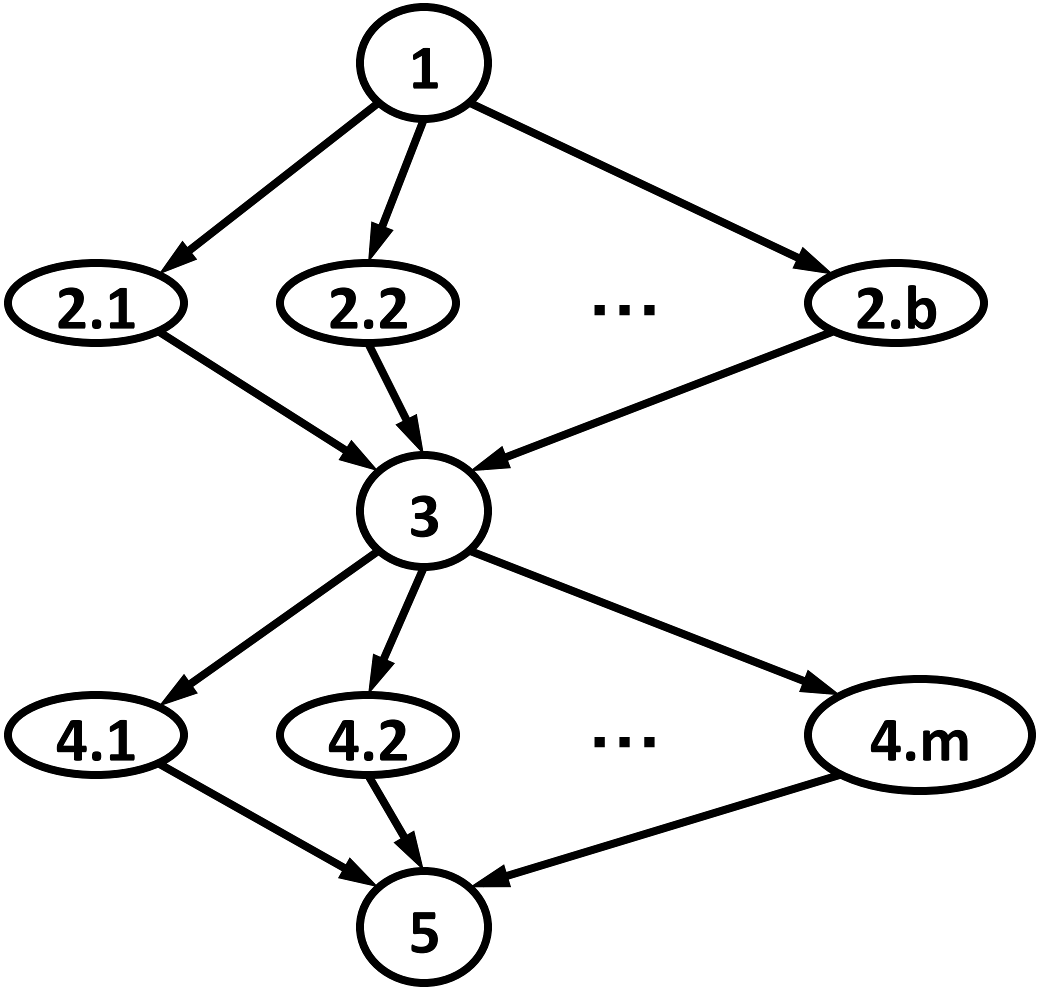
Phase 1: identify parents

Phase 2: compute BICs

Phase 3: store BICs

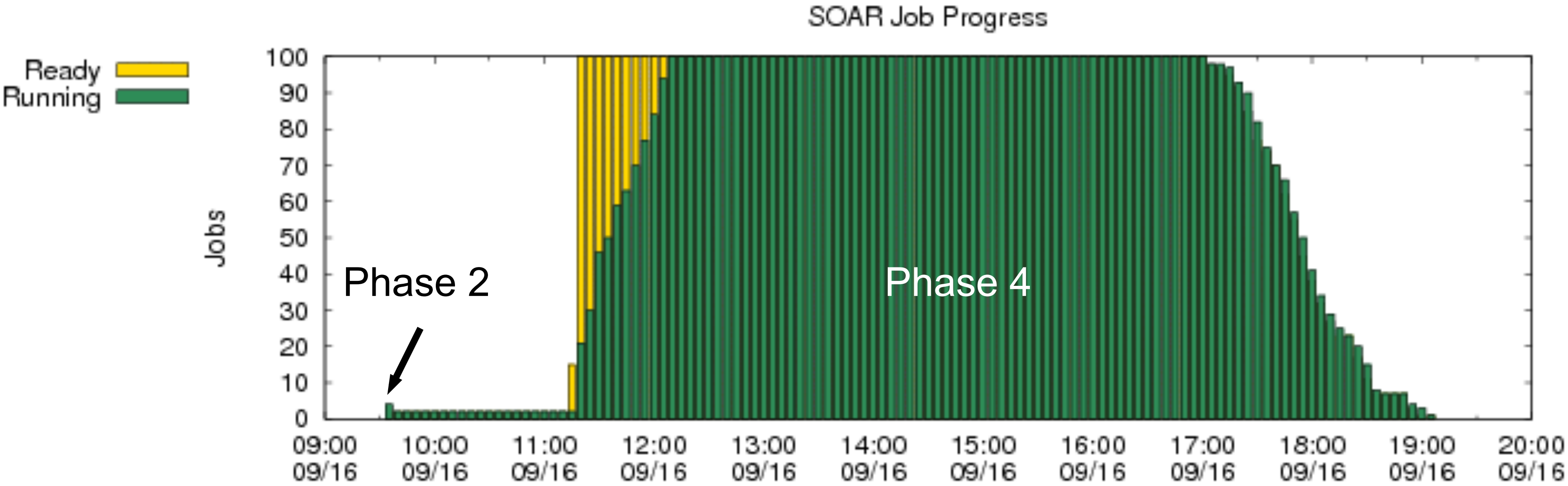
Phase 4: run Markov chains

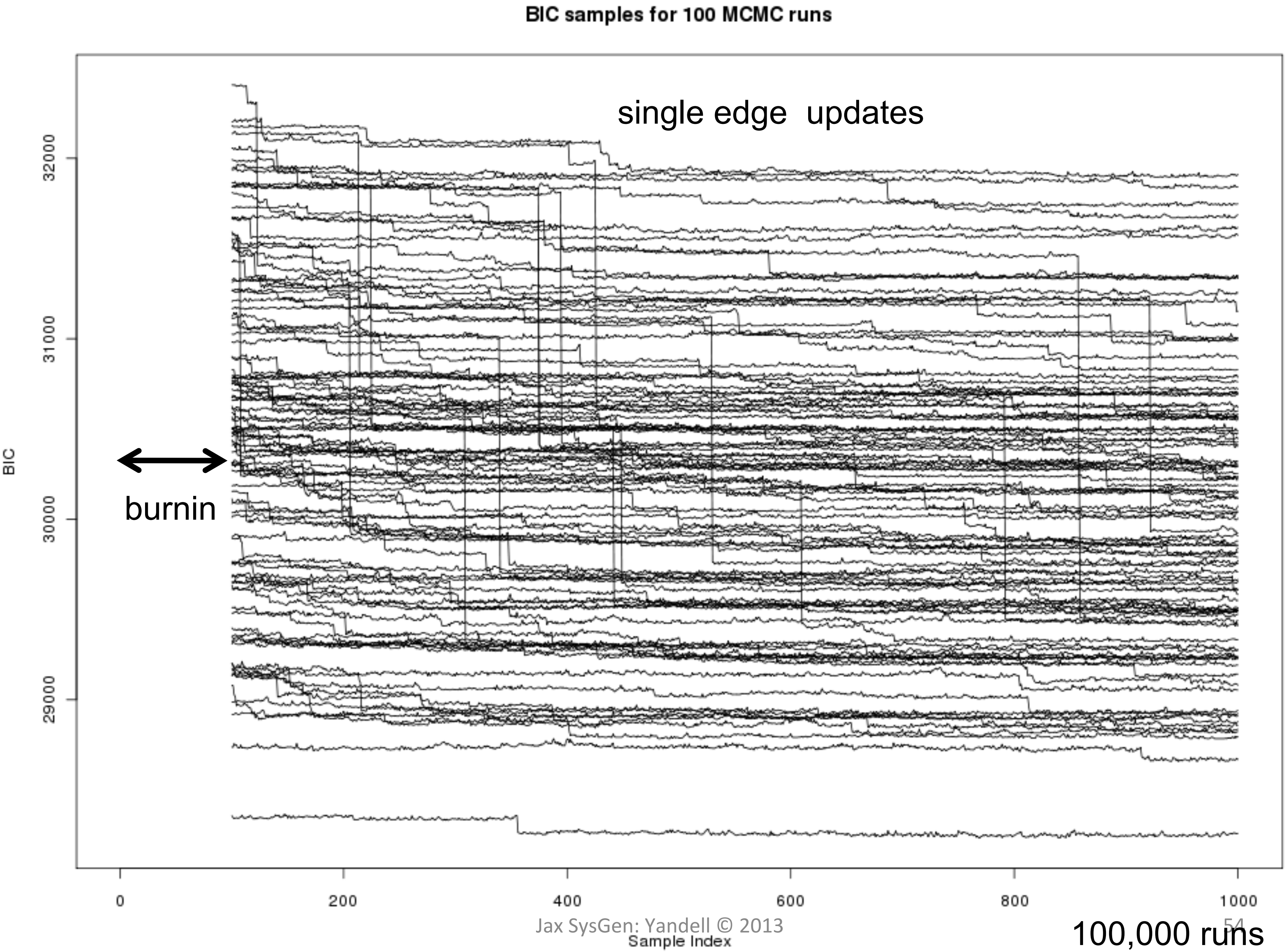
Phase 5: combine results



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





single edge updates

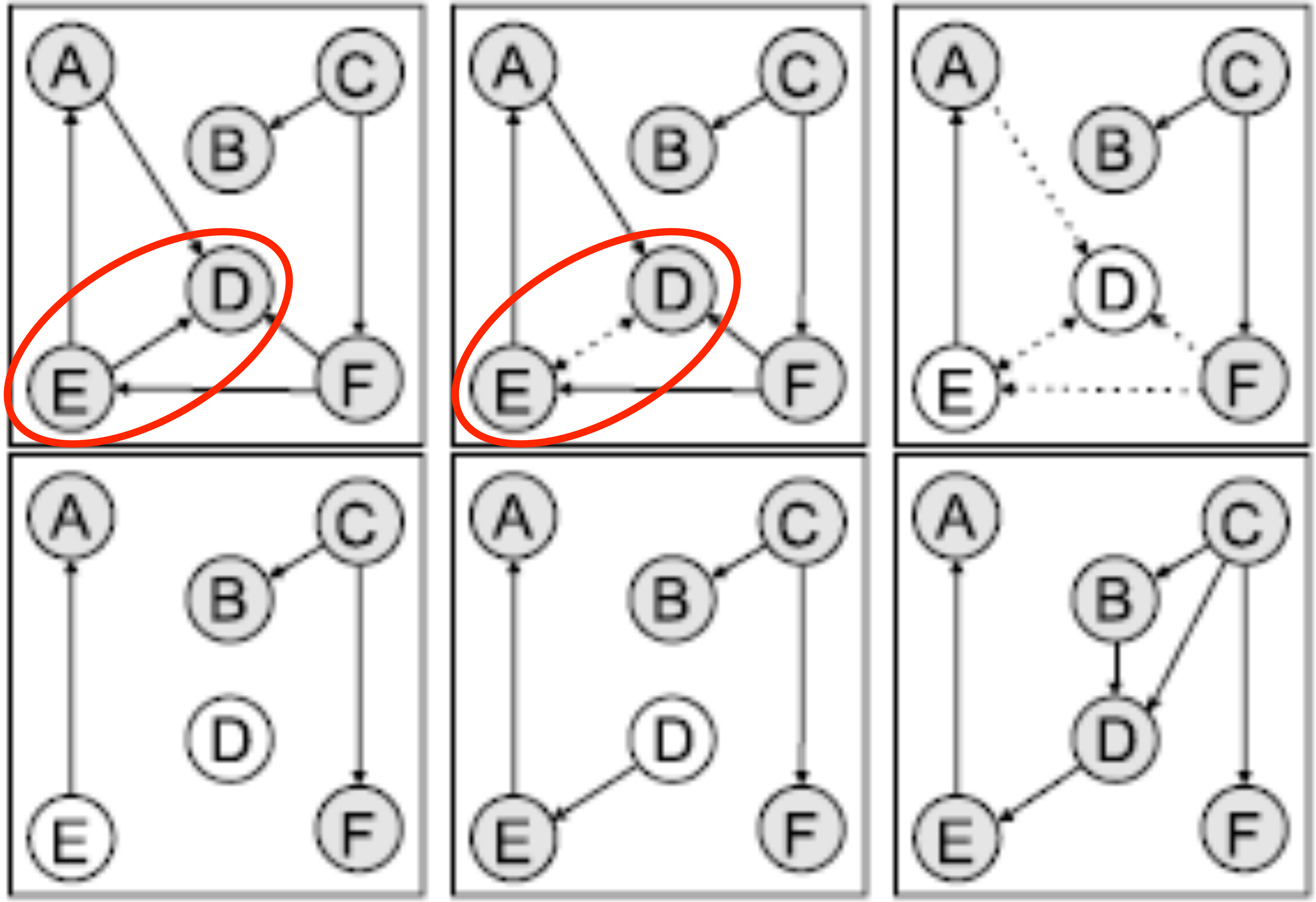
burnin

100,000 runs

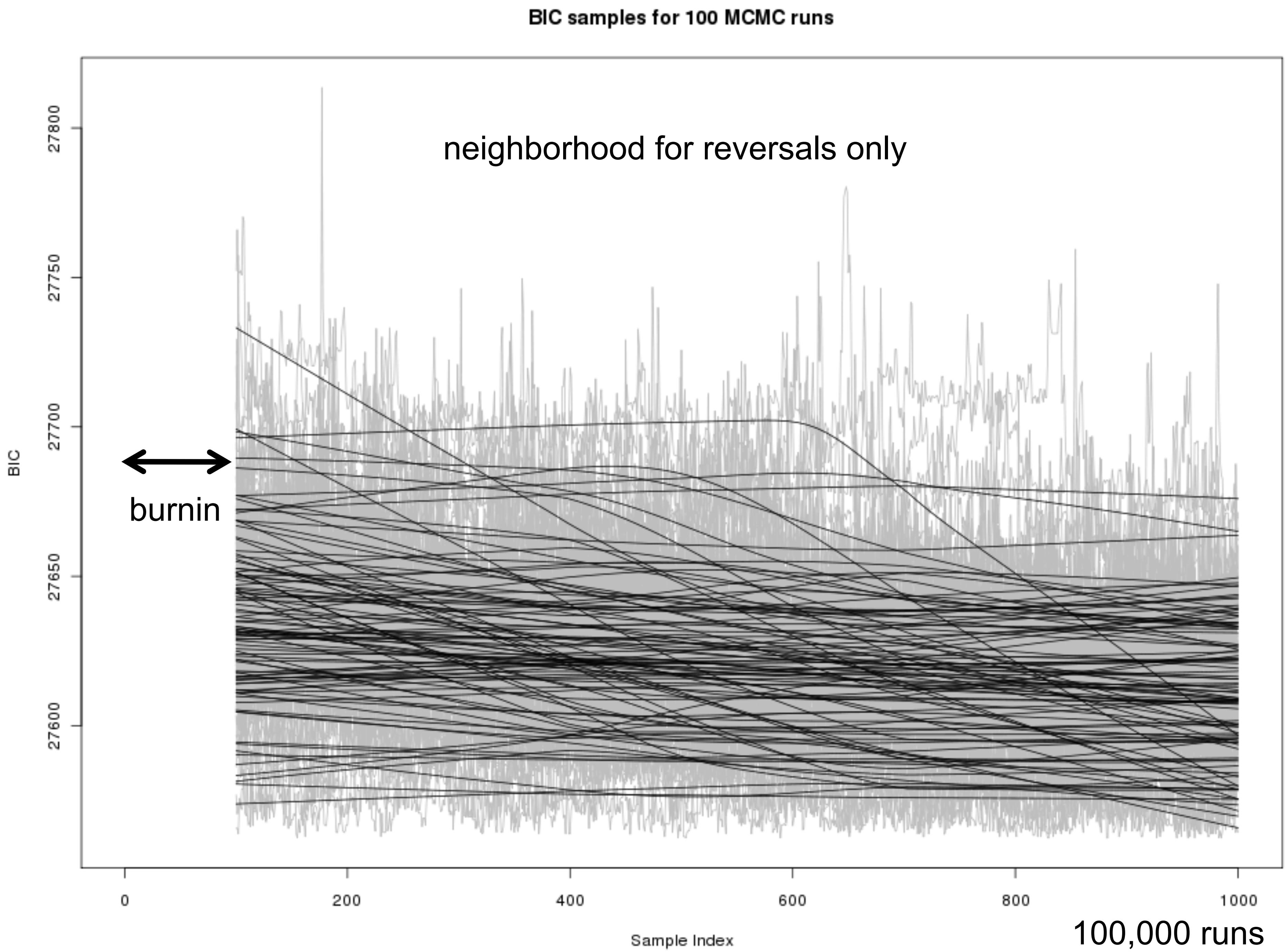
neighborhood edge reversal

select edge
drop edge
identify parents

orphan nodes
reverse edge
find new parents



Grzegorczyk M. and Husmeier D. (2008) *Machine Learning* 71 (2-3), 265-305.



neighborhood for reversals only

burnin

100,000 runs

limits of causal inference

unfaithful: false positive edges

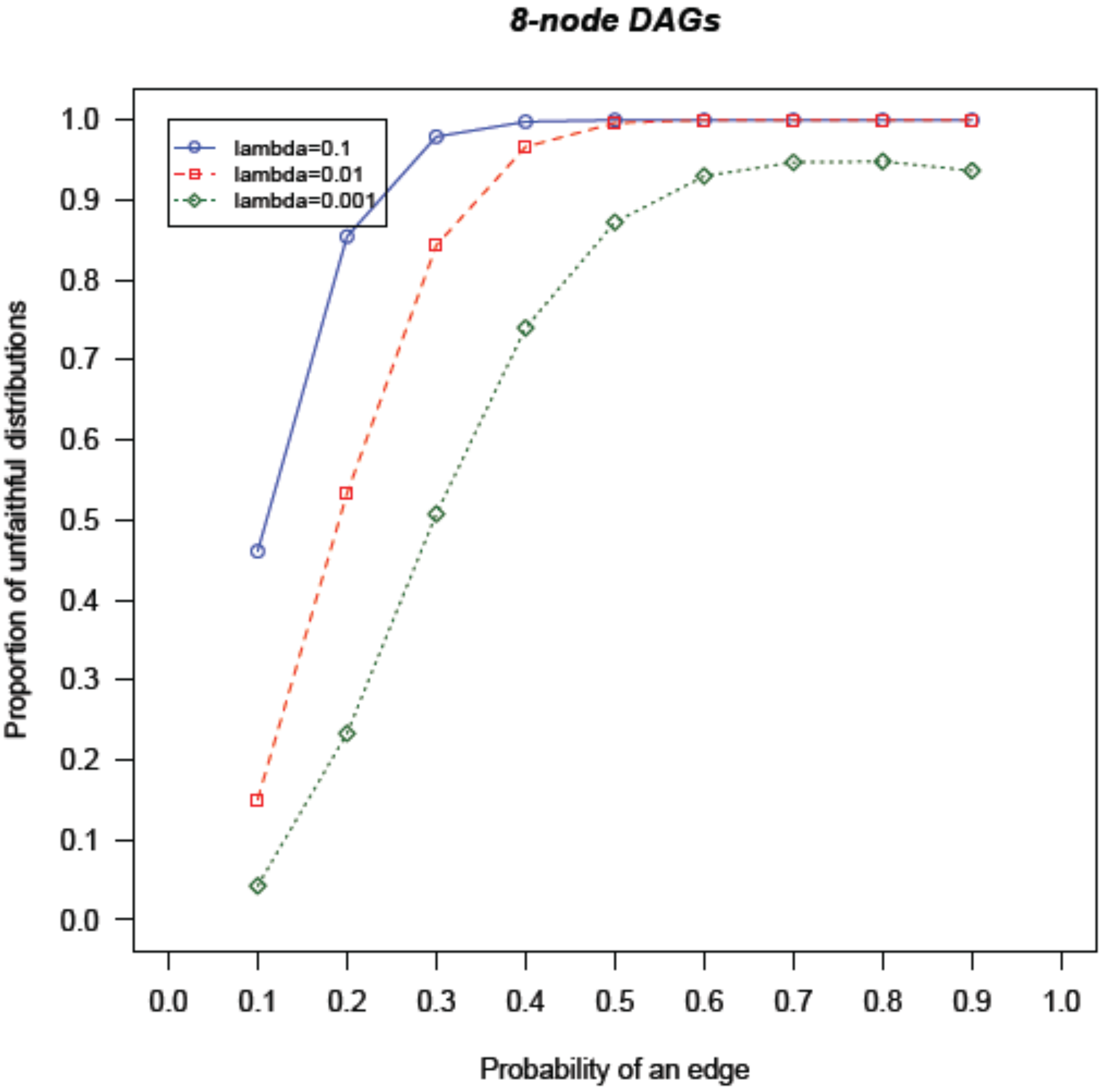
λ =min|cor(Y_i,Y_j)|

λ=c•sqrt(dp/n)

d=max degree

p=# nodes

n=sample size



Uhler, Raskutti, Buhlmann, Yu (2012 arxiv)

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* | *B, W*) = exp(– *W* * |*B*–*G*|) / 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
  - – *P*(*W*|*B*) = ϕ exp(- ϕ *W*) exponential
- 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

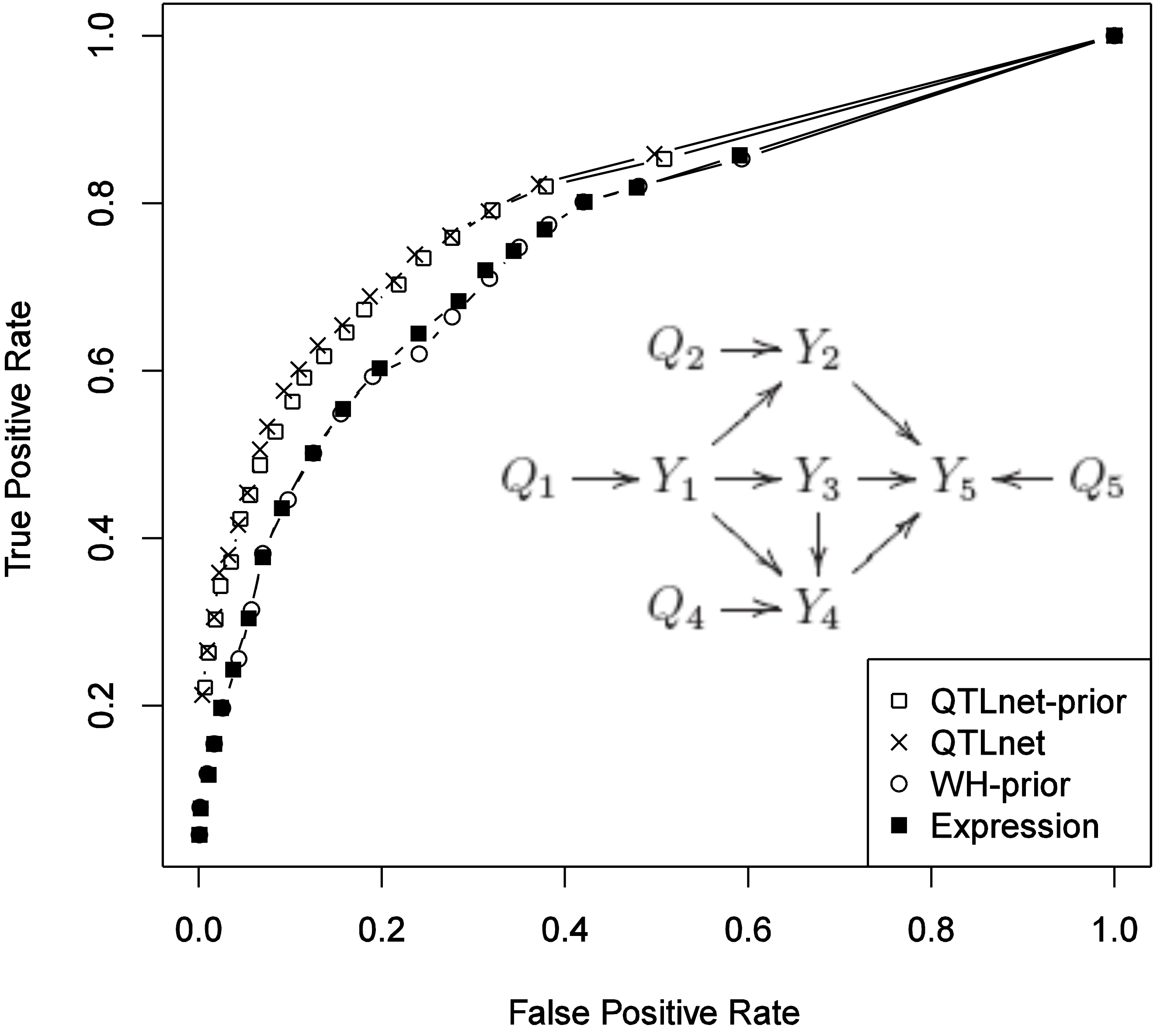
$$\Pr(G, W \mid Y, Q, B) = \Pr(Y|G,Q)\Pr(G|B,W)\Pr(W|B)$$

  - $\Pr(Y|G,Q)$ 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
- related ref: Kim et al. Przytycka (2010) *RECOMB*

ROC curve simulation

open = QTLnet

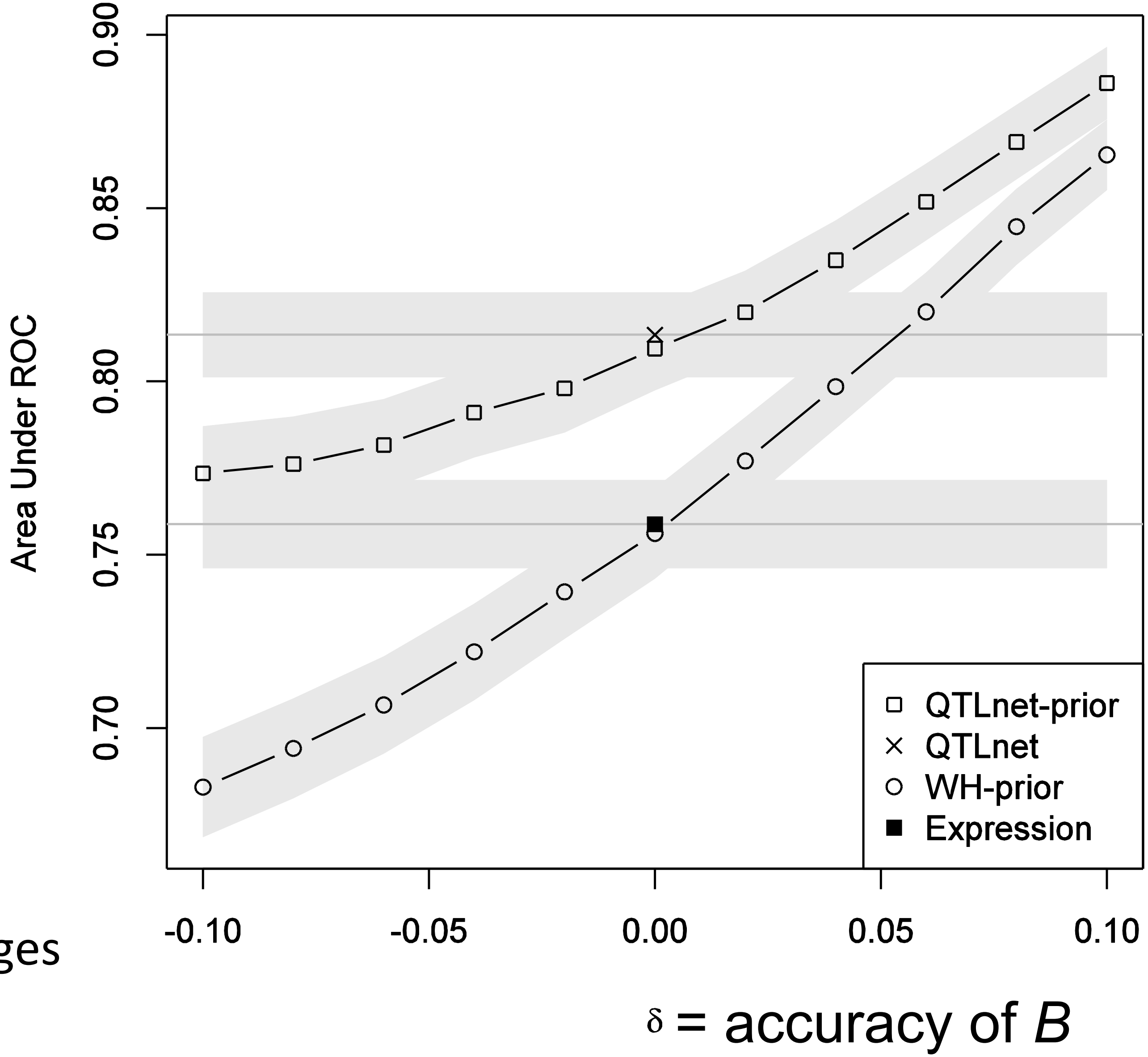
closed = phenotypes only



integrated ROC curve

2x2: genetics pathways

probability classifier
ranks true > false edges



QTL software on CRAN

- R/qtlhot: hotspots & causal architecture
  - map hotspots, permutation tests
  - causal model selection tests
- R/qtlnet: QTL-driven phenotype networks
  - infer QTLs and directed graphs
  - coming: prior biological information
- R/qtlbim: Bayesian Interval Mapping for QTL
  - multiple QTL inference, graphical diagnostics
  - see earlier Jax talks for details

many thanks!

- NIH/NIGMS with Karl Broman, Nengjun Yi
- NIH/NIDDK with Alan Attie, Mark Keller
- Other collaborators:
  - Mark Keller (Attie Lab Scientist)
  - Chris Plaisier (Institute for Systems Biology, Seattle)
  - Elias Chaibub Neto (Sage Bionetworks, Seattle)
  - Jee Young Moon (grad student)
  - Xinwei Deng (VA Tech Asst Prof)
  - and many more!