Causal Network Models for Correlated Quantitative Traits

Brian S. Yandell

UW-Madison

October 2011

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

outline

- Correlation and causation
- Correlated traits in organized groups
  - modules and hotspots
  - Genetic vs. environmental correlation
- QTL-driven directed graphs
  - Assume QTLs known, causal network unknown
- Causal graphical models in systems genetics
  - QTLs unknown, causal network unknown
- Scaling up to larger networks
  - Searching the space of possible networks
  - Dealing with computation

“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*
- Population-based estimates (from a sample of individuals) may not reflect processes within an individual.

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 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 $\rightarrow$ B, B $\rightarrow$ C, then A $\rightarrow$ C
  - local (Markov): events have only proximate causes
  - asymmetric: if A $\rightarrow$ B, then B cannot $\rightarrow$ 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
  - $Y_1 = \mu_1 + e$
  - $Y_2 = \mu_2 + \beta_1 \bullet Y_2 + e$
- adding in QTLs: $Q_1 \rightarrow Y_1 \rightarrow Y_2 \leftarrow Q_2$
  - $Y_1 = \mu_1 + \theta_1 \bullet Q_1 + e$
  - $Y_2 = \mu_2 + \beta_1 \bullet Y_1 + \theta_2 \bullet Q_2 + e$

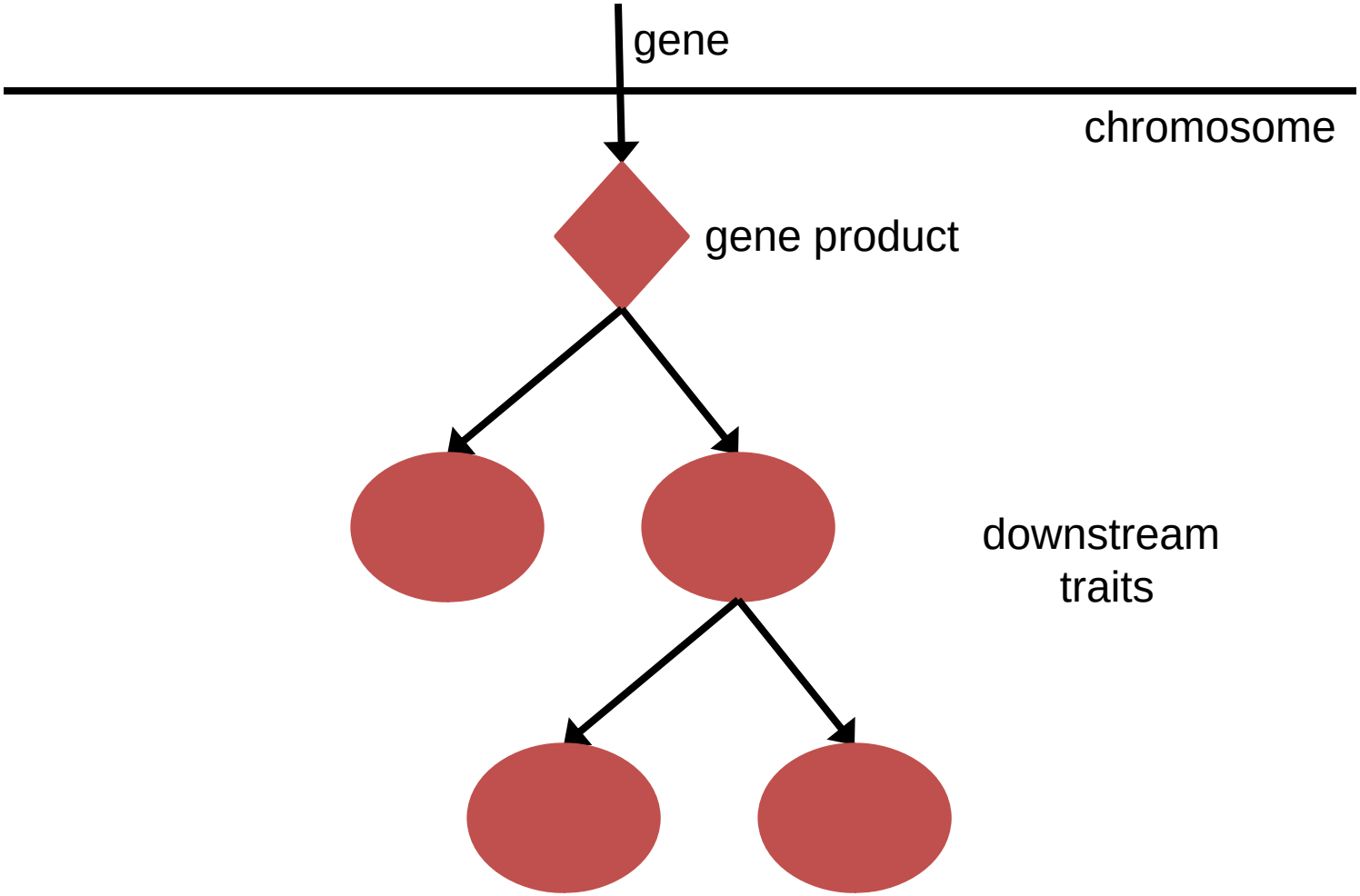
organizing correlated traits

- functional grouping from prior studies
  - GO, KEGG; KO panels; TF and PPI databases
- co-expression modules (Horvath talk today)
- 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

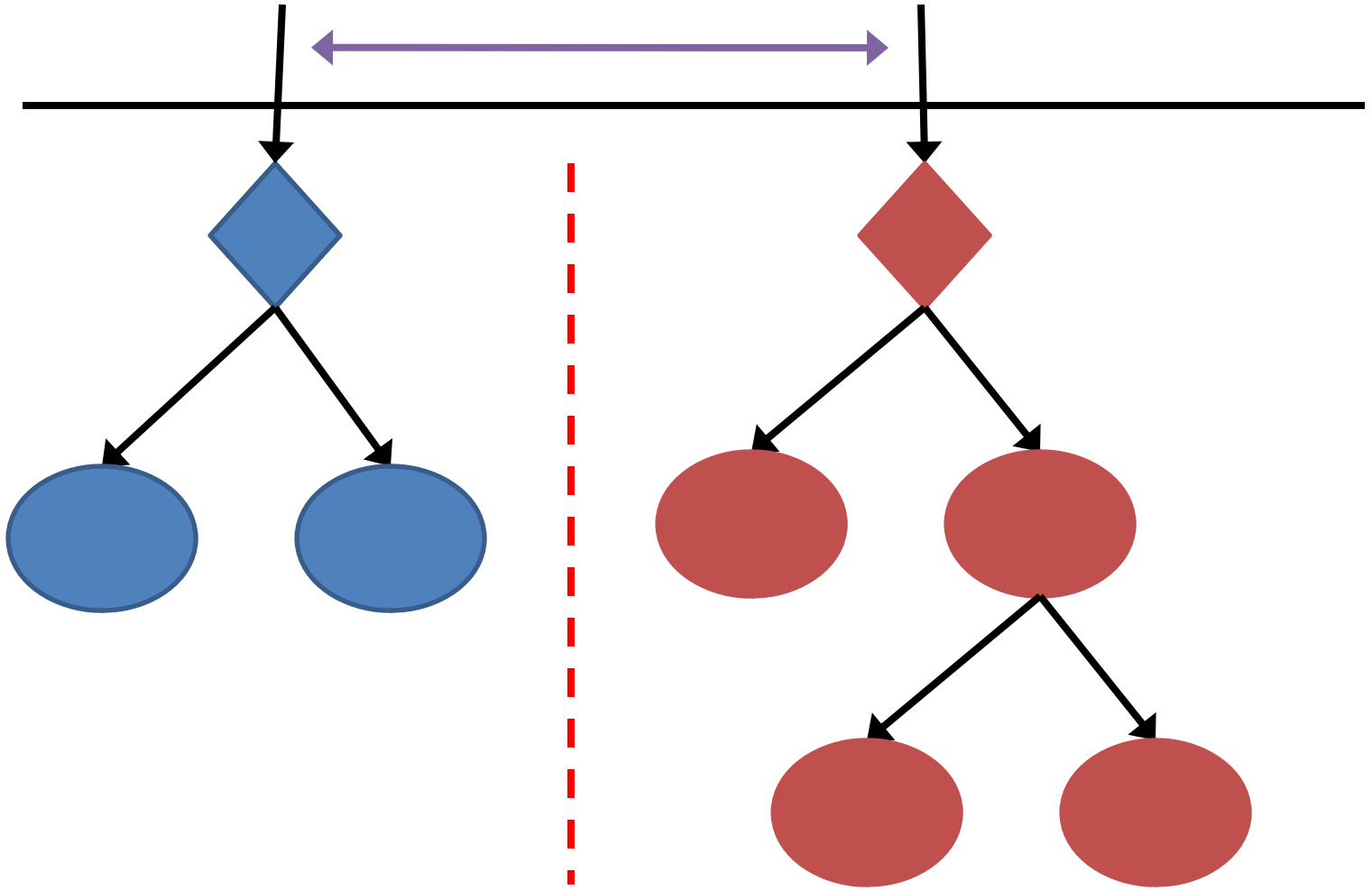
Correlated traits in a hotspot

- why are traits correlated?
  - Environmental: hotspot is spurious
  - 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



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).

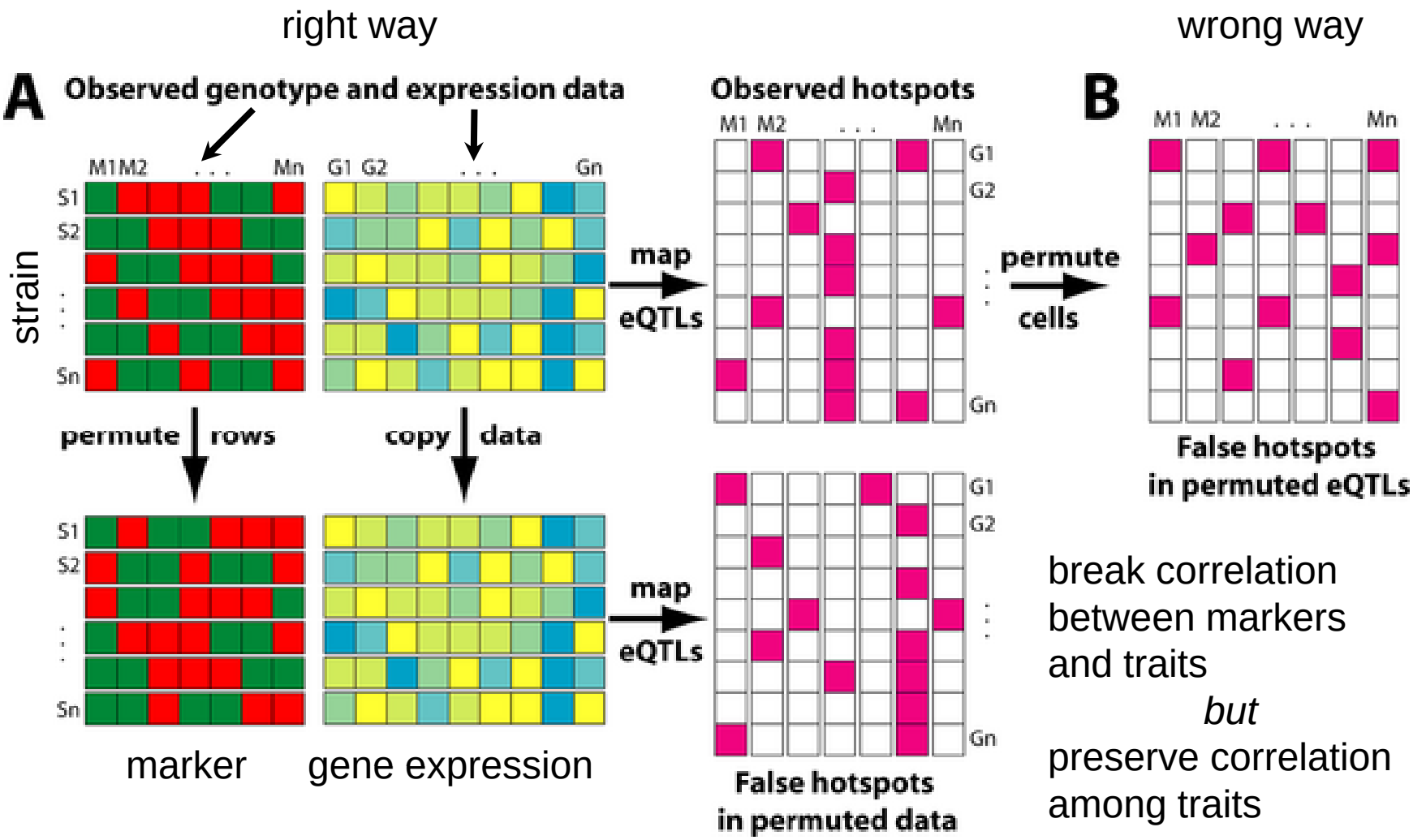
hotspot permutation test

(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*)



break correlation between markers and traits

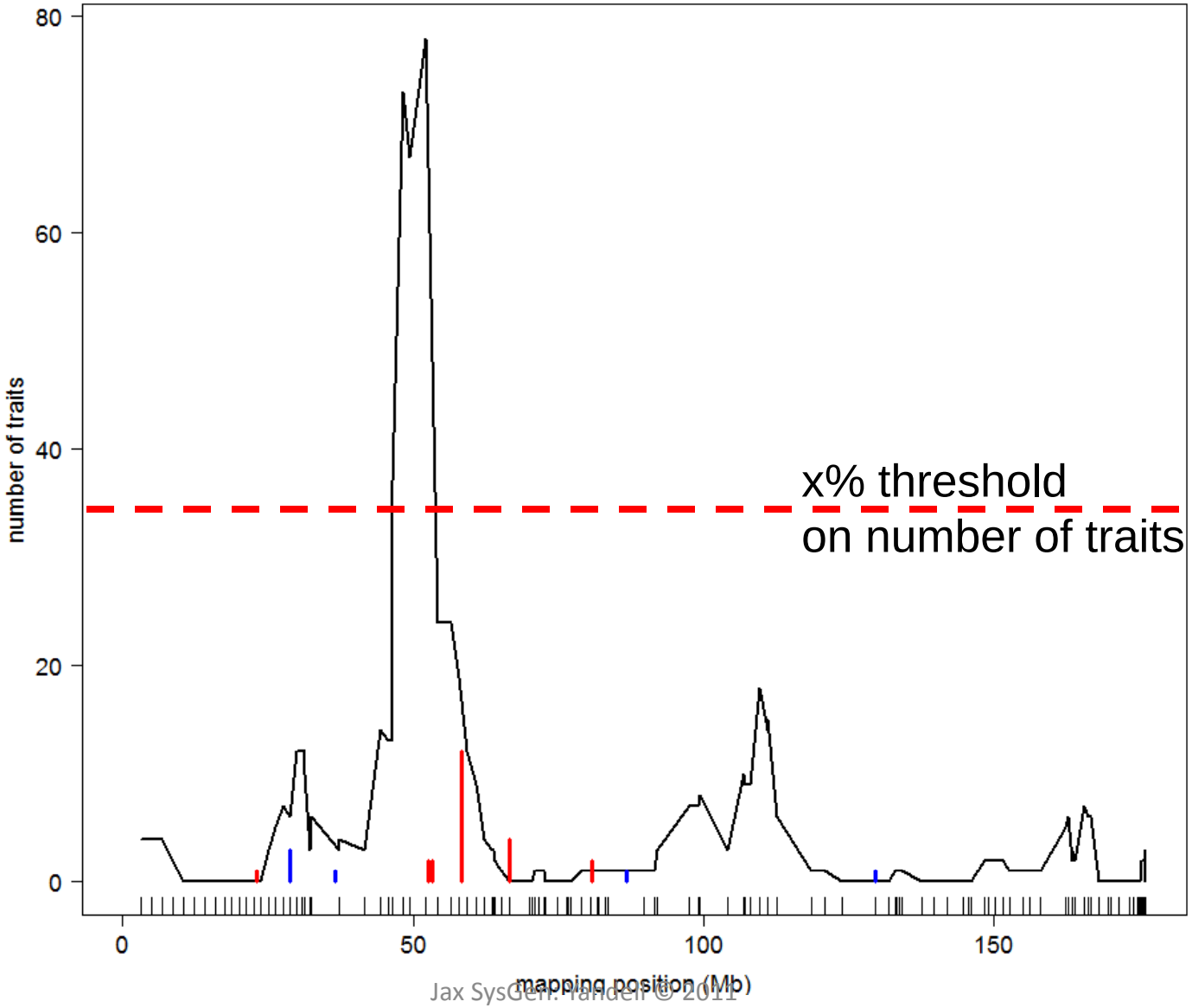
but

preserve correlation among traits

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

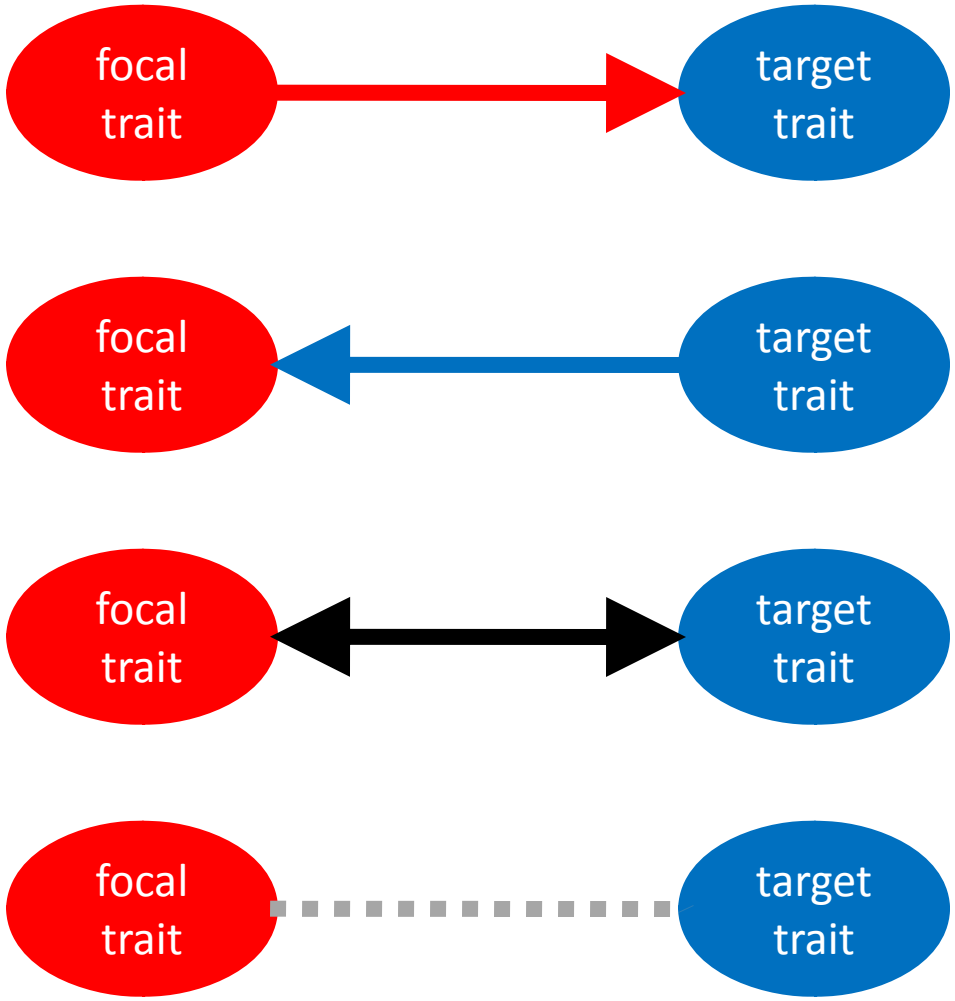
- detecting single trait with very large LOD
  - control FWER across genome
  - control FWER across all traits
- finding small “hotspots” with significant traits
  - all with large LODs
  - could indicate a strongly disrupted signal pathway
- sliding LOD threshold across hotspot sizes

BxH ApoE-/- chr 2: hotspot



causal model selection choices

in context of larger, unknown network



causal

reactive

correlated

uncorrelated

causal architecture

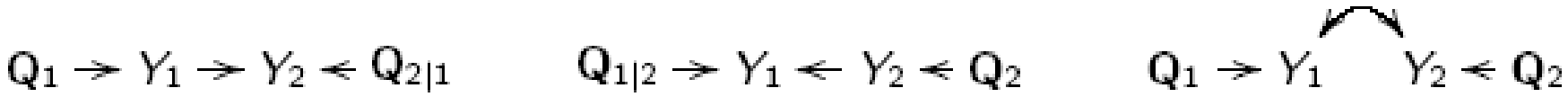
- how many traits are up/downstream of a trait?
  - focal trait causal to downstream target traits
  - record count at Mb position of focal gene
  - red = downstream, blue = upstream
- what set of target traits to consider?
  - all traits
  - traits in module or hotspot

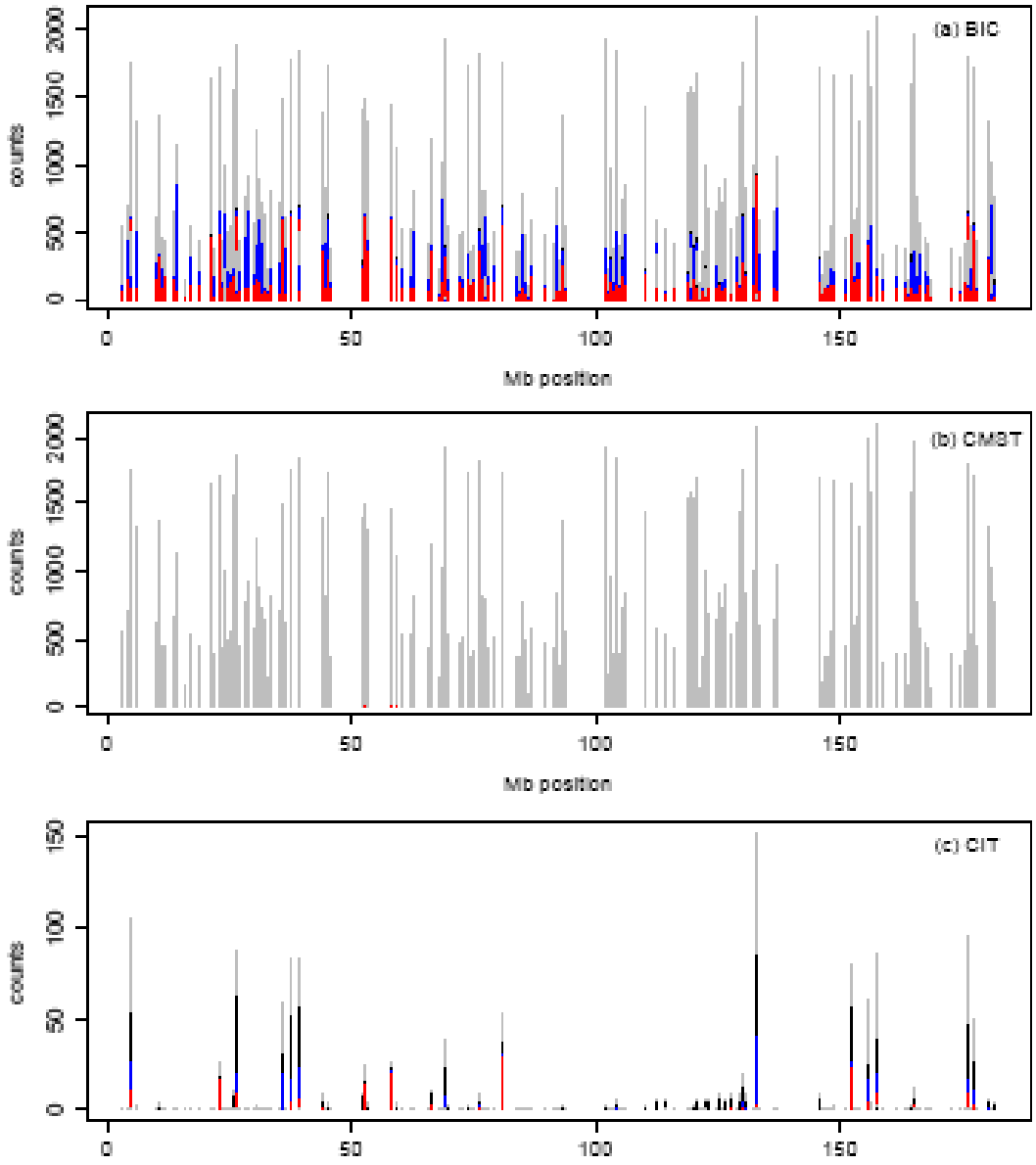
causal architecture references

- 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

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

(M_1) (M_2) (M_3)



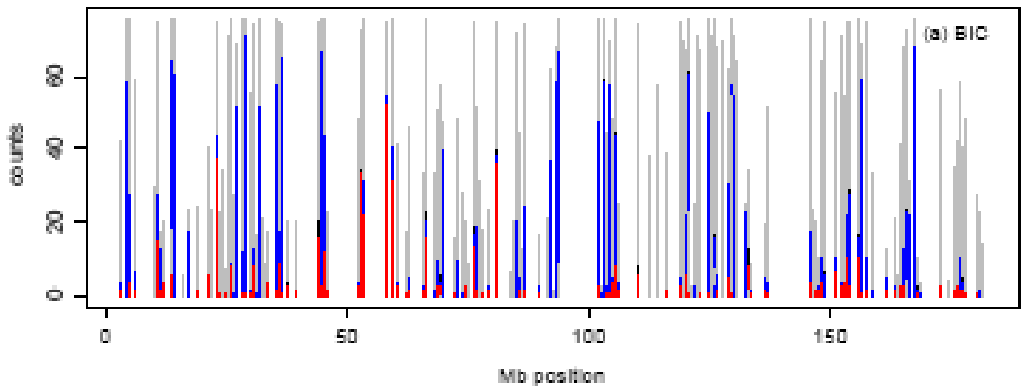


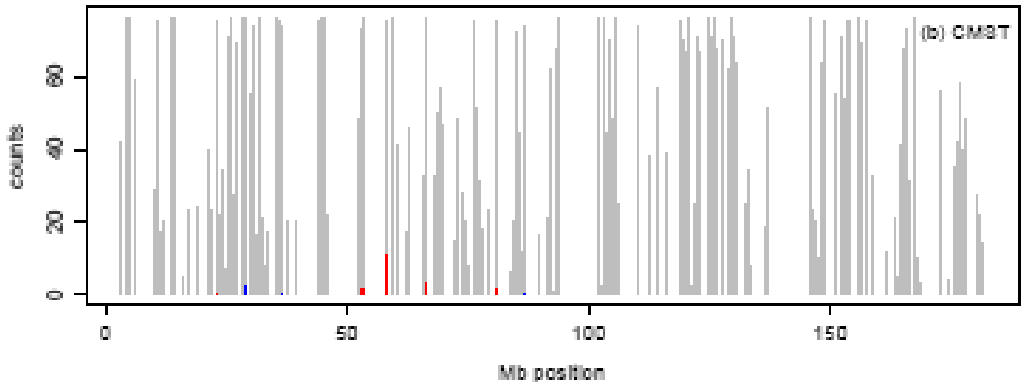
BxH ApoE-/- study
Ghazalpour et al. (2008)
PLoS Genetics

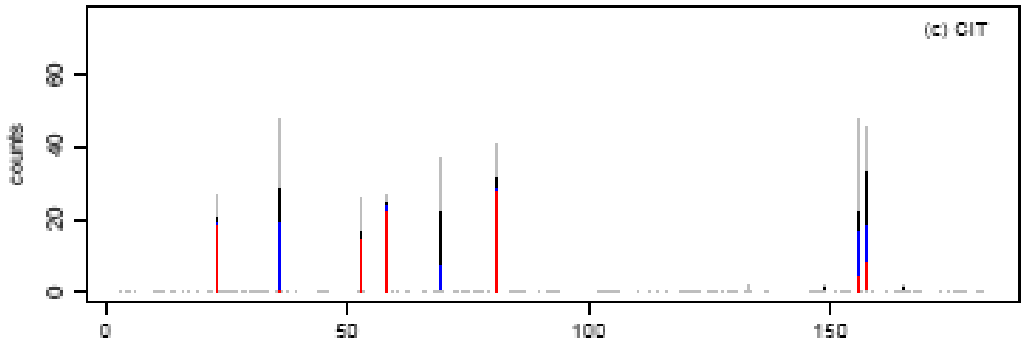
Liver expression data in a mice intercross.

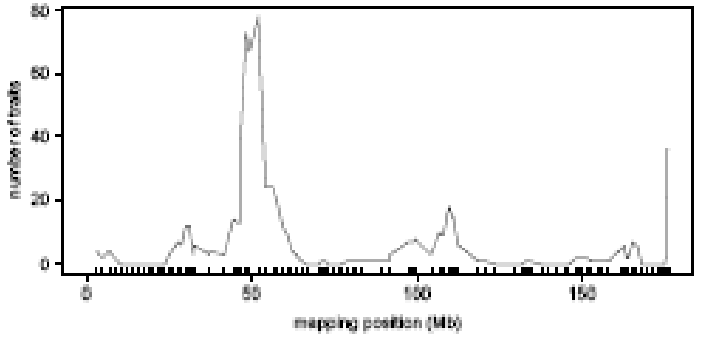
3,421 transcripts and 1,065 markers.

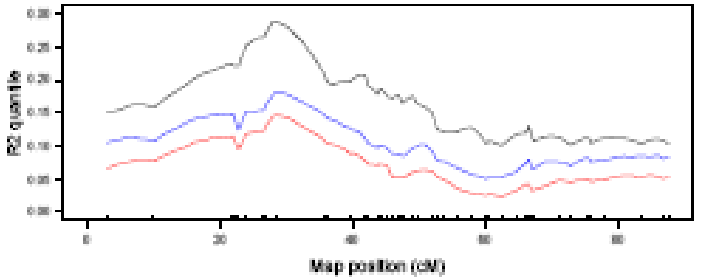
261 transcripts physically located on chr 2.











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

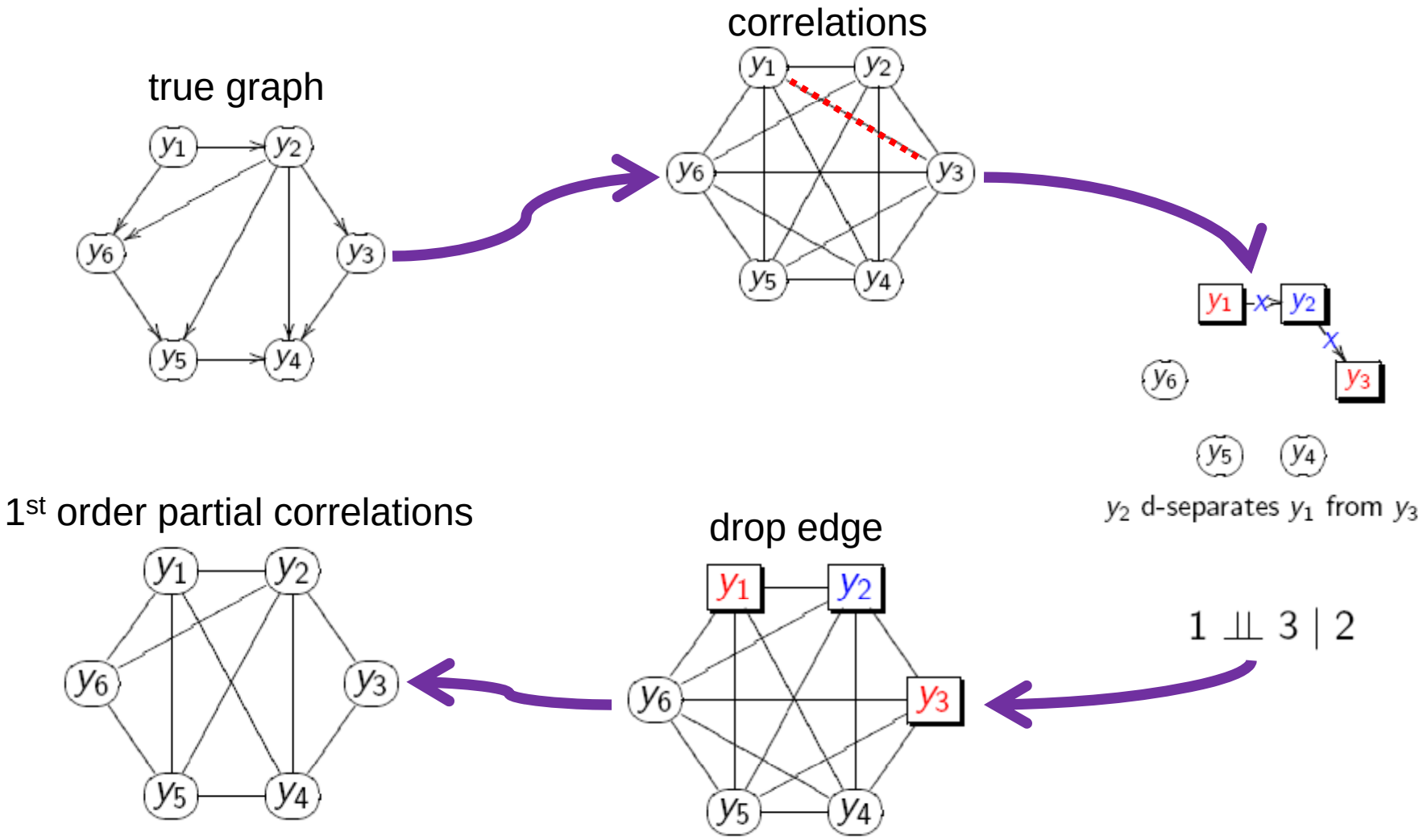
This collection of traits enriches for "immune system process".

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

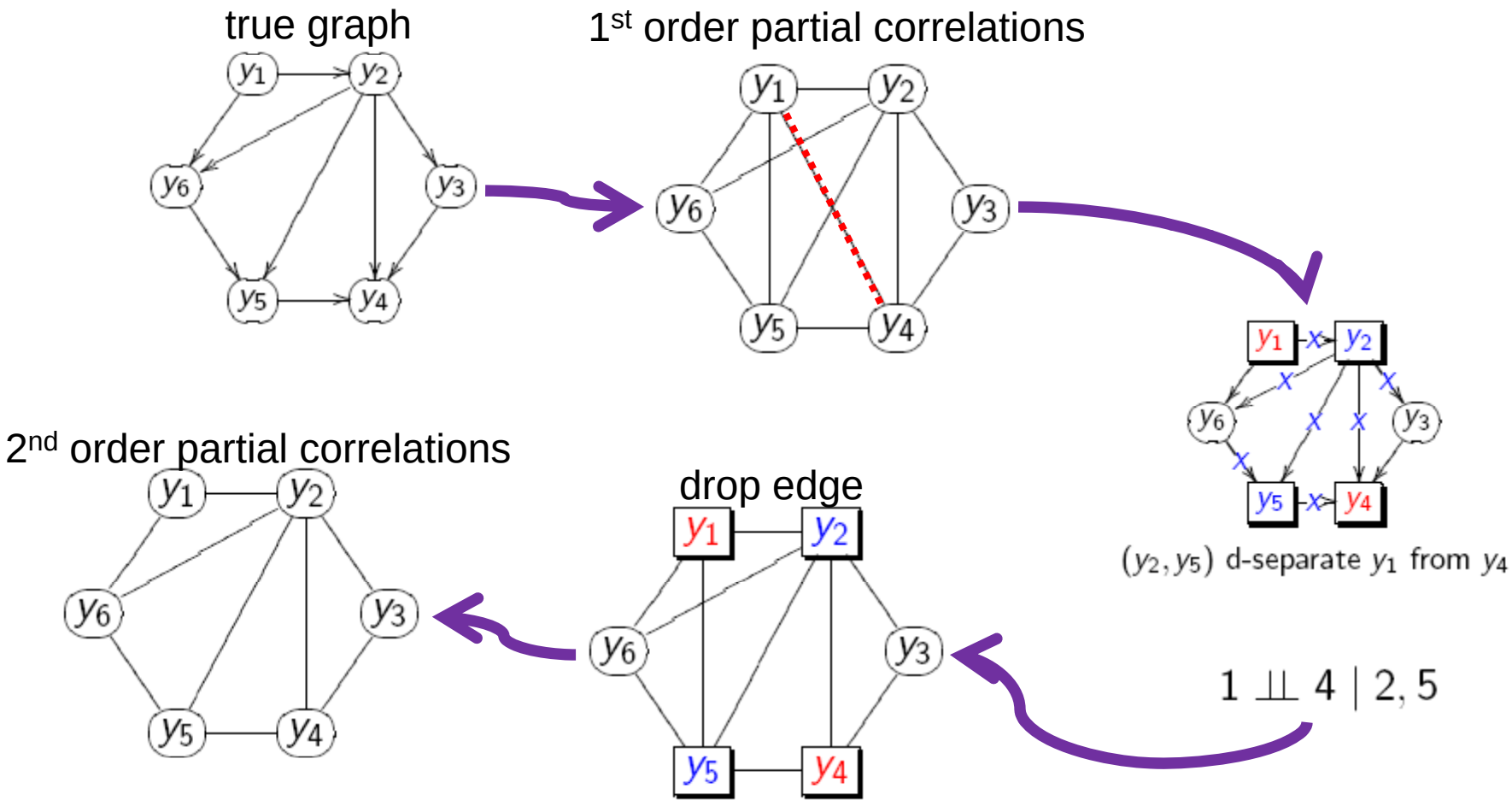
QTL-driven directed graphs

- 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

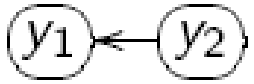


partial correlation (PC) skeleton



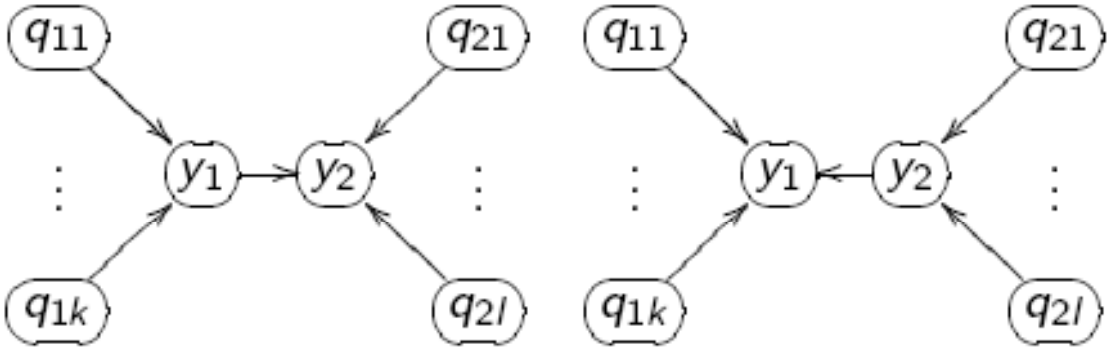
edge direction: which is causal?

M_1 : 

$M2$: 

the above models are likelihood equivalent,

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

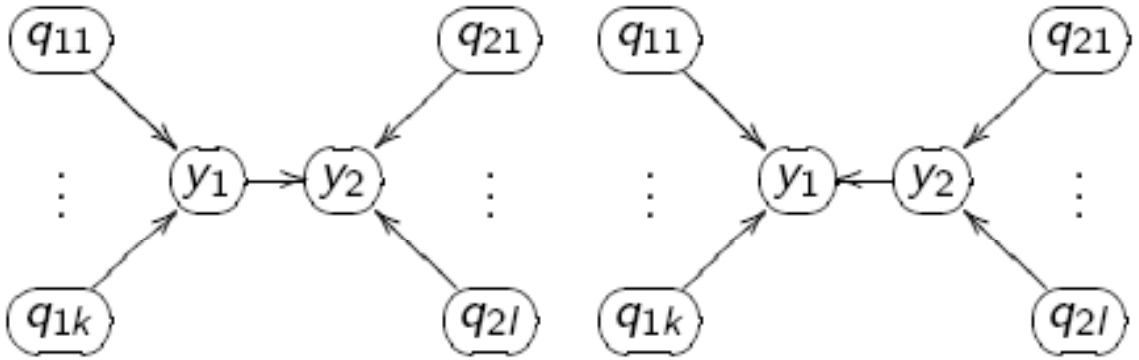


not likelihood equivalent due to QTL

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

test edge direction using LOD score

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

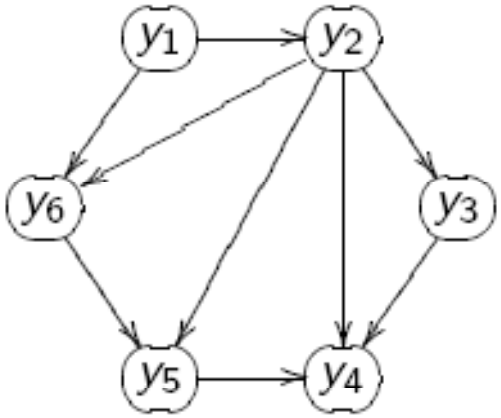


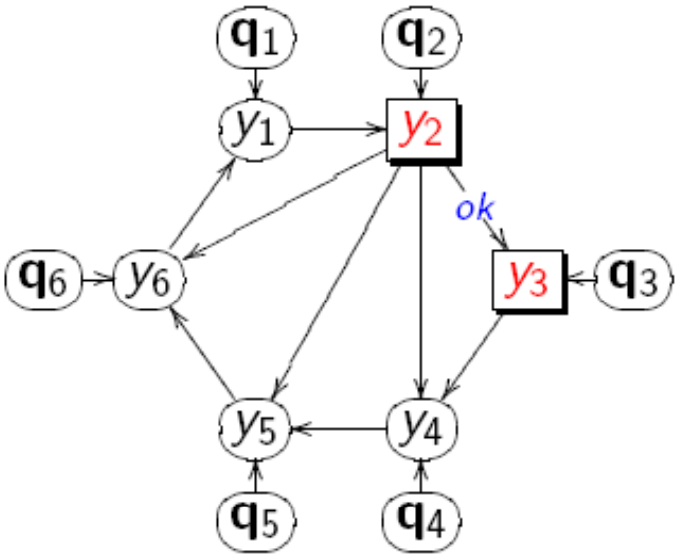
not likelihood equivalent because

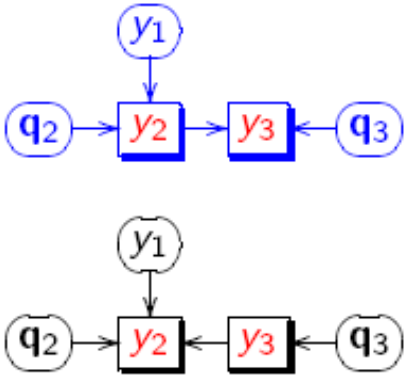
$$f(\mathbf{q}_1) f(y_1 \mid \mathbf{q}_1) f(y_2 \mid y_1, \mathbf{q}_2) f(\mathbf{q}_2)$$
$$\neq$$
$$f(\mathbf{q}_2) f(y_2 \mid \mathbf{q}_2) f(y_1 \mid y_2, \mathbf{q}_1) f(\mathbf{q}_1)$$

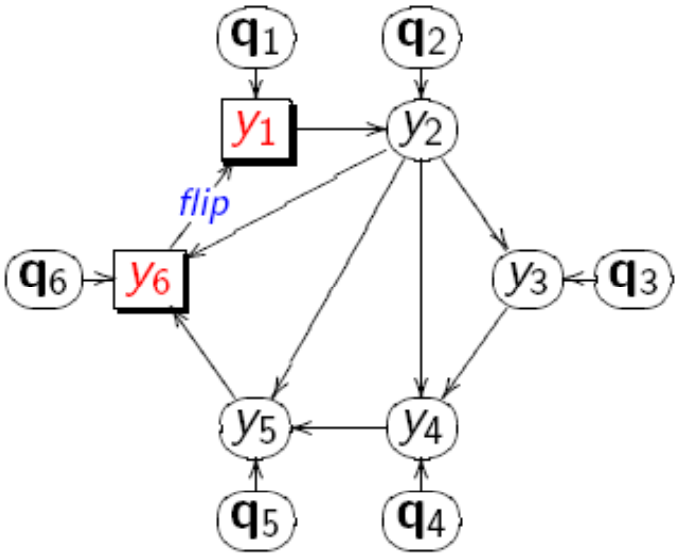
reverse edges using QTLs

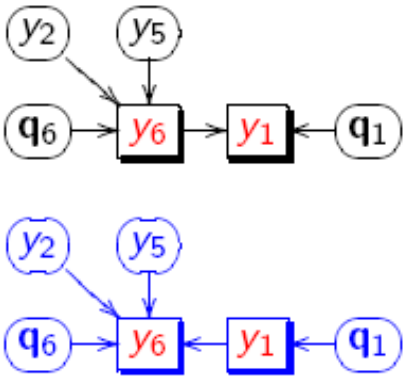
true graph

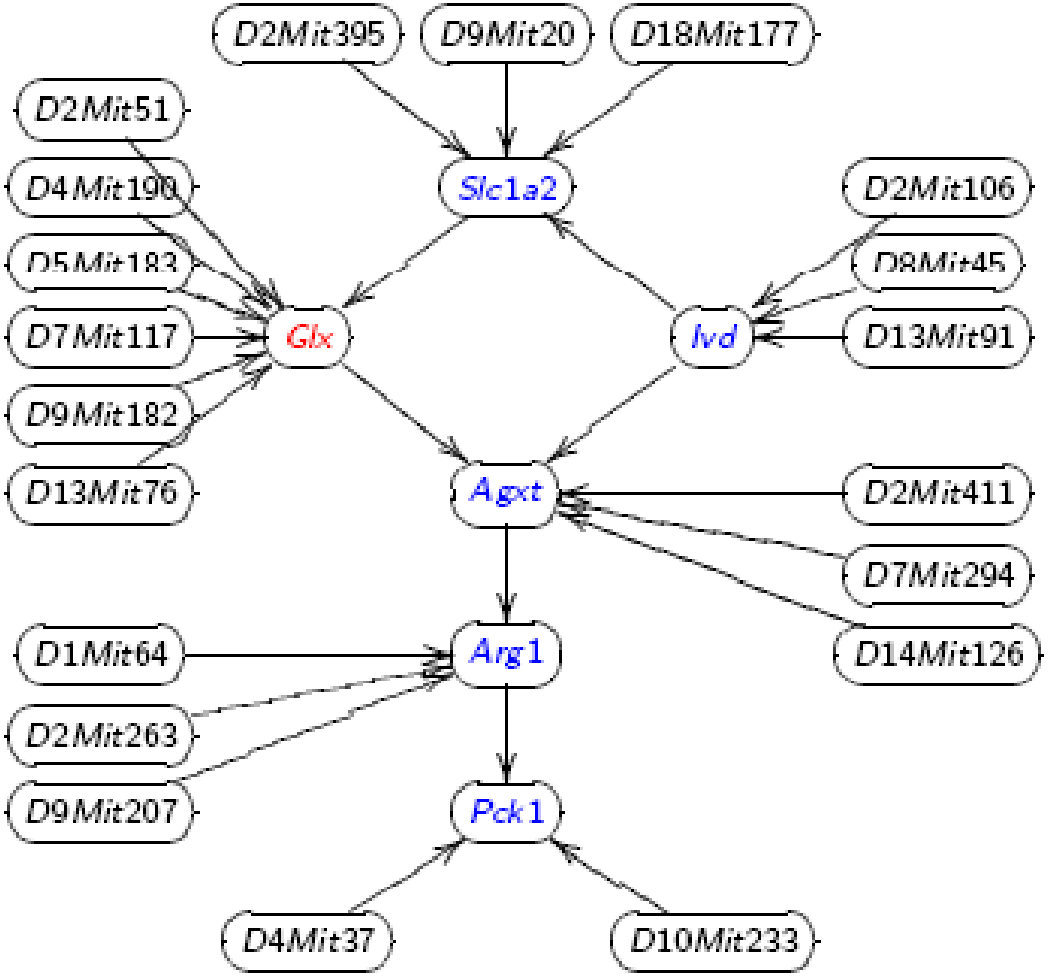












- We constructed a network from metabolites and transcripts involved in liver metabolism.
- We validated this network with *in vitro experiments* (Ferrara et al 2008). Four out of six predictions were confirmed.

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

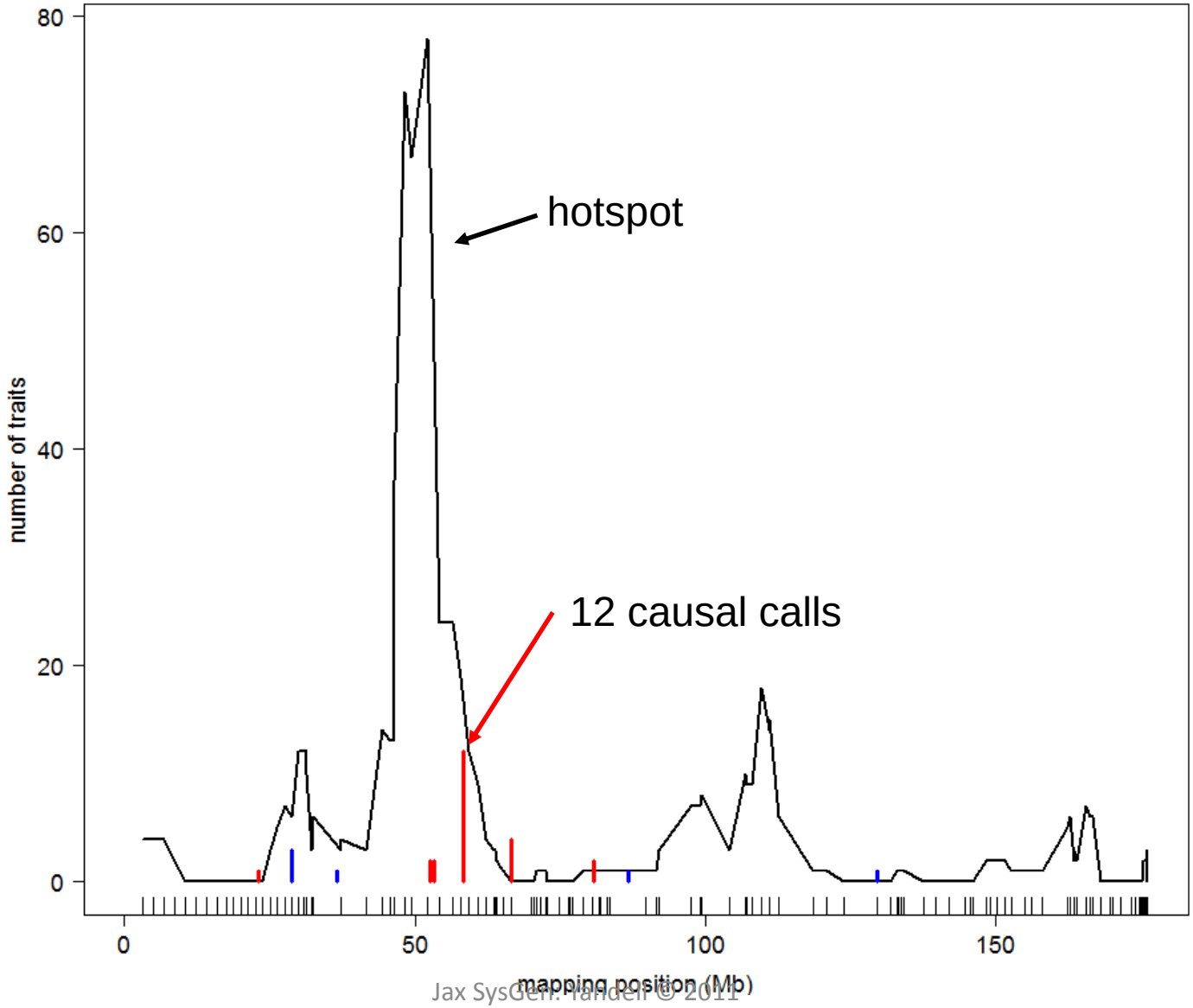
missing data method: MCMC

- known phenotypes Y, genotypes Q
- unknown graph G
- want to study $\Pr(Y \mid G, Q)$
- 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

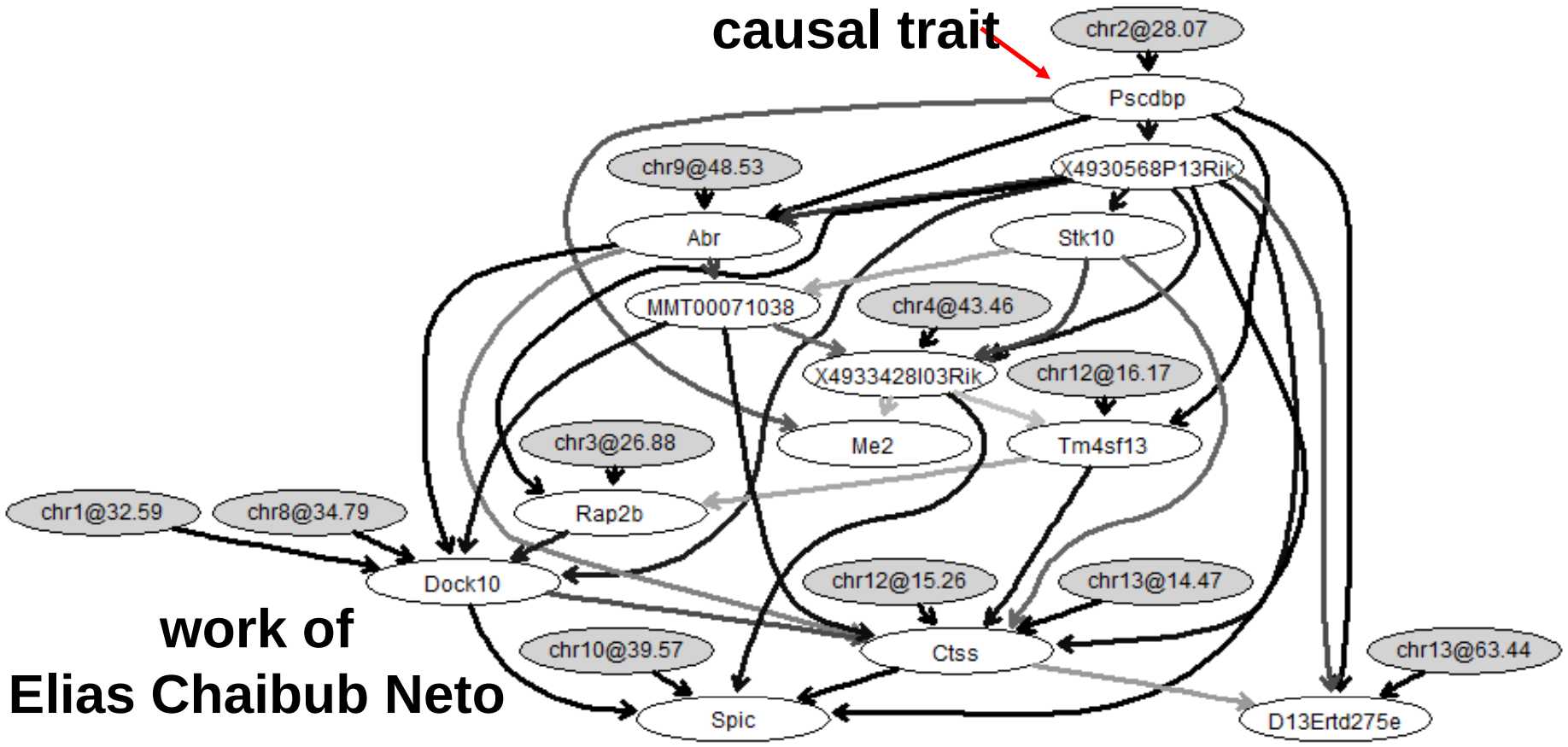
MCMC steps for QTLnet

- 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-/- chr 2: causal architecture



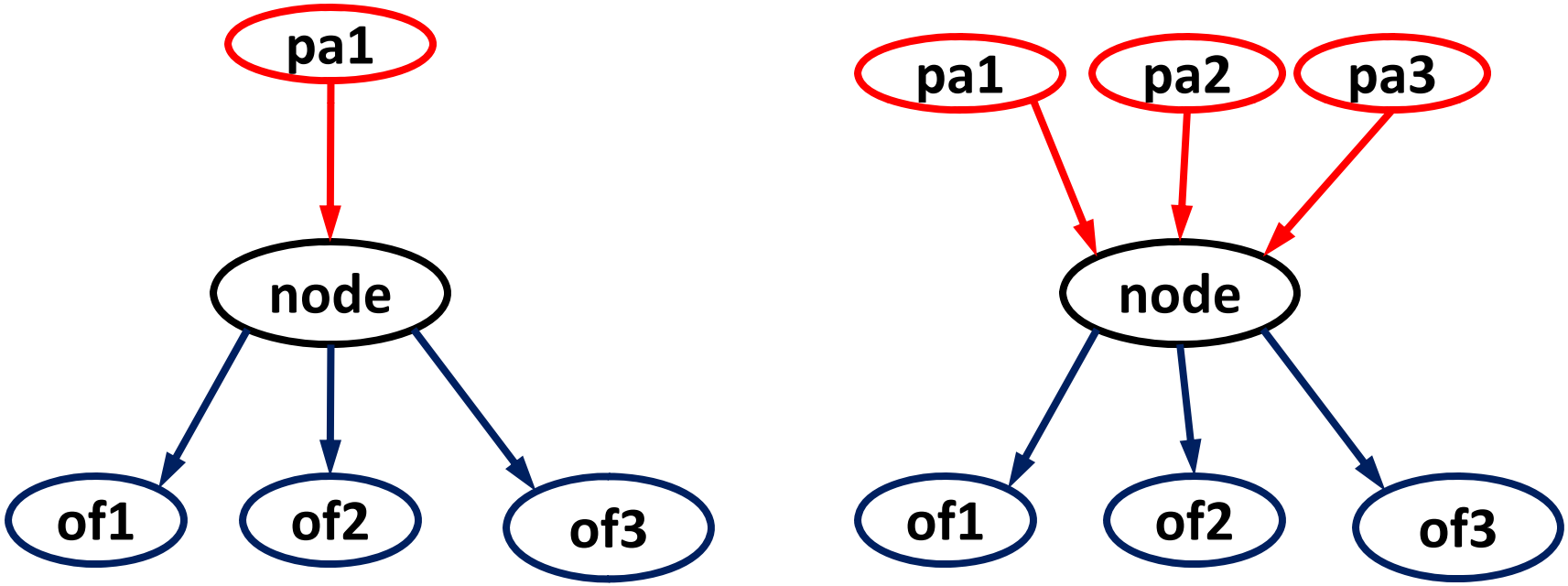
BxH ApoE-/- causal network for transcription factor Pscdbp



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

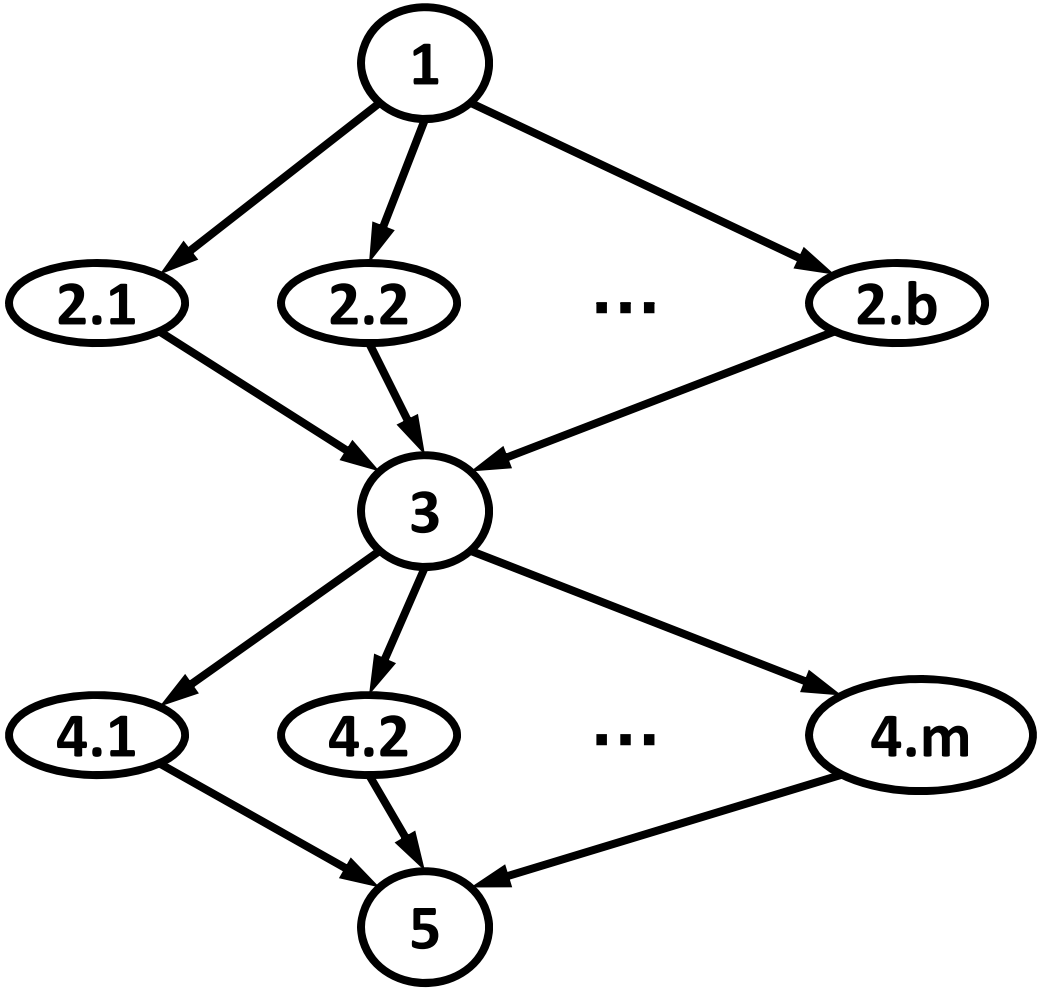
Phase 1: identify parents

Phase 2: compute BICs

Phase 3: store BICs

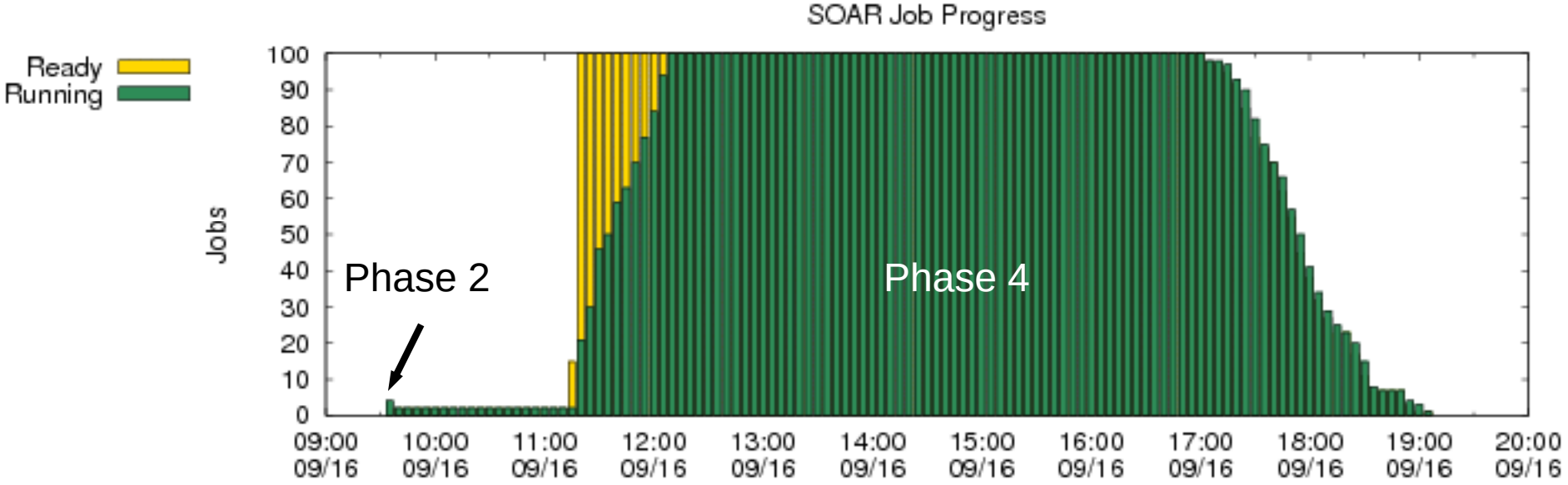
Phase 4: run Markov chains

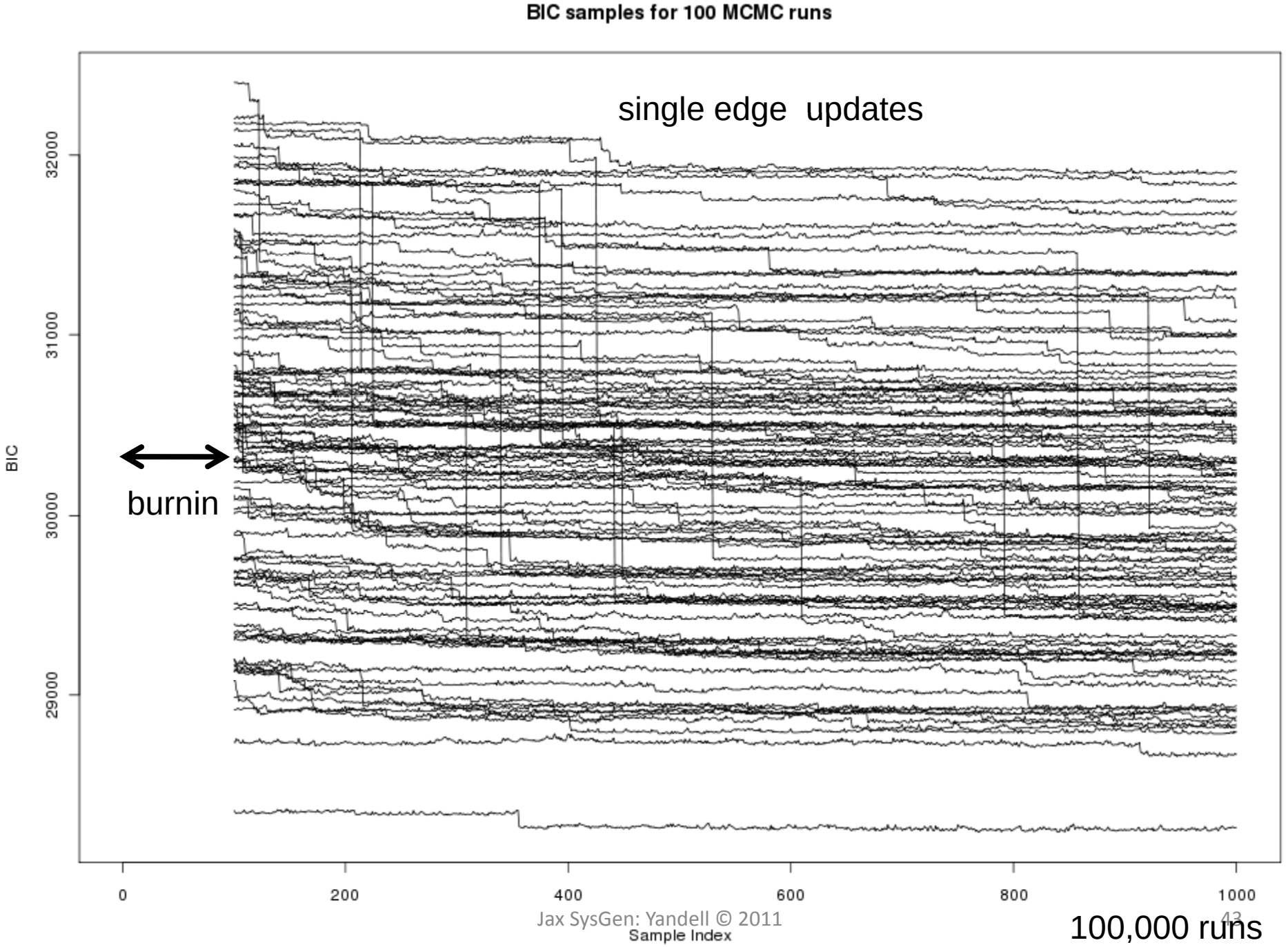
Phase 5: combine results



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

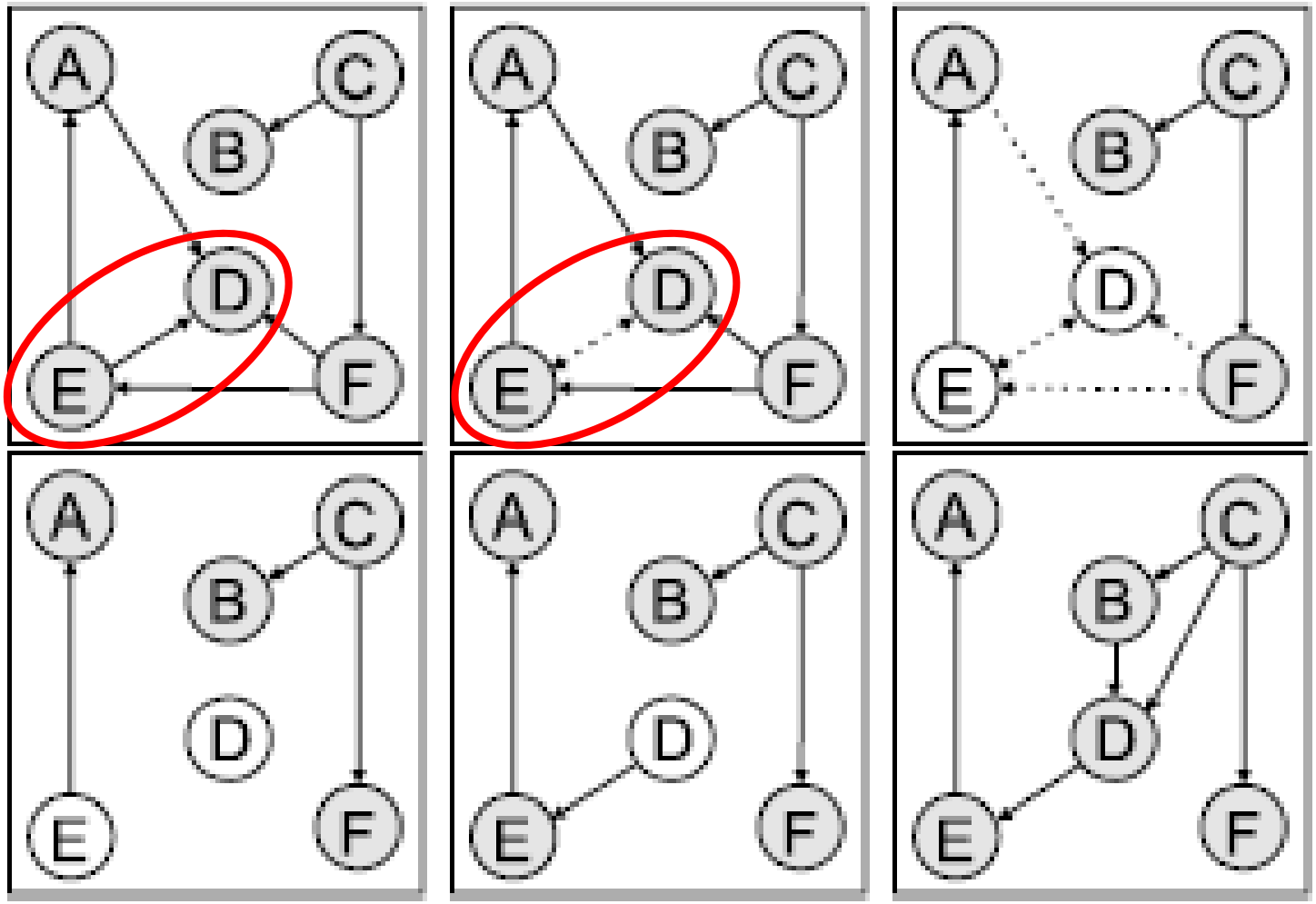



BIC samples for 100 MCMC runs
single edge updates
burnin
BIC
32000
31000
30000
29000
0
200
400
600
800
1000
Sample Index
100,000 runs

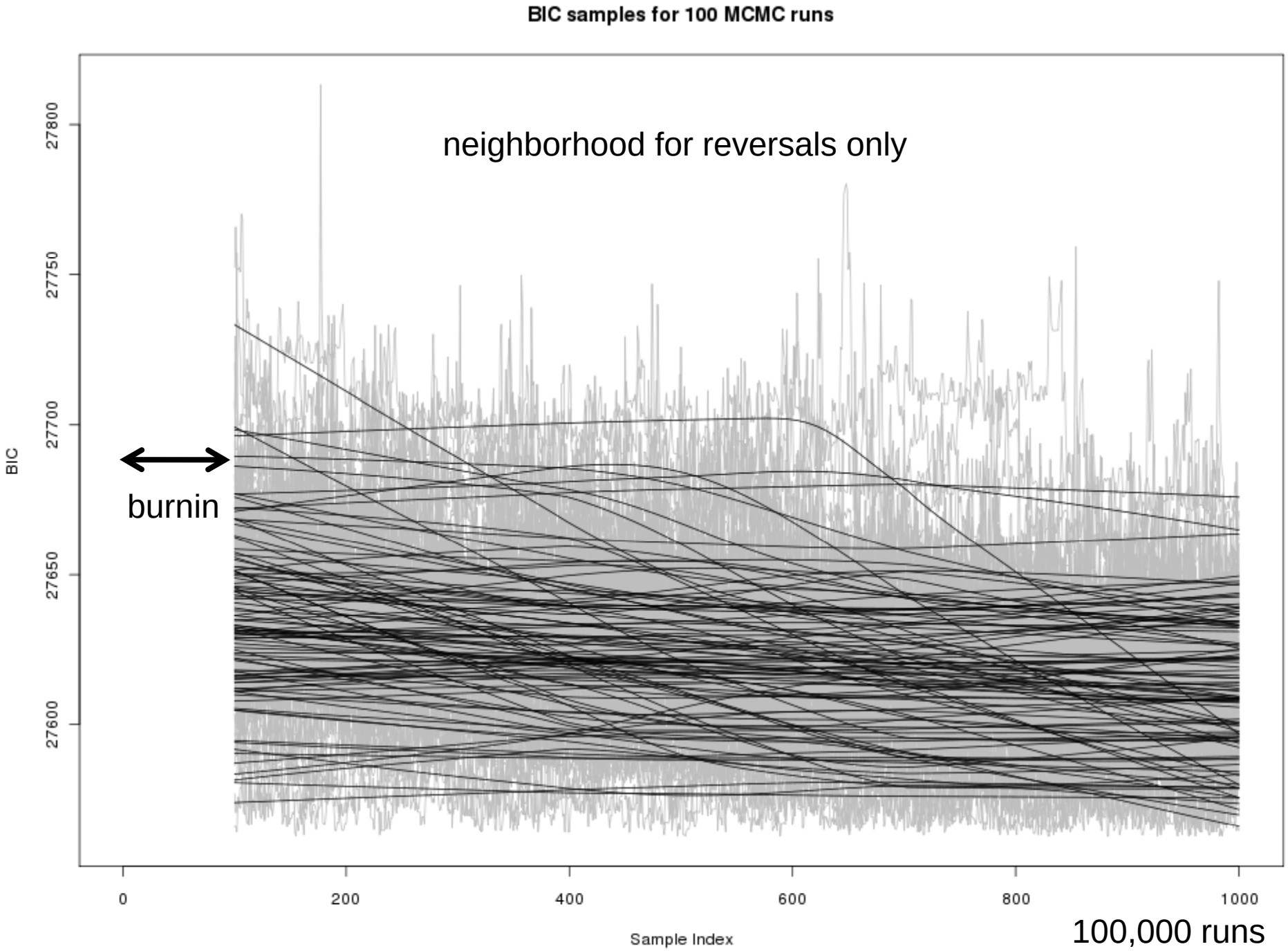
neighborhood edge reversal

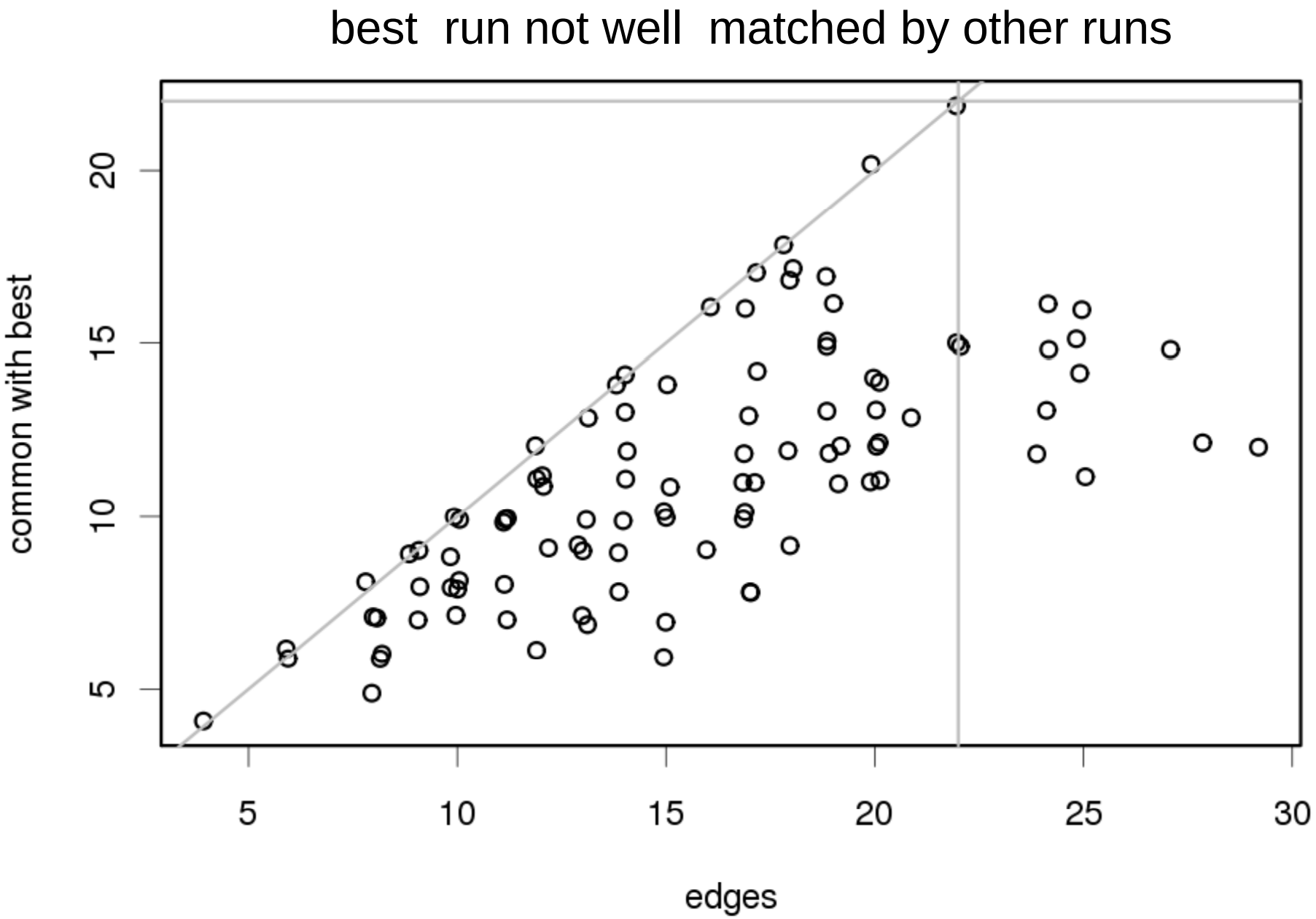
select edge
drop edge
identify parents

orphan nodes
reverse edge
find new parents



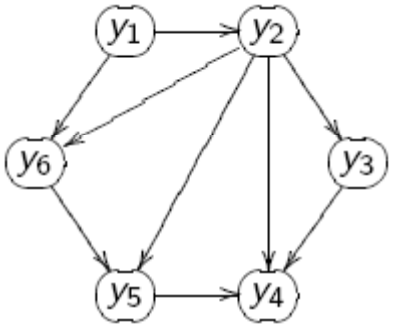
Grzegorczyk M. and Husmeier D. (2008) *Machine Learning* 71 (2-3), 265-305.

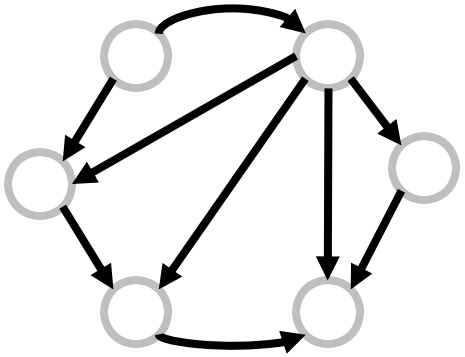

BIC samples for 100 MCMC runs
neighborhood for reversals only
burnin
BIC
27800
27750
27700
27650
27600
0
200
400
600
800
1000
Sample Index
100,000 runs


best run not well matched by other runs
common with best
edges
5
10
15
20
25
30

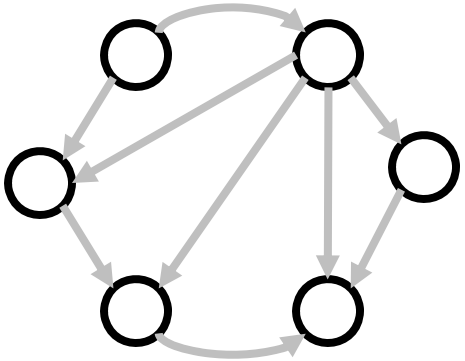
new update scheme MCMC proposals

1. decide to update edge (2) or node (3)

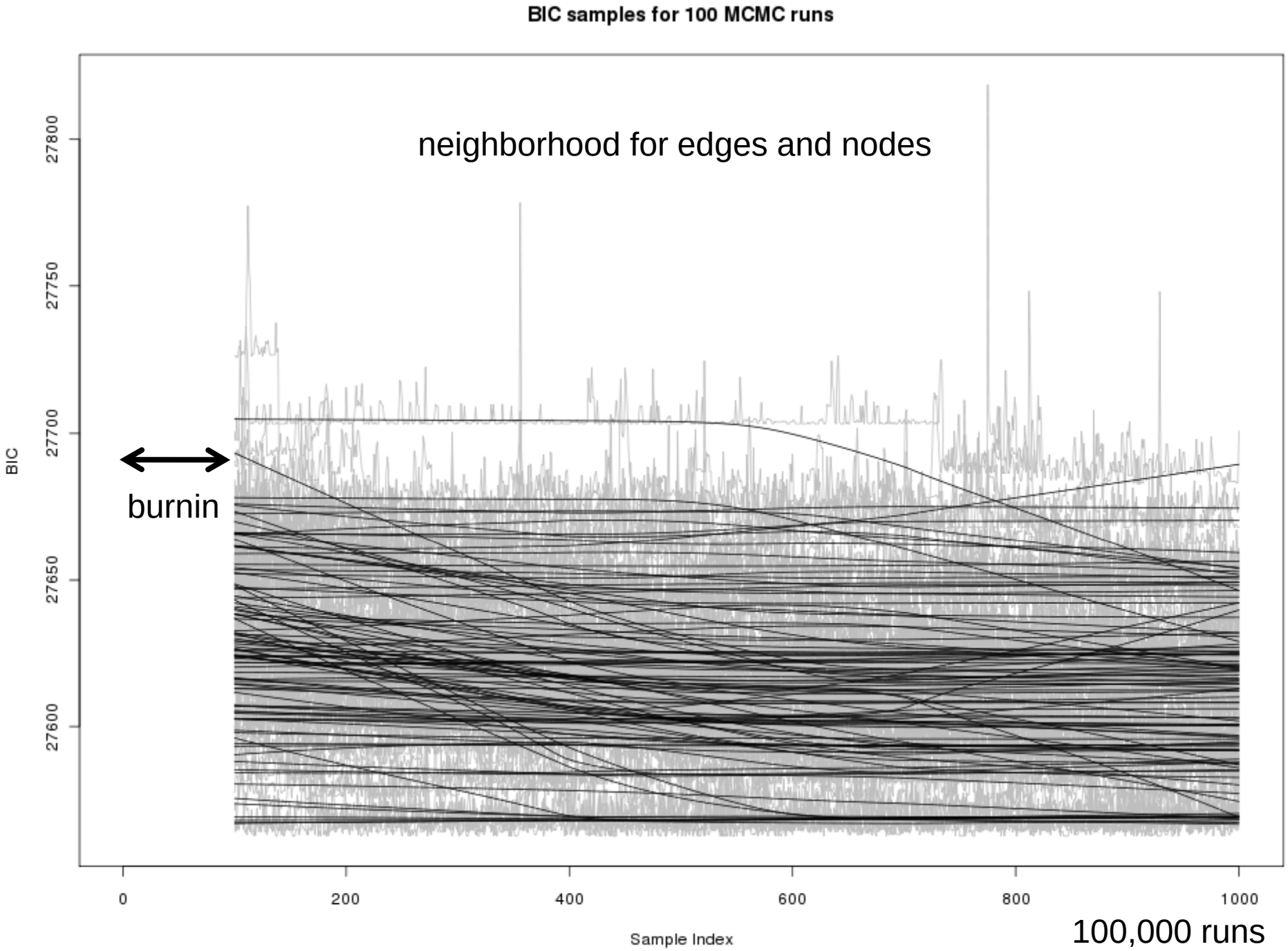




2. pick edge at random
drop or reverse edge
update node parents



3. pick node at random
keep or drop offspring edges
update node parents


BIC samples for 100 MCMC runs
neighborhood for edges and nodes
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* 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
- 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*

future work

- improve algorithm efficiency
  - Ramp up to 100s of phenotypes
- develop visual diagnostics to explore estimates
- incorporate latent variables
  - Aten et al. Horvath (2008 *BMC Sys Biol*)
- extend to outbred crosses, humans