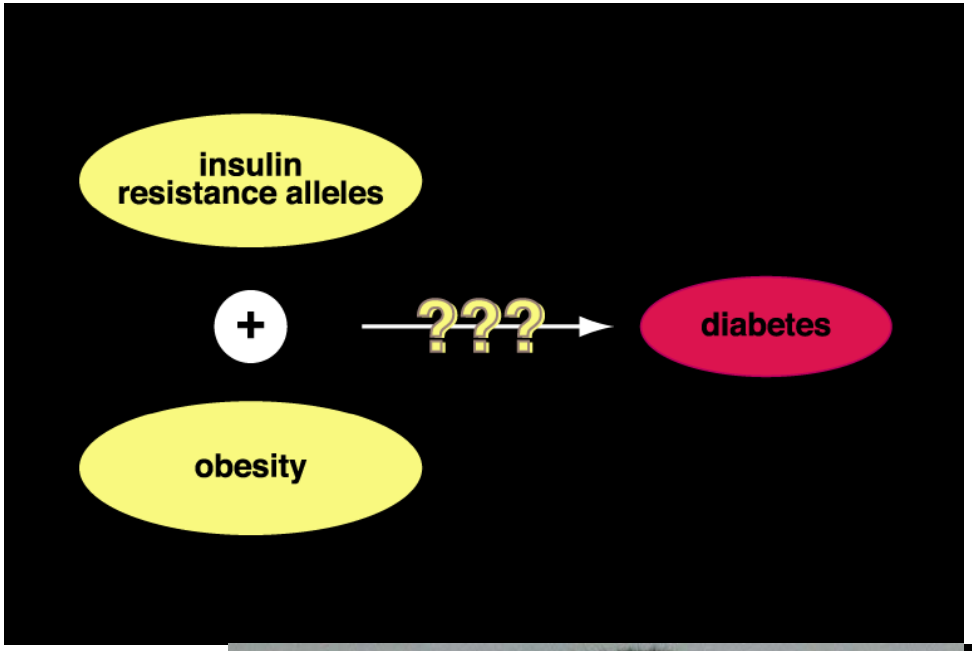
Bayesian causal phenotype network incorporating genetic variation and biological knowledge

Brian S Yandell, Jee Young Moon
University of Wisconsin-Madison
Elias Chaibub Neto, Sage Bionetworks
Xinwei Deng, VA Tech

http://www.stat.wisc.edu/~yandell/talk/2012.oslo.pdf



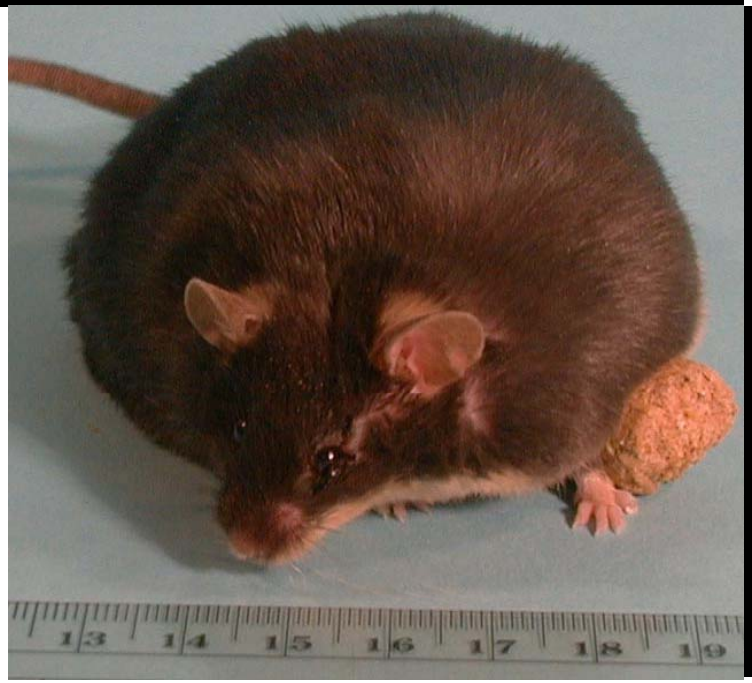
BTBR mouse is insulin resistant

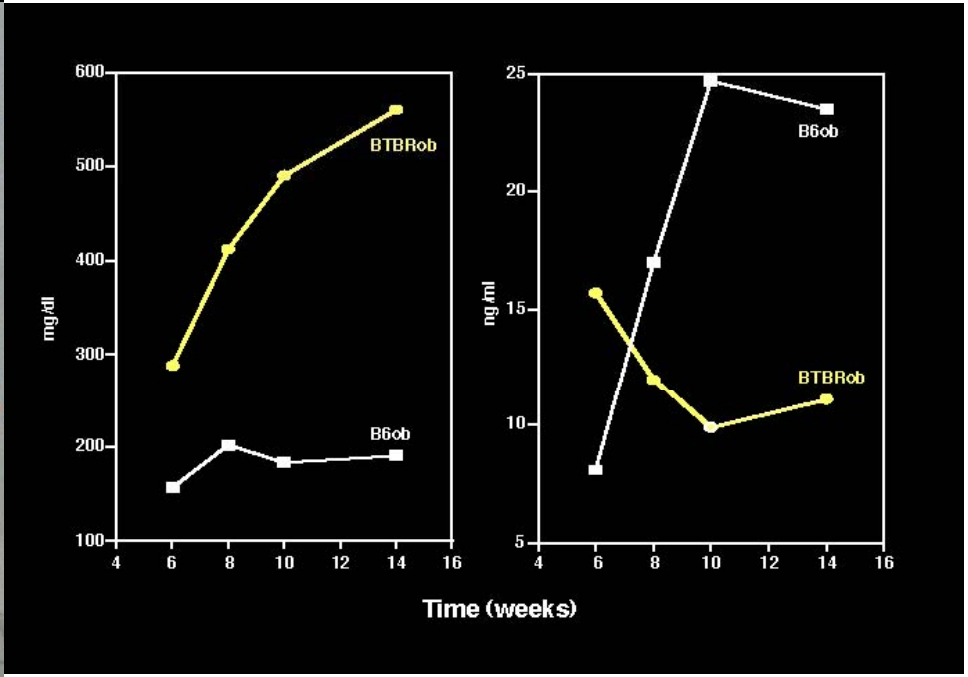
B6 is not

make both obese…

Alan Attie
Biochemistry

glucose insulin

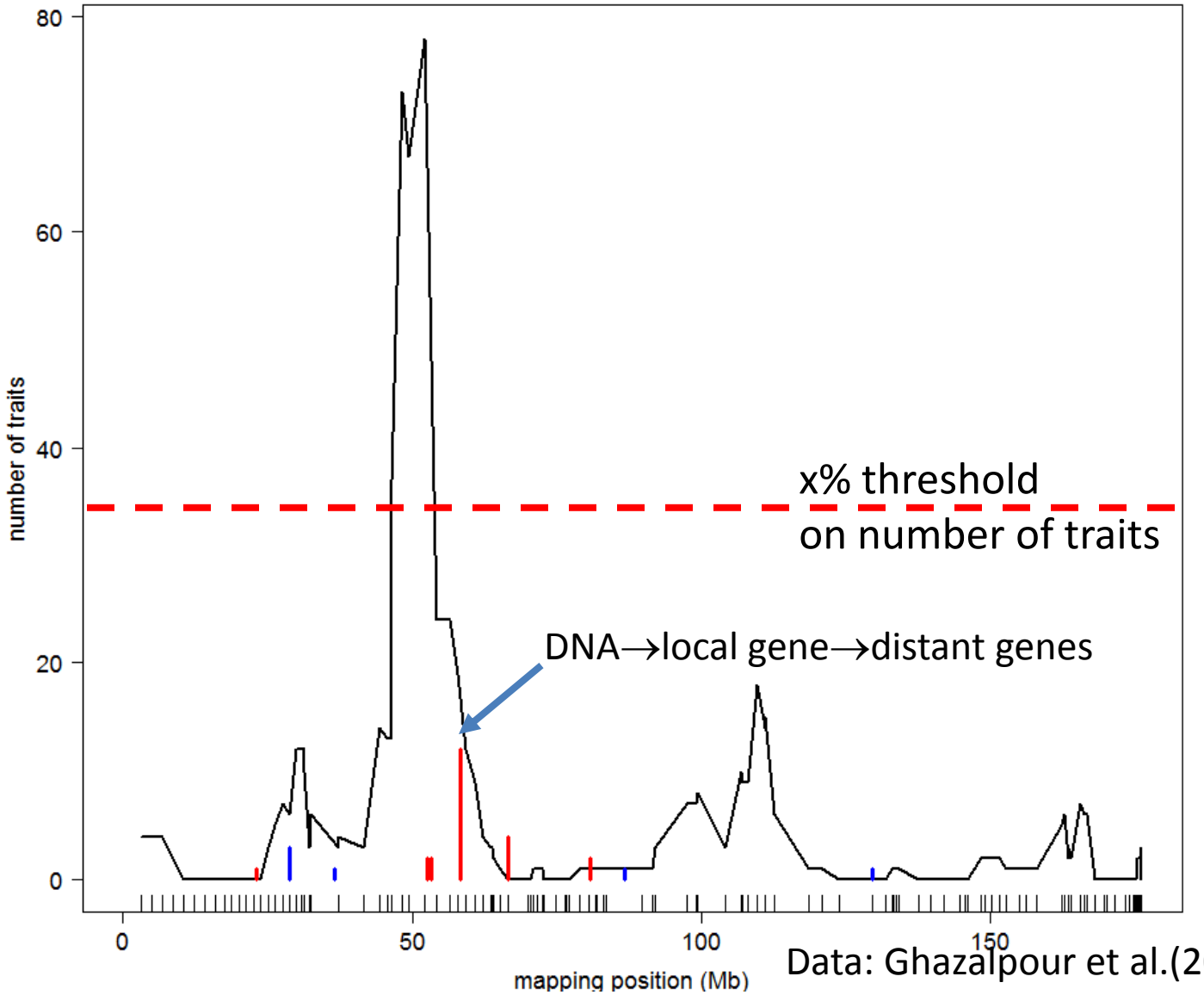




bigger picture

- how do DNA, RNA, proteins, metabolites regulate each other?
- regulatory networks from microarray expression data
  - time series measurements or transcriptional perturbations
  - segregating population: **genotype as driving perturbations**
- goal: discover causal regulatory relationships among phenotypes
- use knowledge of regulatory relationships from databases
  - how can this improve causal network reconstruction?

BxH ApoE-/- chr 2: hotspot

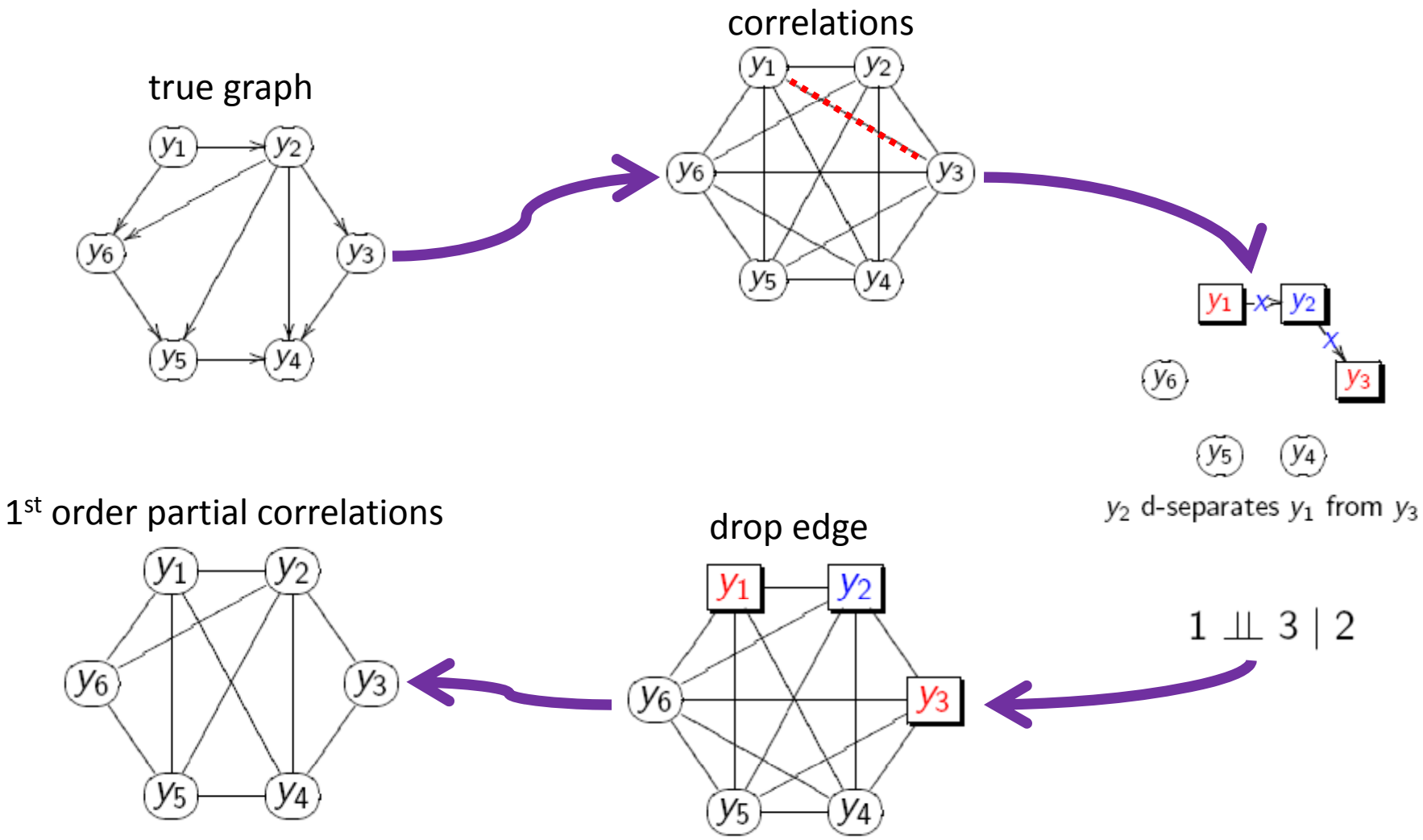


Data: Ghazalpour et al.(2006) *PLoS Genetics*

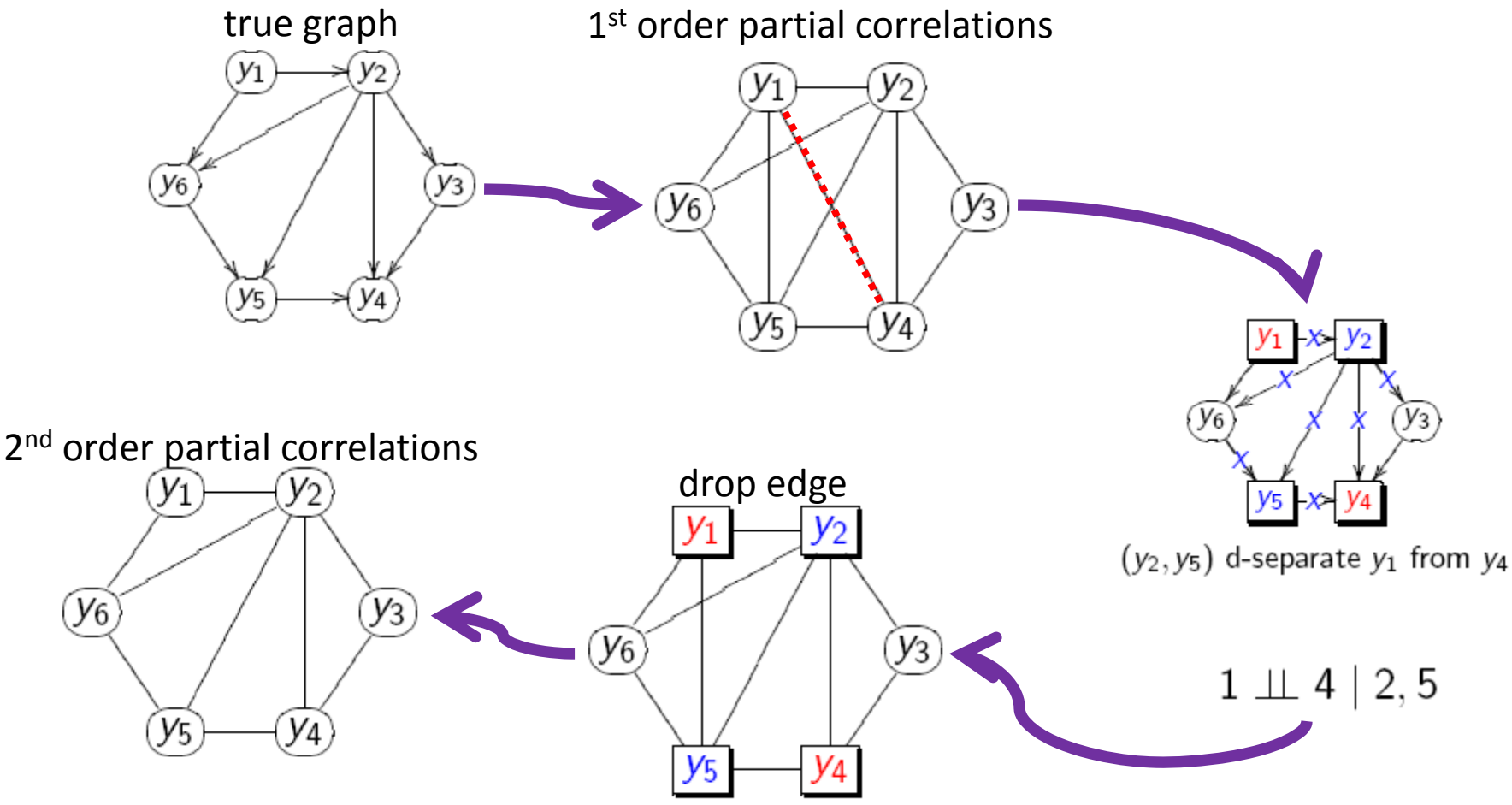
QTL-driven directed graphs

- 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



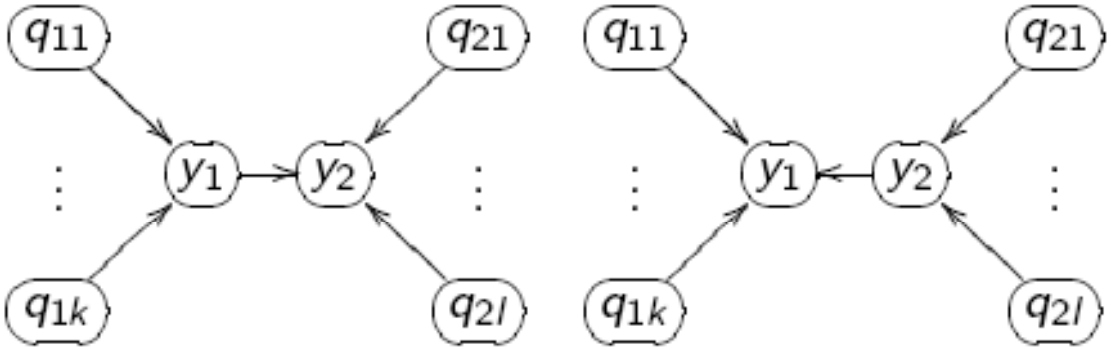
partial correlation (PC) skeleton



edge direction: which is causal?

the above models are likelihood equivalent,

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

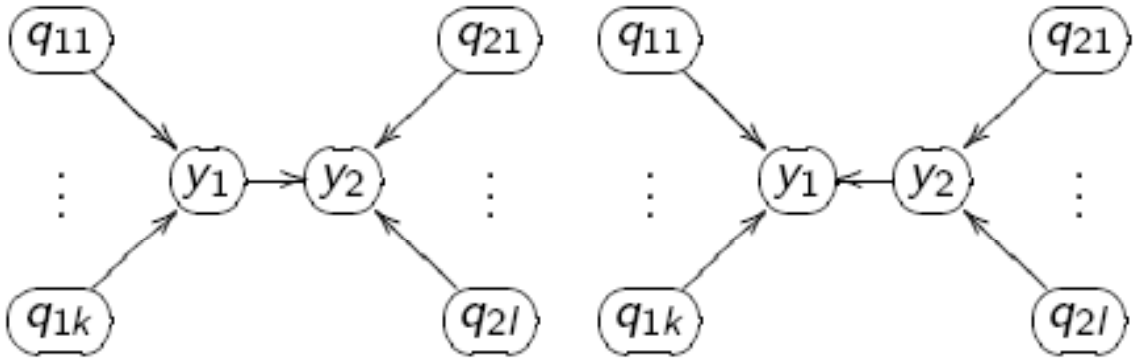


not likelihood equivalent due to QTL

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

test edge direction using LOD score

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

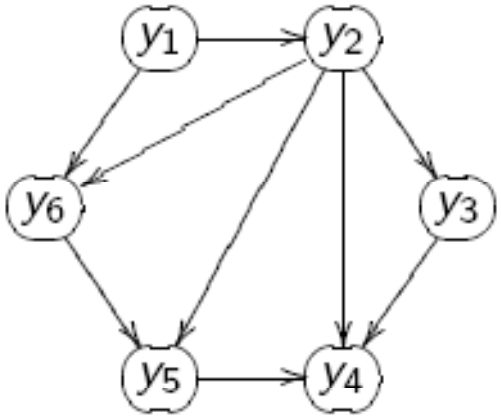


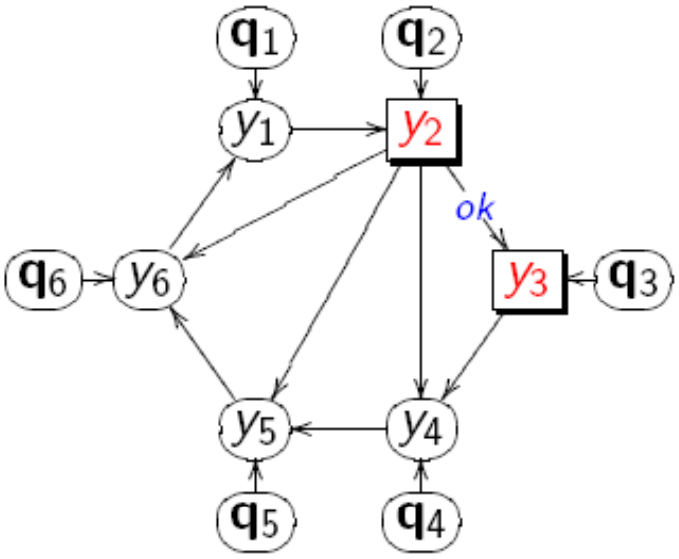
not likelihood equivalent because

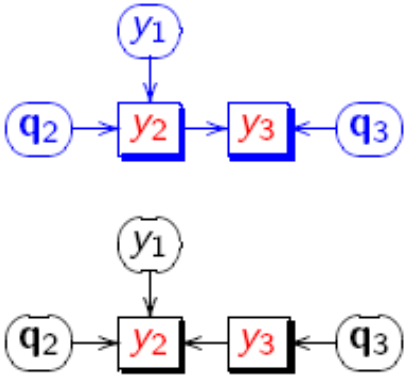
$$f(\mathbf{q}_1) f(y_1 \mid \mathbf{q}_1) f(y_2 \mid y_1, \mathbf{q}_2) f(\mathbf{q}_2)$$
$$\neq$$
$$f(\mathbf{q}_2) f(y_2 \mid \mathbf{q}_2) f(y_1 \mid y_2, \mathbf{q}_1) f(\mathbf{q}_1)$$

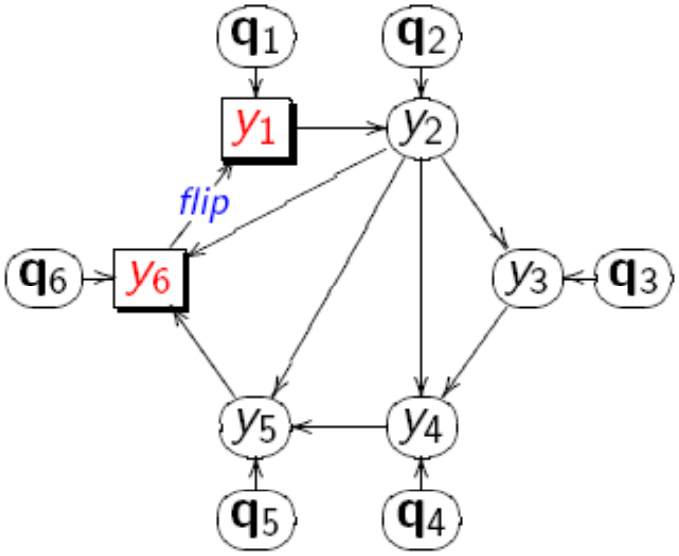
reverse edges using QTLs

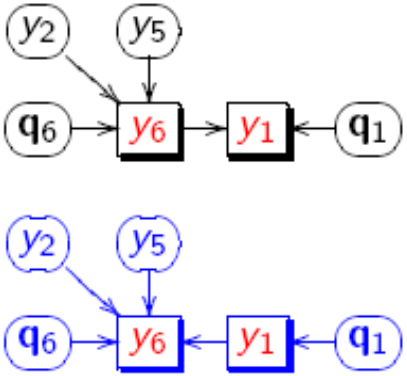
true graph











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

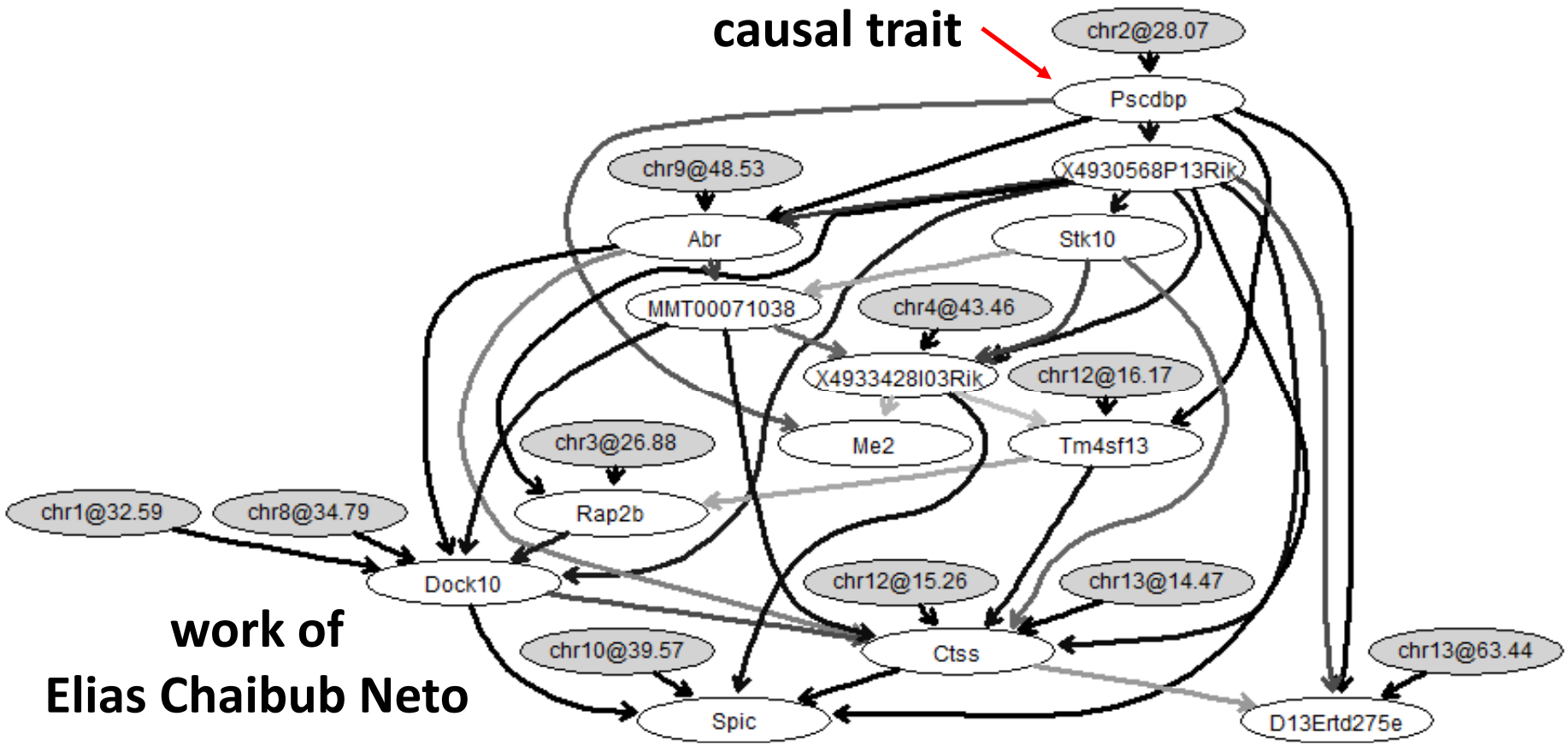
missing data method: MCMC

- known phenotypes Y, genotypes Q
- unknown graph G
- want to study $\Pr(Y \mid G, Q)$
- break down in terms of individual edges
  - $\Pr(Y|G,Q)$ = sum of $\Pr(Y_i \mid \text{pa}(Y_i), Q)$
- 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-/- causal network for transcription factor Pscdbp

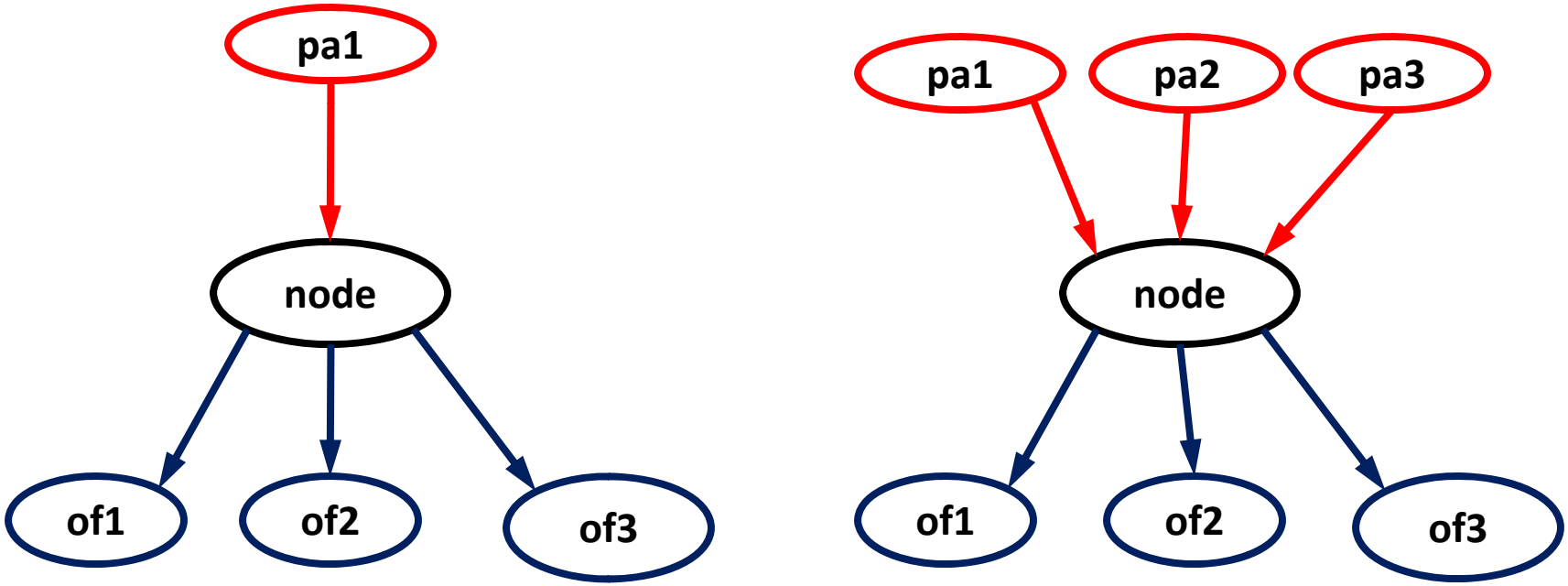


Data: Ghazalpour et al.(2006) *PLoS Genetics*

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

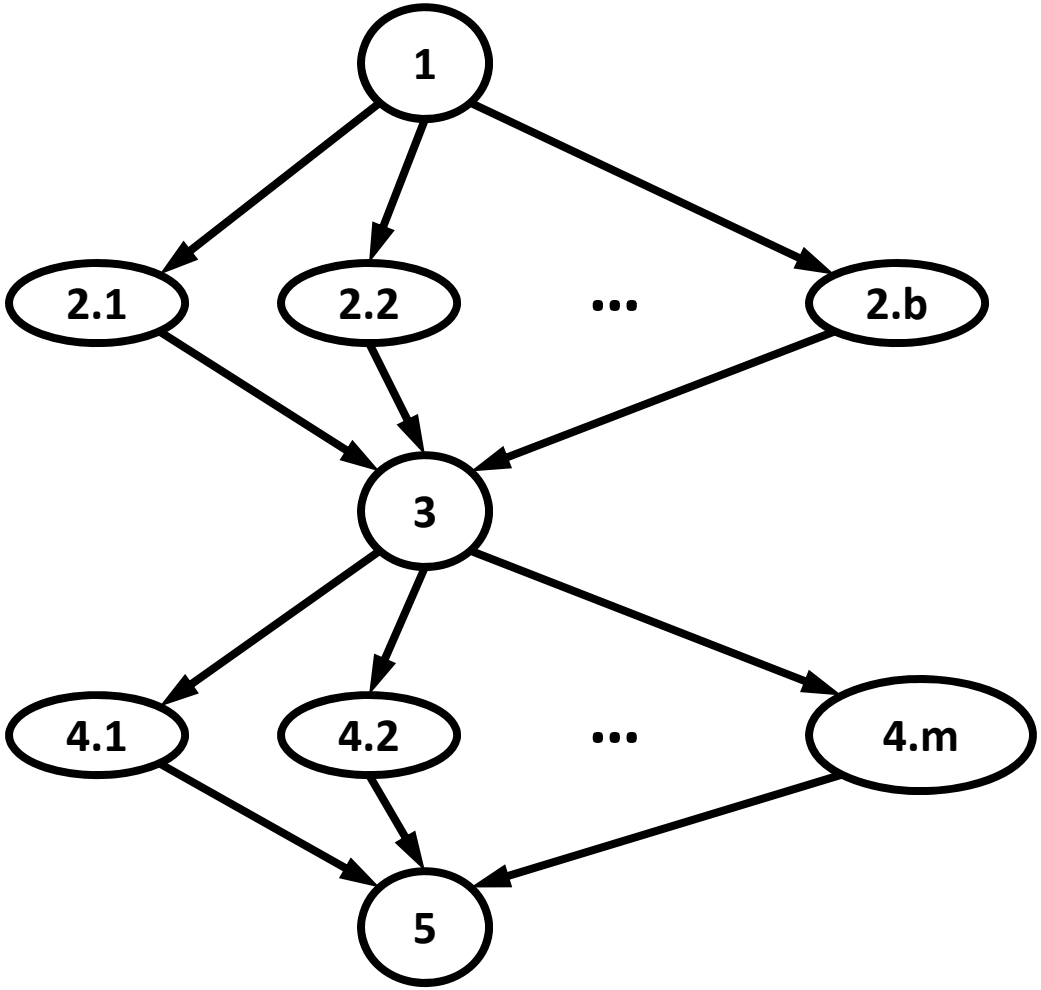
BIC = LOD – penalty

all possible parents to all nodes

Phase 3: store BICs

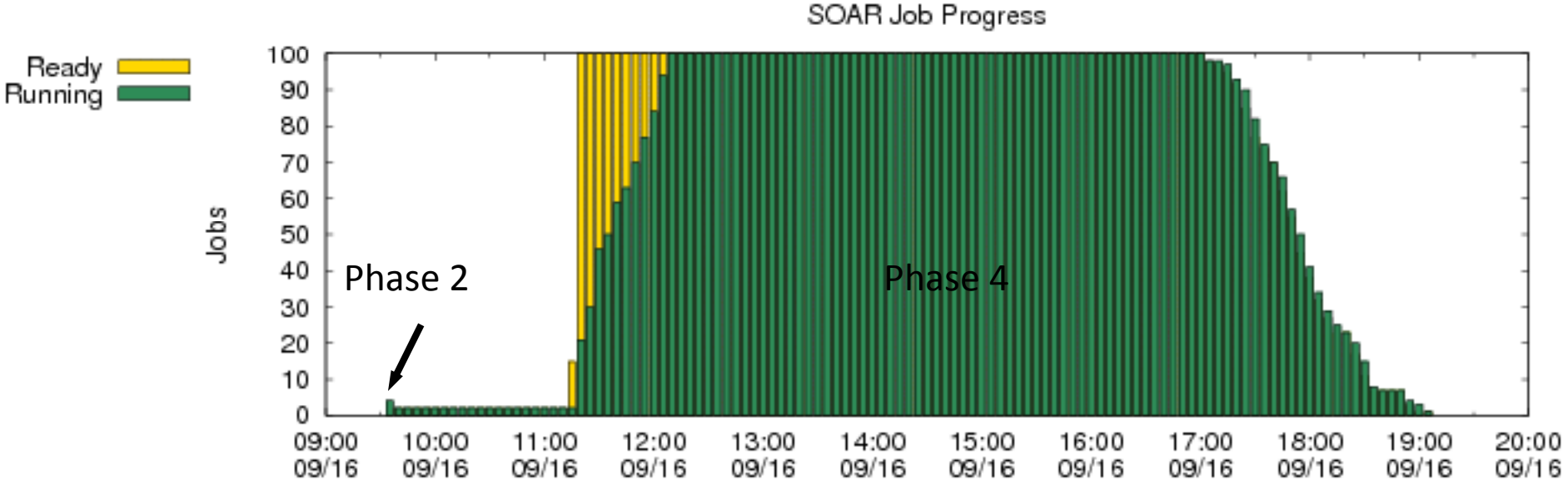
Phase 4: run Markov chains

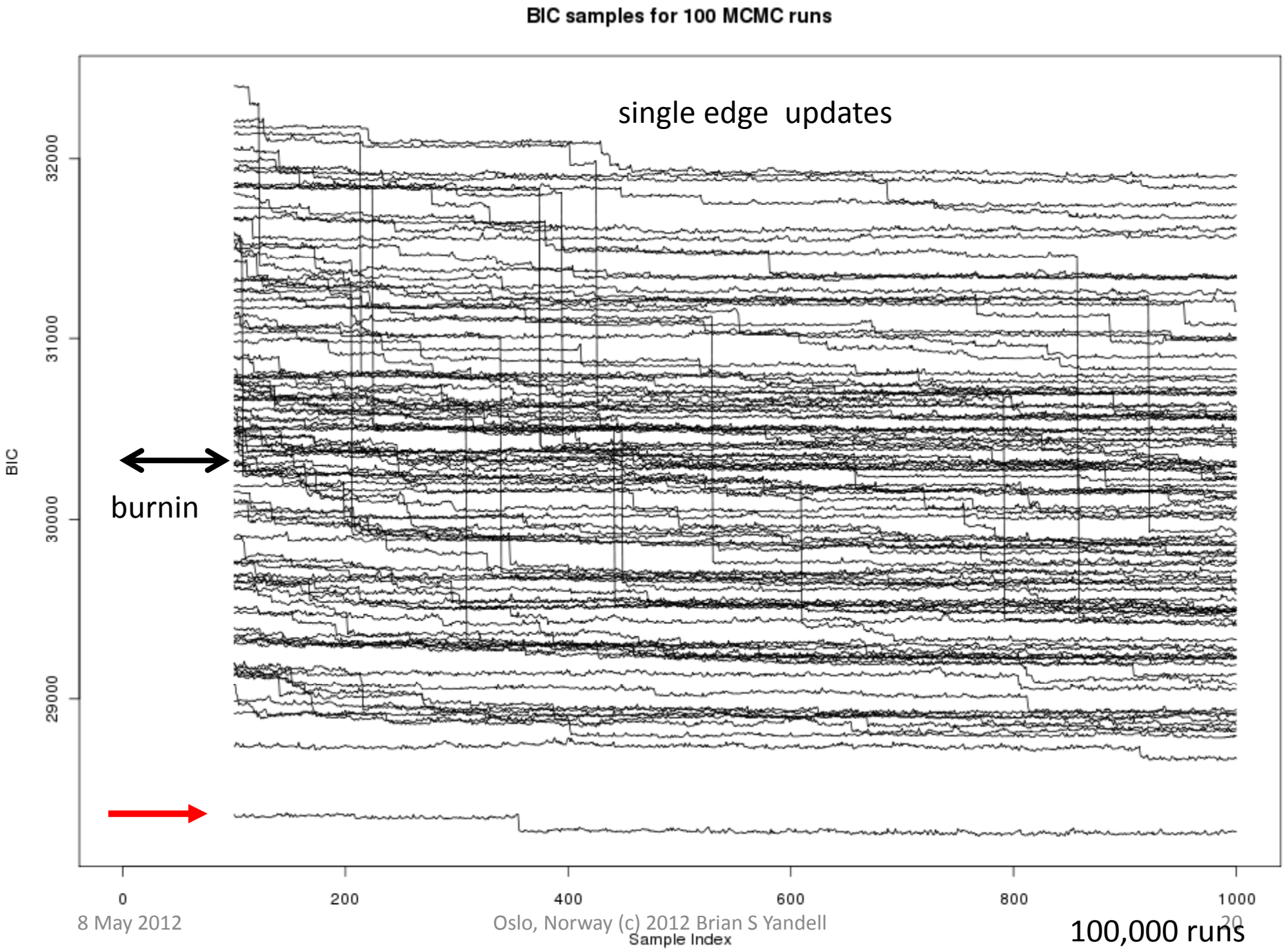
Phase 5: combine results



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



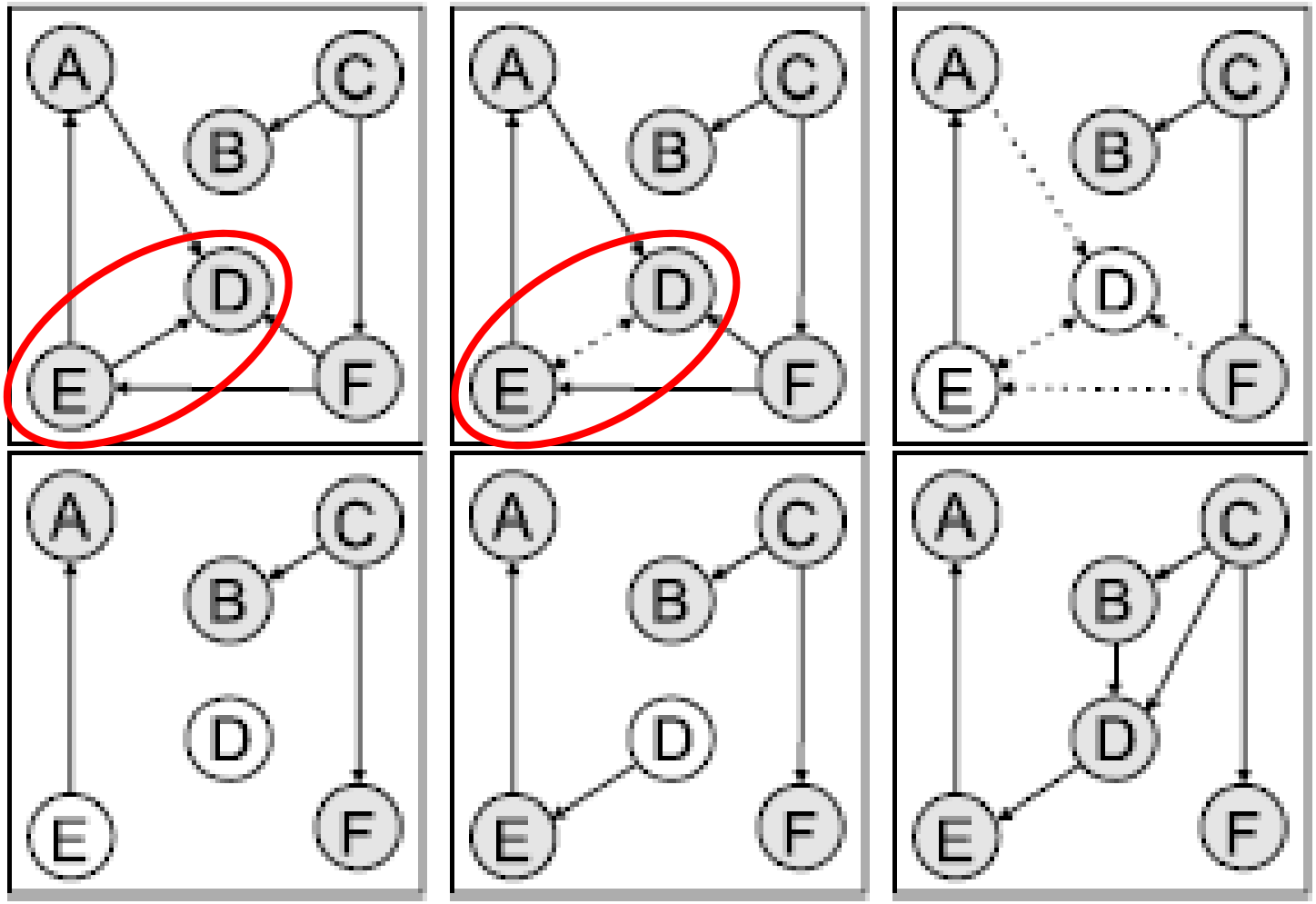

BIC samples for 100 MCMC runs
single edge updates
burnin
BIC
32000
31000
30000
29000
0
200
400
600
800
1000
Sample Index

100,000 runs

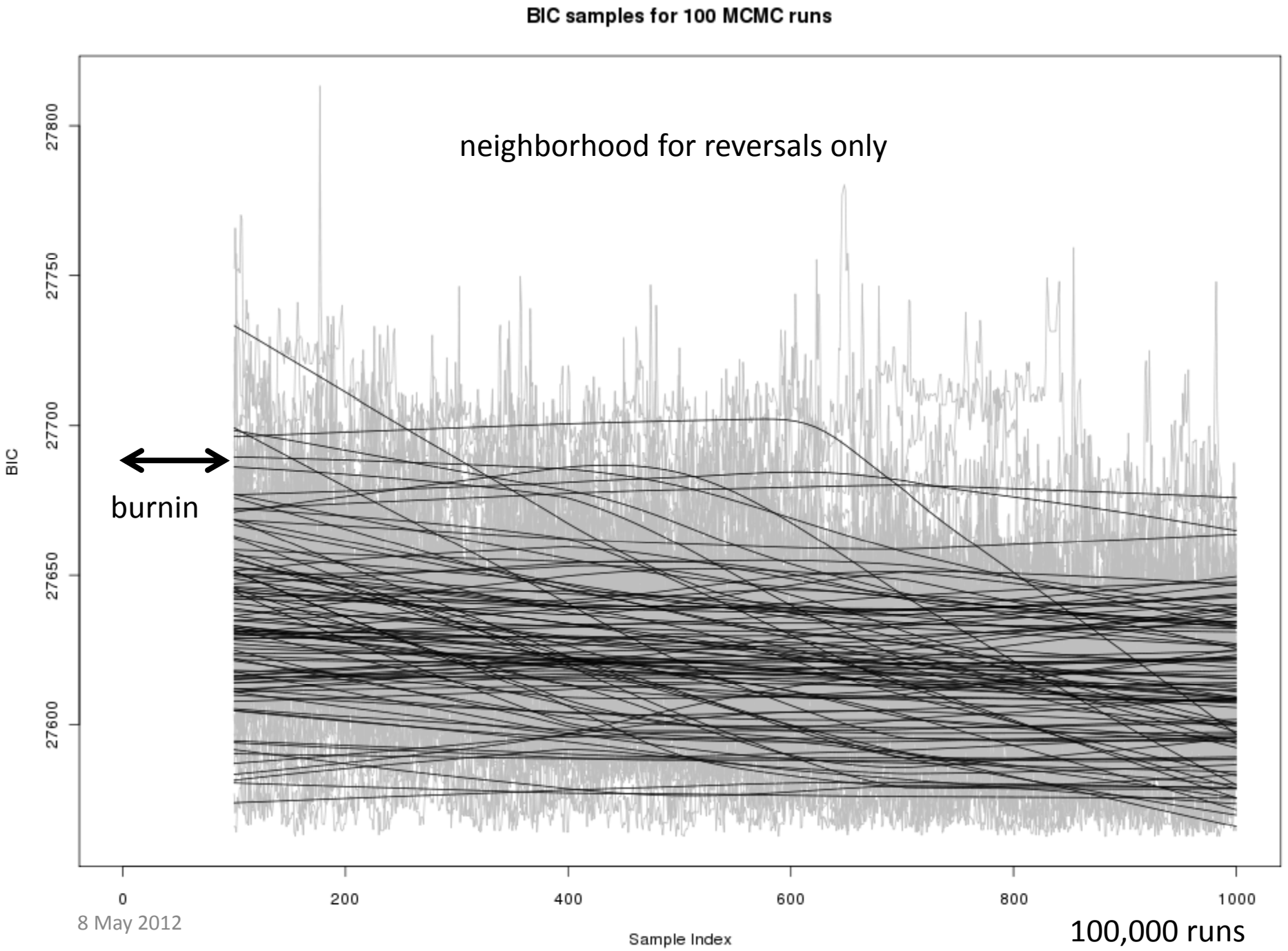
neighborhood edge reversal

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.


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

100,000 runs

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_γ from biological knowledge B
  - $\Pr(G_\gamma \mid B, W) = \exp(-W * |B - G_\gamma|) / \text{constant}$
  - B = prob of edge given TF, PPI, KO database
    - derived using previous experiments, papers, etc.
  - G_γ = 0-1 matrix for graph with directed edges
- W = inferred weight of biological knowledge
  - W=0: no influence; W large: assumed correct
  - $P(W|B) = \phi \exp(-\phi W)$ 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 *Q*, bio info *B*
- $\Pr(G, W \mid Y, Q, B) = c \Pr(Y|G,Q)\Pr(G|B,W,Q)\Pr(W|B)$
  - $\Pr(Y|G,Q)$ is genetic architecture (QTLs)
    - using parent nodes of each trait as covariates
  - $\Pr(G|B,W,Q) = \Pr(G_Y|B,W)\ \Pr(G_{Q\rightarrow Y}|Q)$
    - $\Pr(G_Y|B,W)$ relates graph to biological info
    - $\Pr(G_{Q\rightarrow Y}|Q)$ relates genotype to phenotype

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)

encoding biological knowledge *B*

transcription factors, DNA binding (causation)

$$B_{ij} = \frac{\lambda e^{-\lambda p}}{\lambda e^{-\lambda p} + (1 - e^{-\lambda})}$$

- p = p-value for TF binding of $i \rightarrow j$
- truncated exponential (λ) when TF $i \rightarrow j$
- uniform if no detection relationship
- Bernard, Hartemink (2005) *Pac Symp Biocomp*

encoding biological knowledge *B*

protein-protein interaction (association)

$$B_{ij} = B_{ji} = \frac{\text{posterior odds}}{1 + \text{posterior odds}}$$

- post odds = prior odds * LR
- use positive and negative gold standards
- Jansen et al. (2003) *Science*

encoding biological knowledge *B*
gene ontology(association)

$$B_{ij} = B_{ji} = c \bullet mean(sim(GO_i, GO_j))$$

- GO = molecular function, processes of gene
- sim = maximum information content across common parents of pair of genes
- Lord et al. (2003) *Bioinformatics*

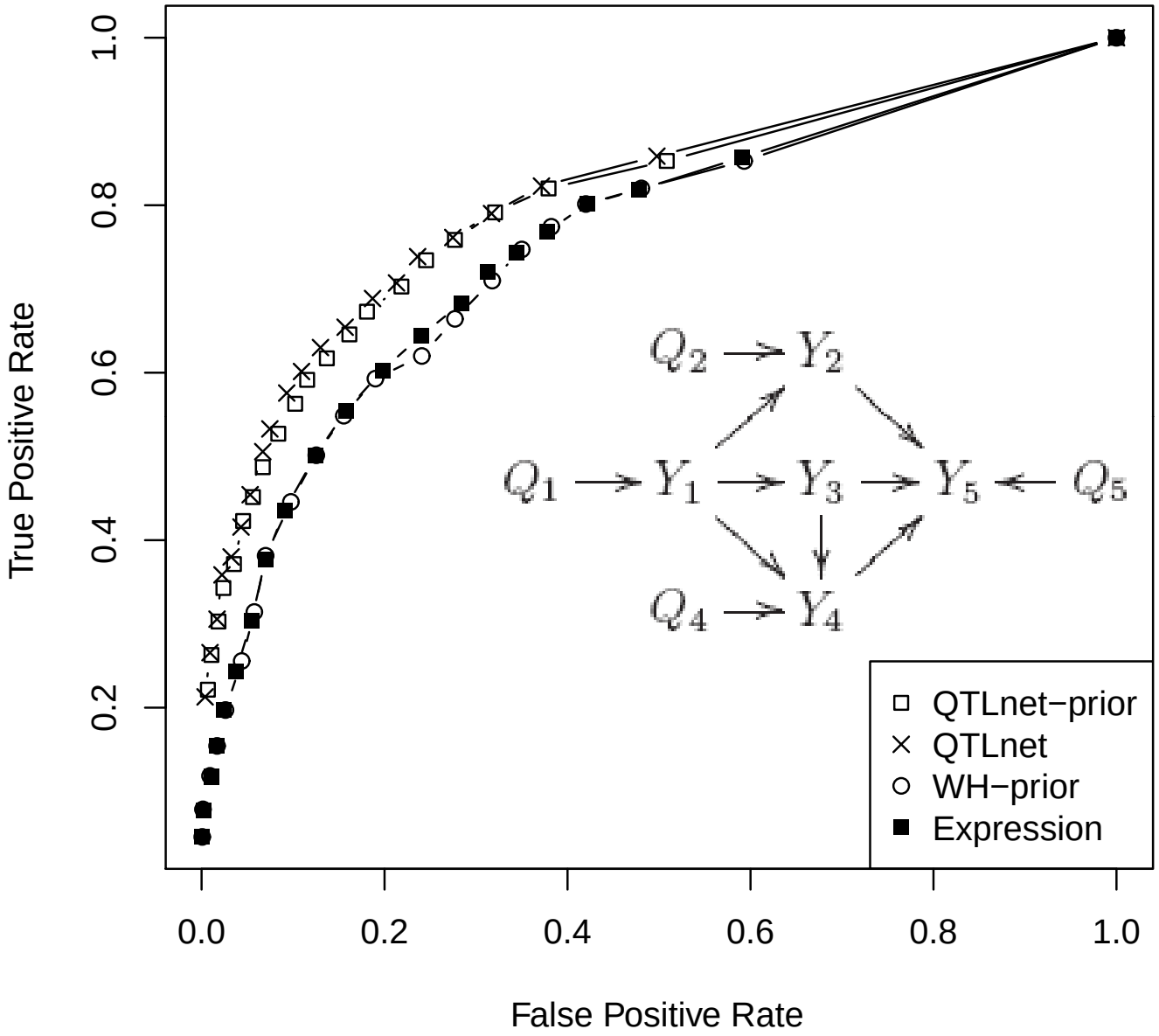
MCMC with pathway information

- sample new network G from proposal $R(G^*|G)$
  - add or drop edges; switch causal direction
- sample QTLs Q from proposal $R(Q^*|Q,Y)$
  - e.g. Bayesian QTL mapping given pa(Y)
- accept new network (G*,Q*) with probability
- A = min(1, $f(G,Q|G^*,Q^*)/ f(G^*,Q^*|G,Q)$)
  - $f(G,Q|G^*,Q^*)$ = Pr(Y|G*,Q*)Pr(G*|B,W,Q*)/$R(G^*|G)R(Q^*|Q,Y)$
- sample W from proposal R(W*|W)
- accept new weight W* with probability …

ROC curve simulation

open = QTLnet

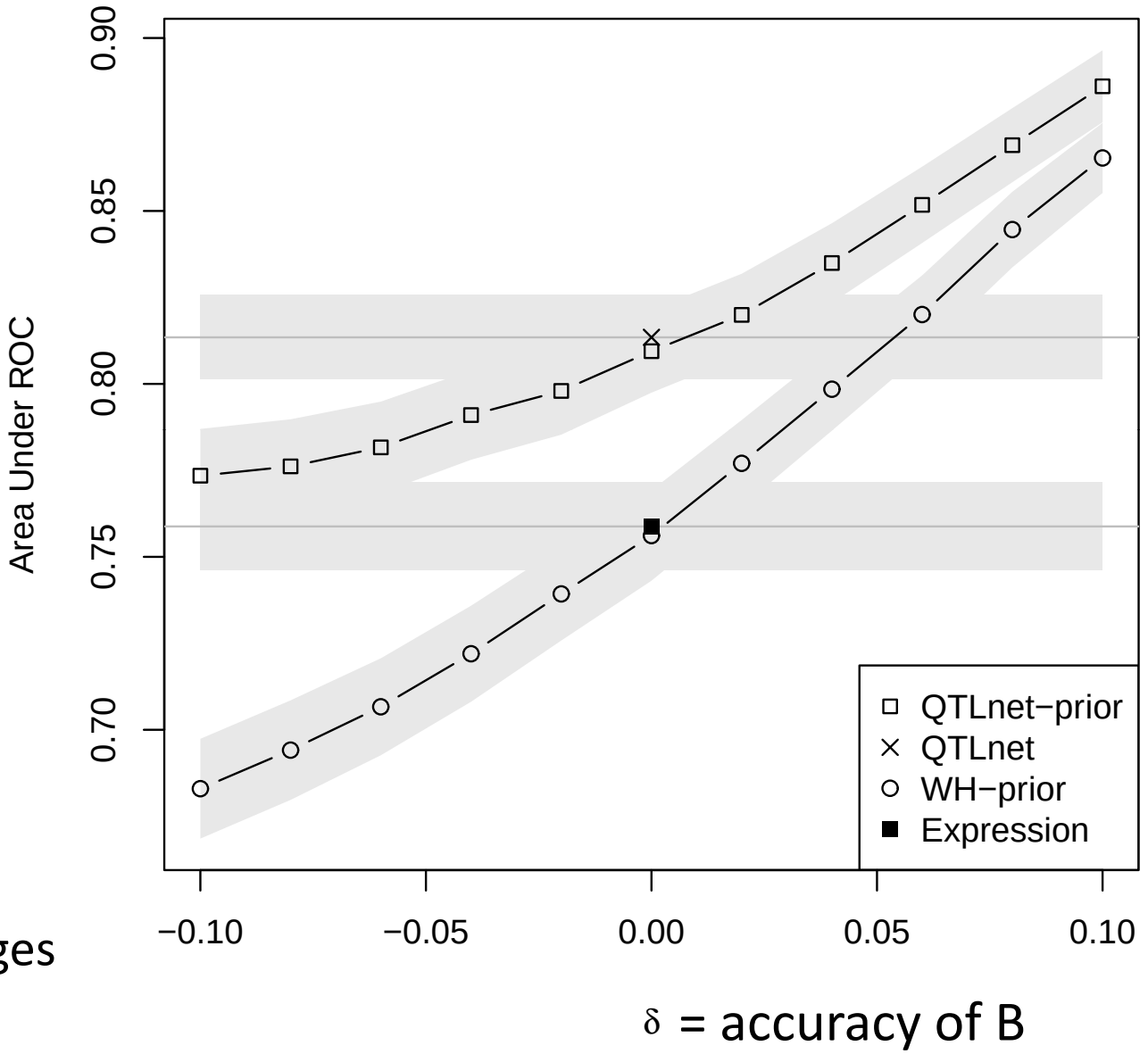
closed = phenotypes only



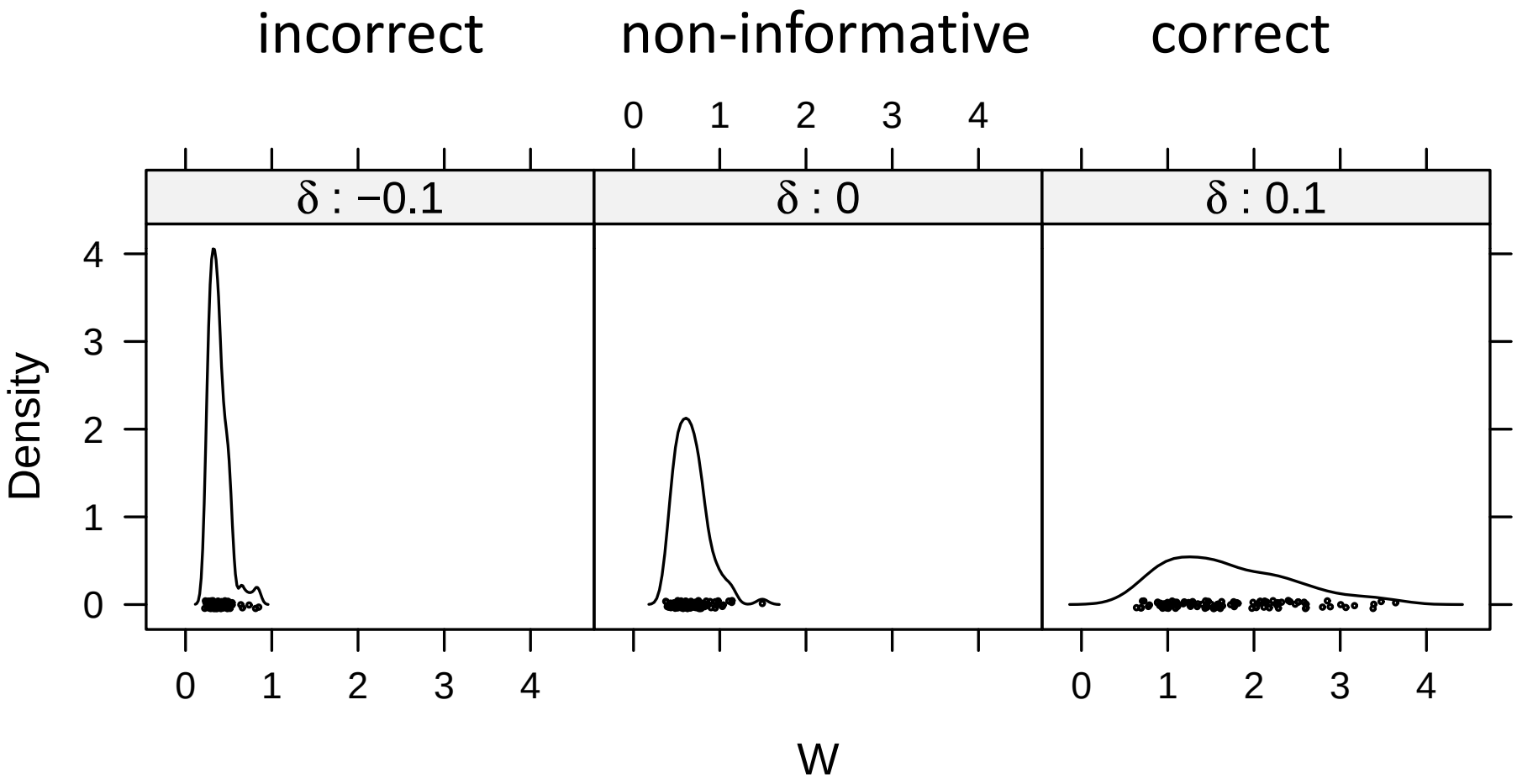
integrated ROC curve

2x2: genetics pathways

probability classifier
ranks true > false edges

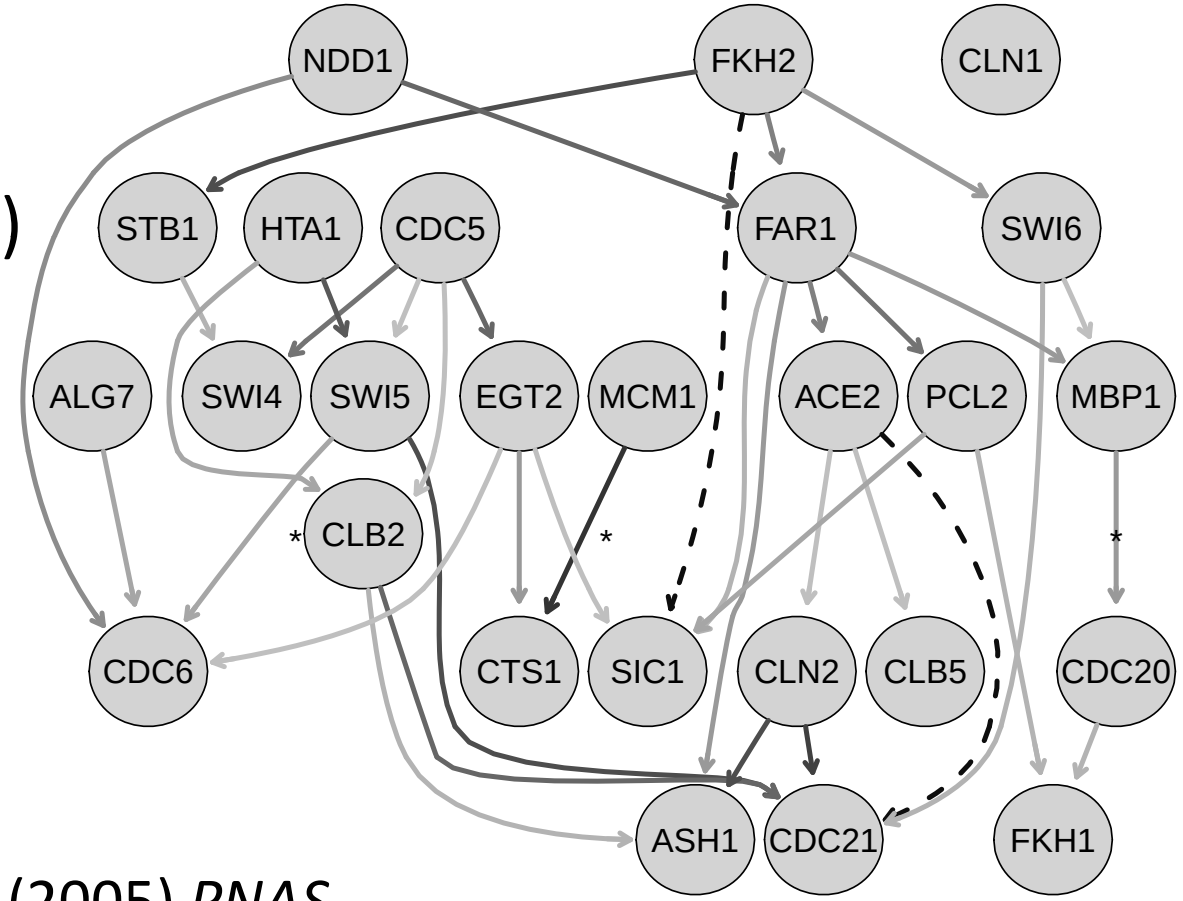


weight on biological knowledge



yeast data—partial success

26 genes
36 inferred edges
dashed: indirect (2)
starred: direct (3)
missed (39)
reversed (0)

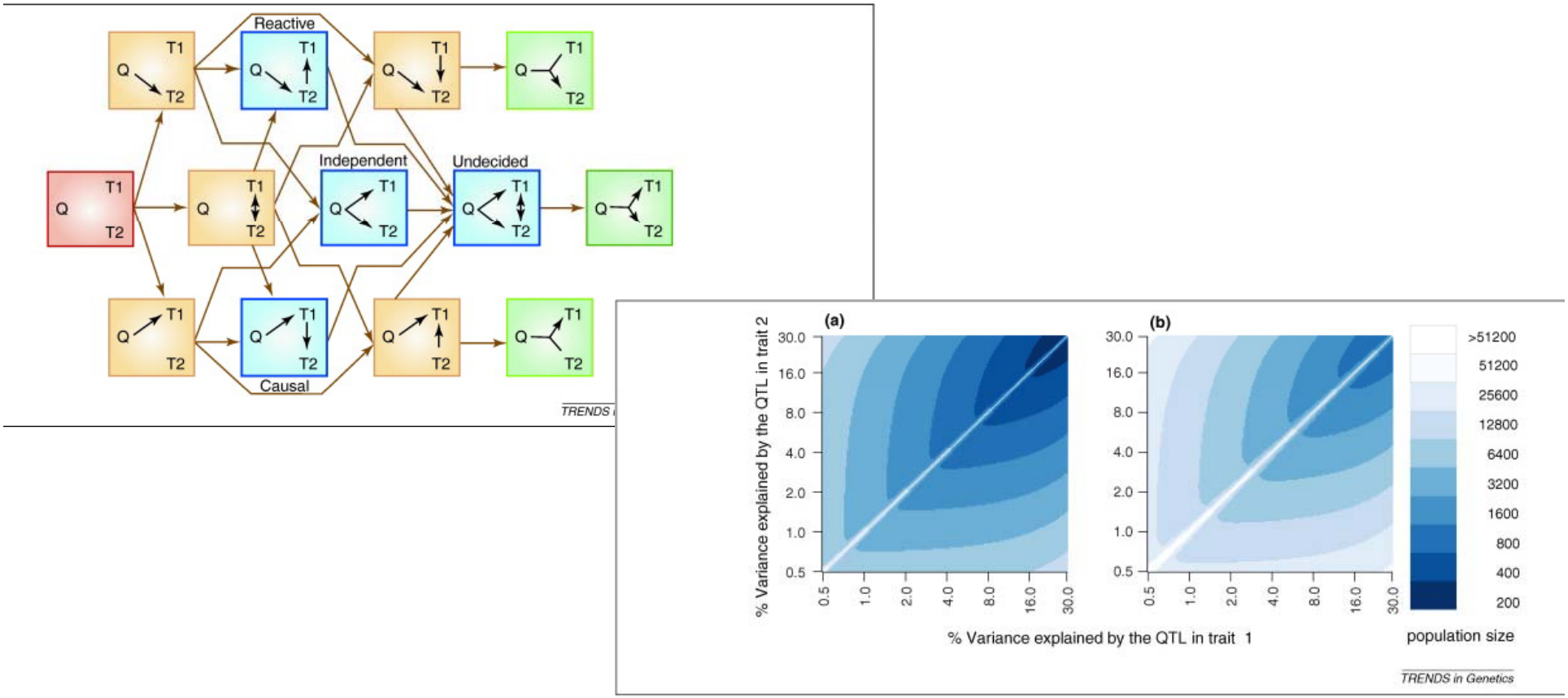


Data: Brem, Kruglyak (2005) *PNAS*

limits of causal inference

- Computing costs already discussed
- Noisy data leads to false positive causal calls
  - Unfaithfulness assumption violated
  - Depends on sample size and omic technology
  - And on graph complexity (d = maximal path length i$\rightarrow$j)
  - Profound limits
- Uhler C, Raskutti G, Buhlmann P, Yu B (2012 in prep) Geometry of faithfulness assumption in causal inference.
- Yang Li, Bruno M. Tesson, Gary A. Churchill, Ritsert C. Jansen (2010) Critical reasoning on causal inference in genome-wide linkage and association studies. *Trends in Genetics 26*: 493-498.

sizes for reliable causal inference genome wide linkage & association



Li, Tesson, Churchill, Jansen (2010) *Trends in Genetics*

limits of causal inference

unfaithful: false positive edges

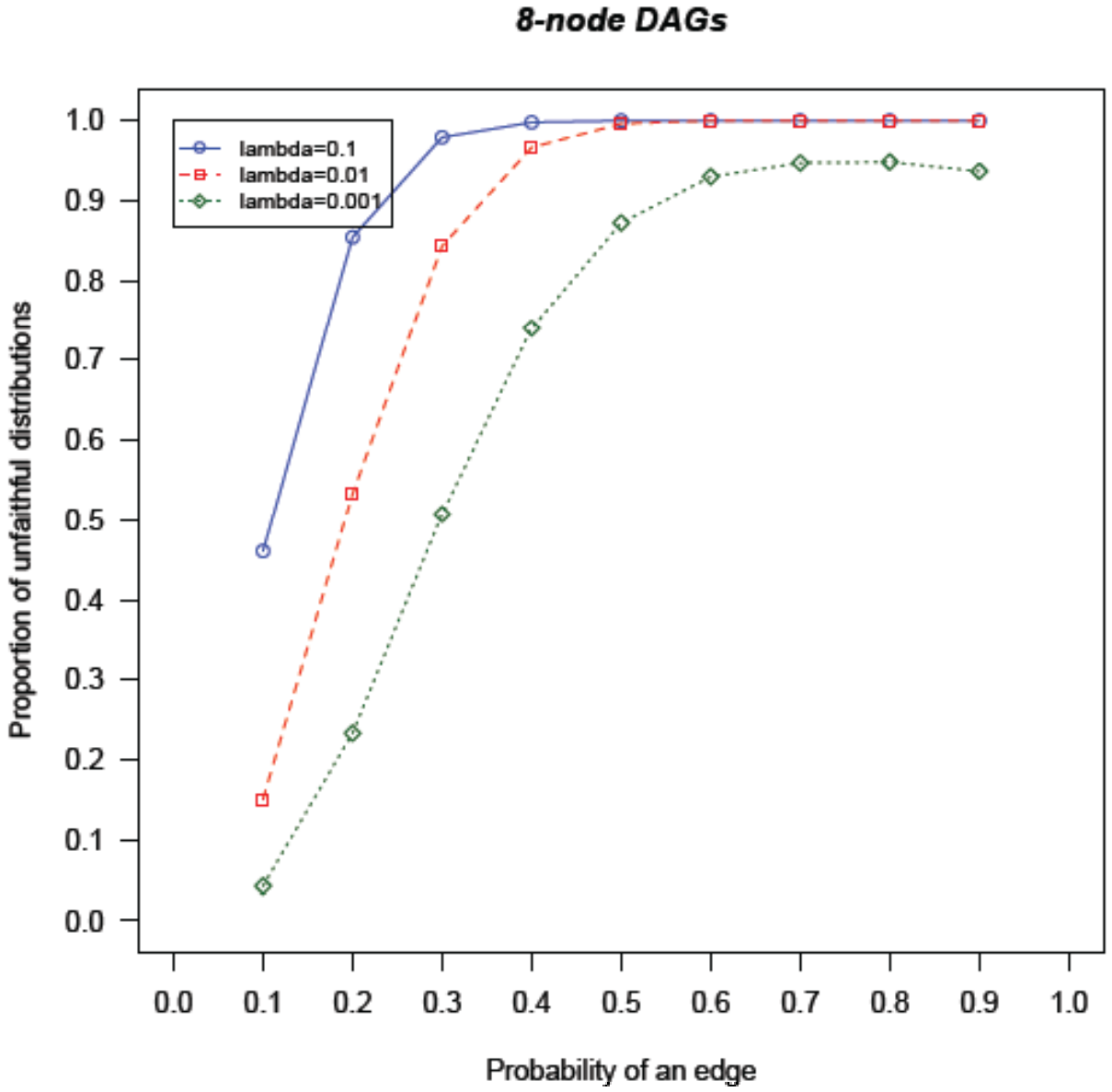
$\lambda = \min|\mathrm{cor}(Y_i, Y_j)|$

$\lambda = c \bullet \mathrm{sqrt}(dp/n)$

d=max degree

p=# nodes

n=sample size

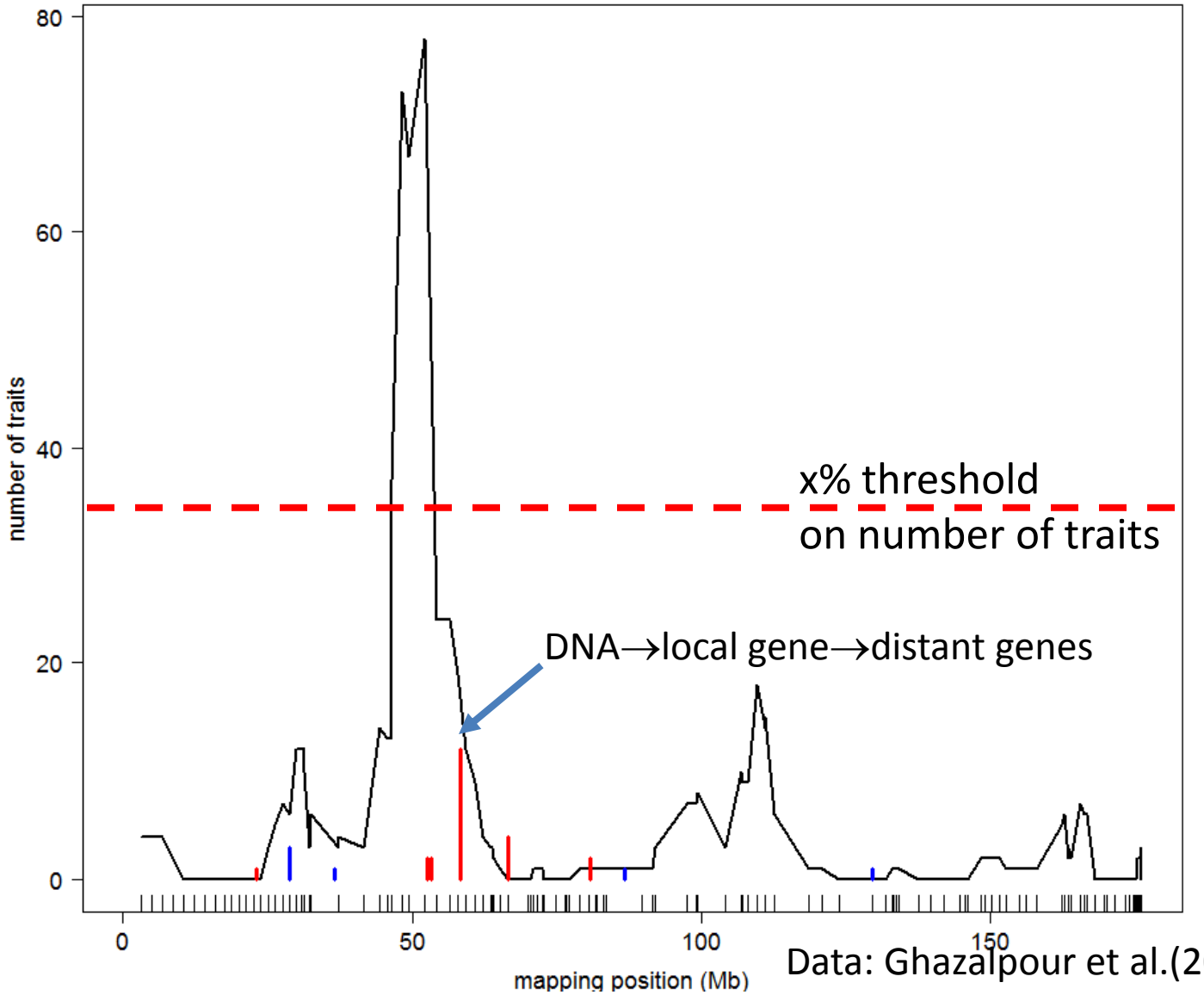


Uhler, Raskutti, Buhlmann, Yu (2012 in prep)

Thanks!

- Grant support
  - NIH/NIDDK 58037, 66369
  - NIH/NIGMS 74244, 69430
  - NCI/ICBP U54-CA149237
  - NIH/R01MH090948
- Collaborators on papers and ideas
  - Alan Attie & Mark Keller, Biochemistry
  - Karl Broman, Aimee Broman, Christina Kendziorski

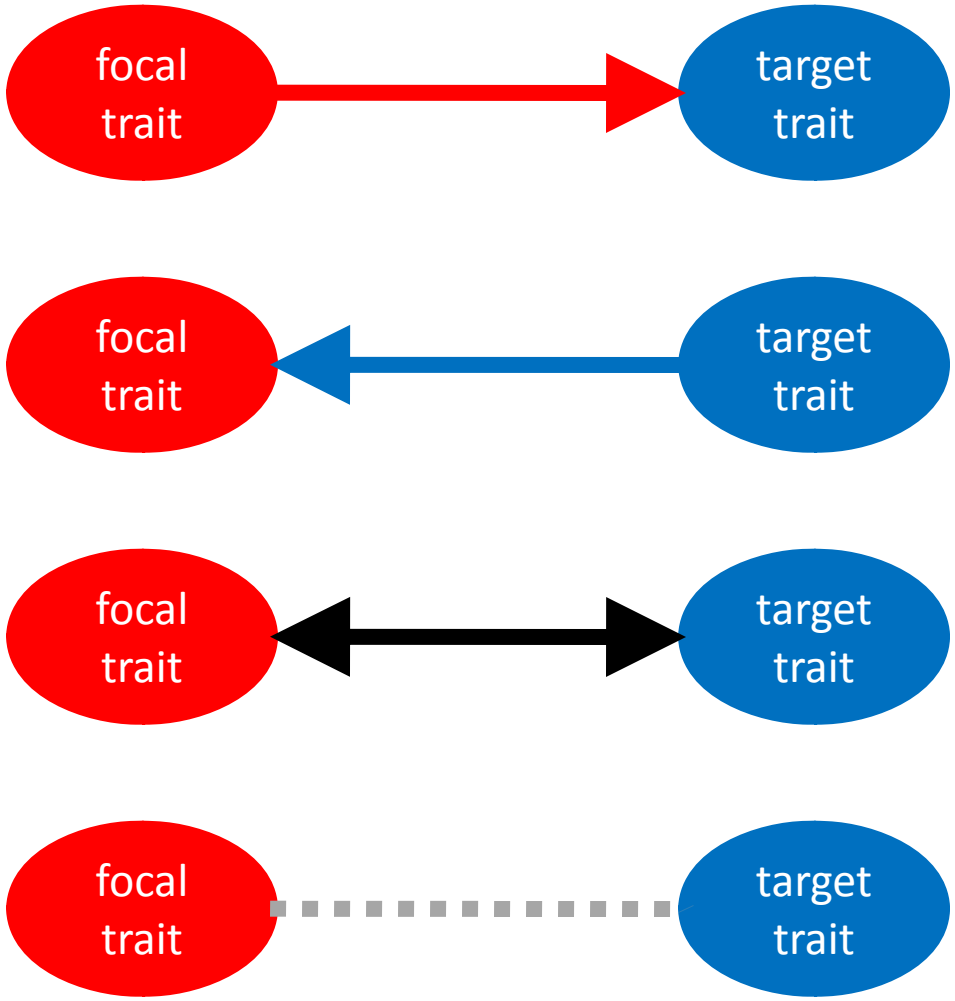
BxH ApoE-/- chr 2: hotspot



Data: Ghazalpour et al.(2006) *PLoS Genetics*

causal model selection choices

in context of larger, unknown network



causal

reactive

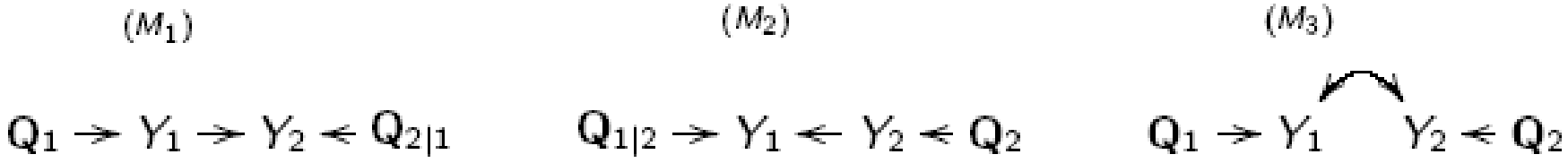
correlated

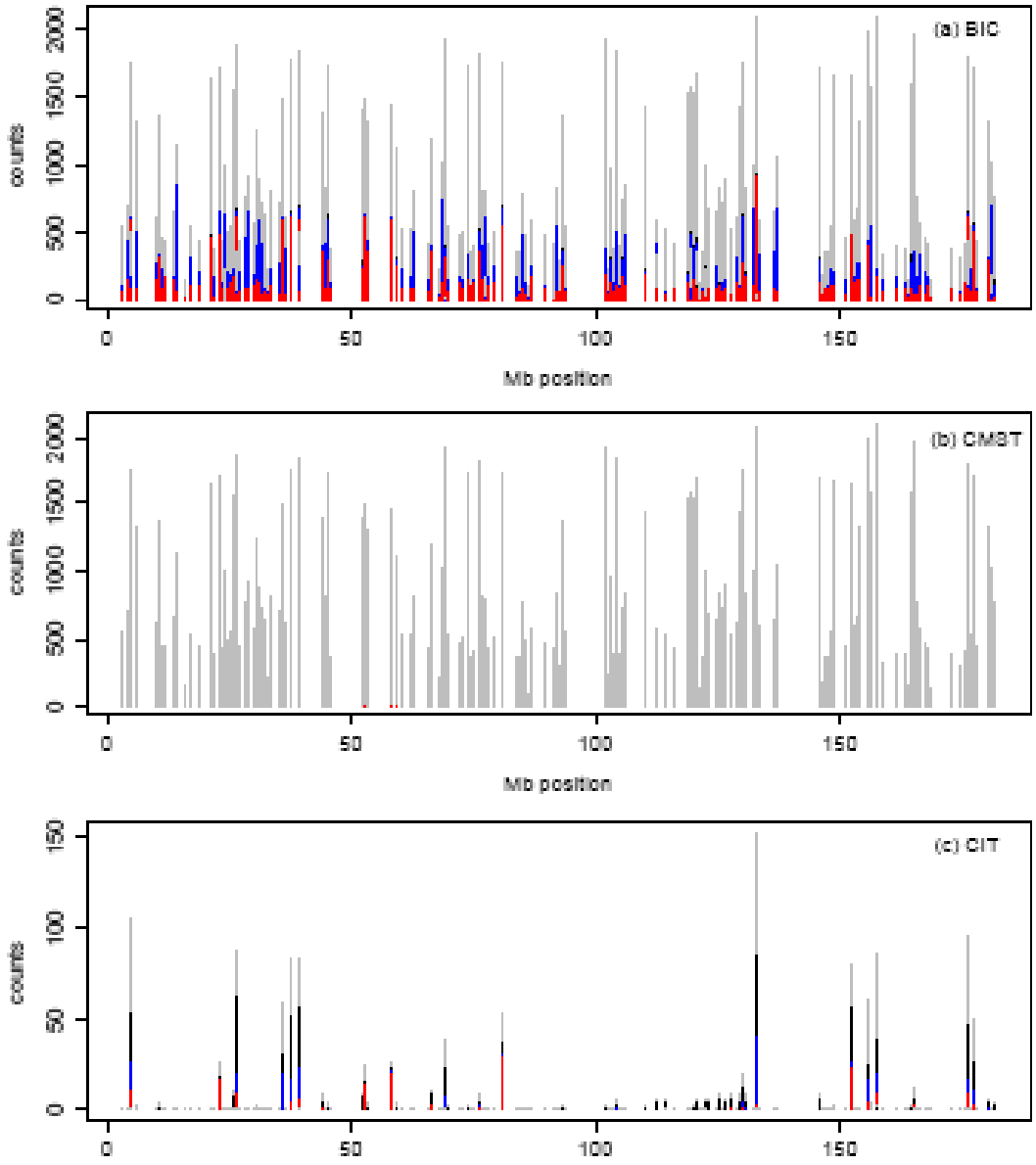
uncorrelated

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. (2010) PhD thesis
  - Chaibub Neto et al. (2012) *Genetics* (in review)

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



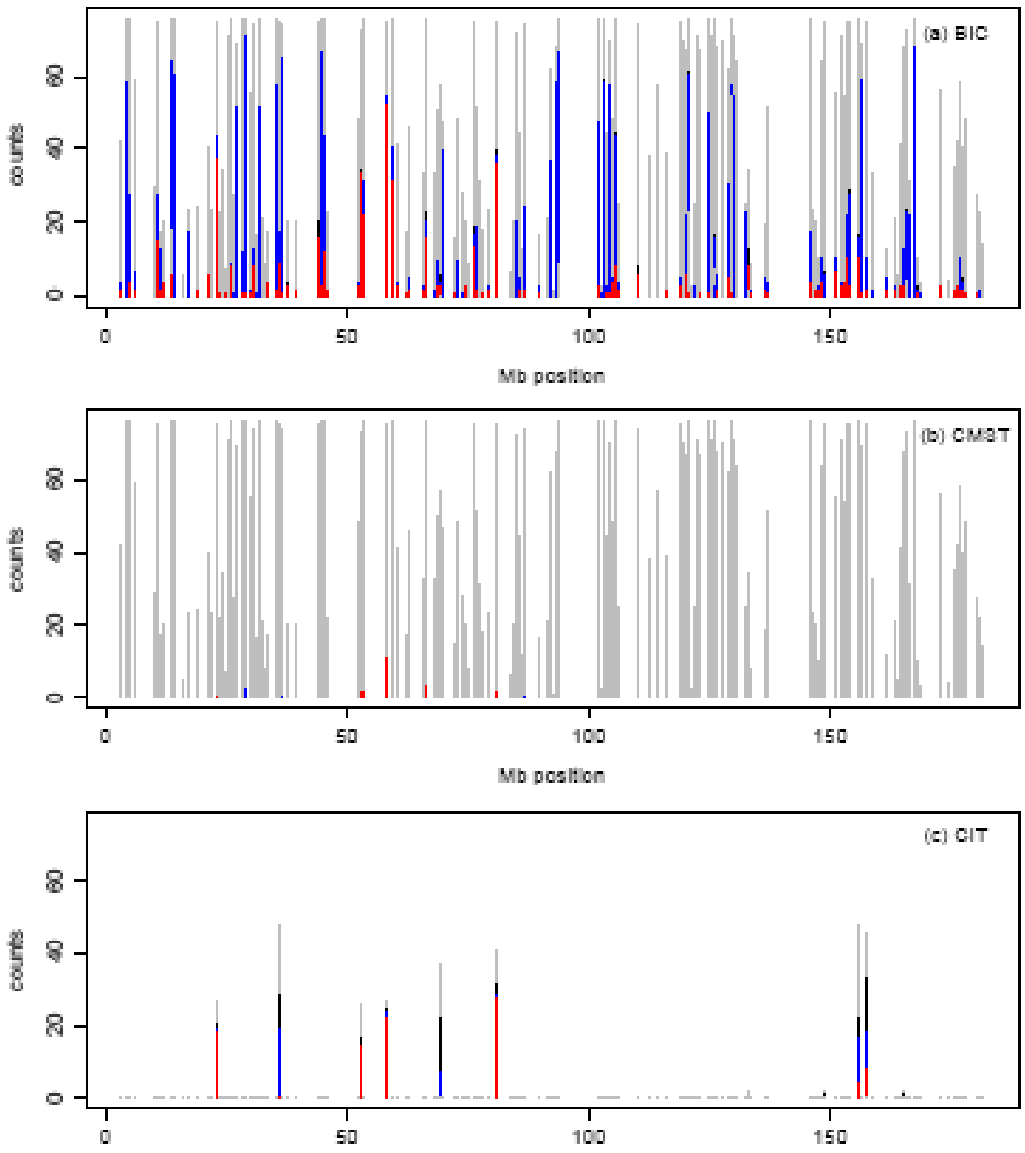


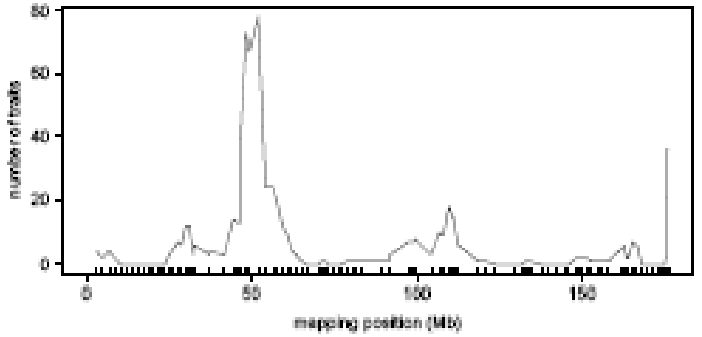
BxH ApoE-/- study
Ghazalpour et al. (2008)
PLoS Genetics

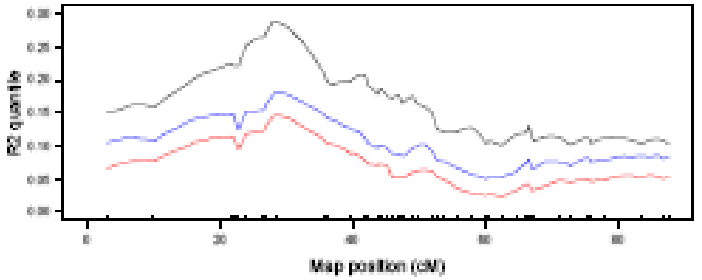
Liver expression data in a mice intercross.

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.