Seattle Summer Institute 2012

15: Systems Genetics for Experimental Crosses

Brian S. Yandell, UW-Madison

Elias Chaibub Neto, Sage Bionetworks

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

Real knowledge is to know the extent of one's ignorance.
Confucius (on a bench in Seattle)

Daily Schedule

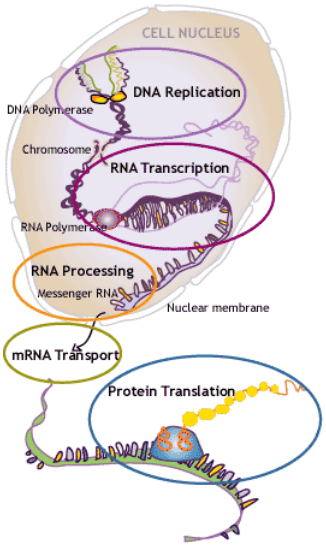
Monday		
8:30-10	Introductions; Overview of System Genetics	1-50
10:30-12	QTL Model Selection	51-100
1:30-3	Gene Mapping for Multiple Correlated Traits	101-150
3:30-5	Hands On Lab: R/qtl	151-200
Tuesday		
8:30-10	Permutation Tests for Correlated Traits	201-250
10:30-12	Scanning the Genome for Causal Architecture	251-300
1:30-3	Causal Phenotype Models Driven by QTL	301-350
3:30-5	Hands On Lab: R/qtlhot, R/qtlnet	351-400
Wednesday		
8:30-10	Incorporating Biological Knowledge	401-450
10:30-12	Platforms for eQTL Analysis	451-500

Overview of Systems Genetics

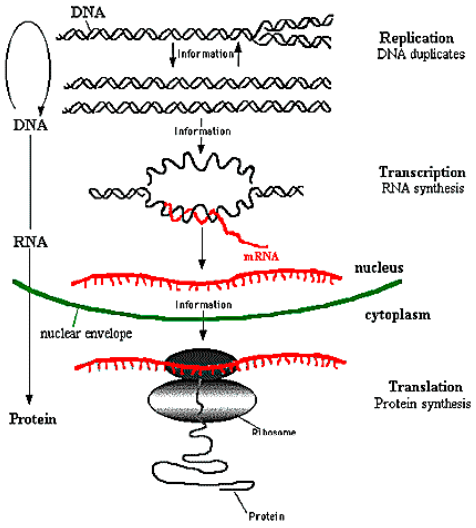
- Big idea: how do genes affect organisms?
- Measuring system(s) state(s) of an organism
- QTL mapping as tool toward goal
- Making sense of multiple traits
- Connecting traits to biochemical pathways
- Putting it all together: workflows

How do genes affect organisms?

- Dogma (with exceptions)
  - DNA -> RNA -> protein -> phenotype
  - redundancy/overlap of biochemical pathways
- System state of organism
  - accumulated effects over time of many genes
  - environmental influences



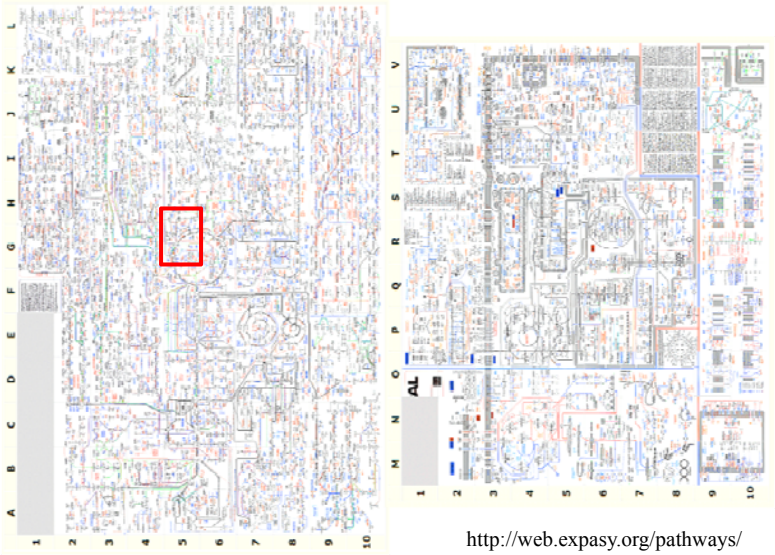
www.nobelprize.org/educational/medicine/dna



The Central Dogma of Molecular Biology

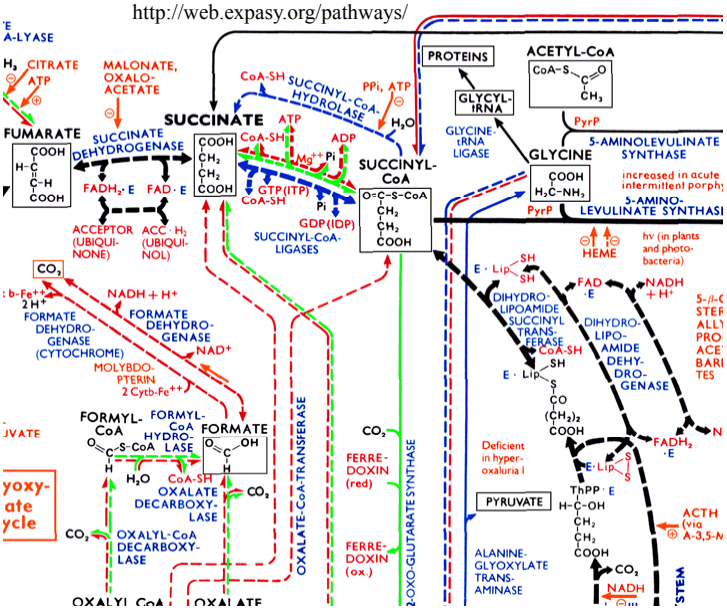
www.accessexcellence.org/RC/VL/GG/central.php

Biochemical Pathways chart, Gerhard Michal, Beohringer Mannheim



http://web.expasy.org/pathways/

http://web.expasy.org/pathways/



systems genetics approach

- study genetic architecture of quantitative traits
  - in model systems, and ultimately humans
- interrogate single resource population for variation
  - DNA sequence, transcript abundance, proteins, metabolites
  - multiple organismal phenotypes
  - multiple environments
- detailed map of genetic variants associated with
  - each organismal phenotype in each environment
- functional context to interpret phenotypes
  - genetic underpinnings of multiple phenotypes
  - genetic basis of genotype by environment interaction

Sieberts, Schadt (2007 *Mamm Genome*); Emilsson et al. (2008 *Nature*) Chen et al. 2008 *Nature*); Ayroles et al. MacKay (2009 *Nature Genetics*)

Measuring an organism

- Phenotype measurement is challenging!
- Cannot measure exactly what is important
- Instead measure multiple related traits
- Multiple traits at one time
- Same trait measured over time

QTL as tool toward goal

- Identifying important genomic region(s)
- But they may contain many genes
- Journey from QTL to gene
  - References…
- Corroborative evidence from multiple traits
  - Reassurance
  - Increased power?
  - Evidence at a system level (pathways, etc.)?

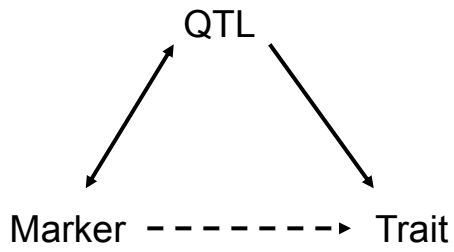
cross two inbred lines

→ linkage disequilibrium

→ associations

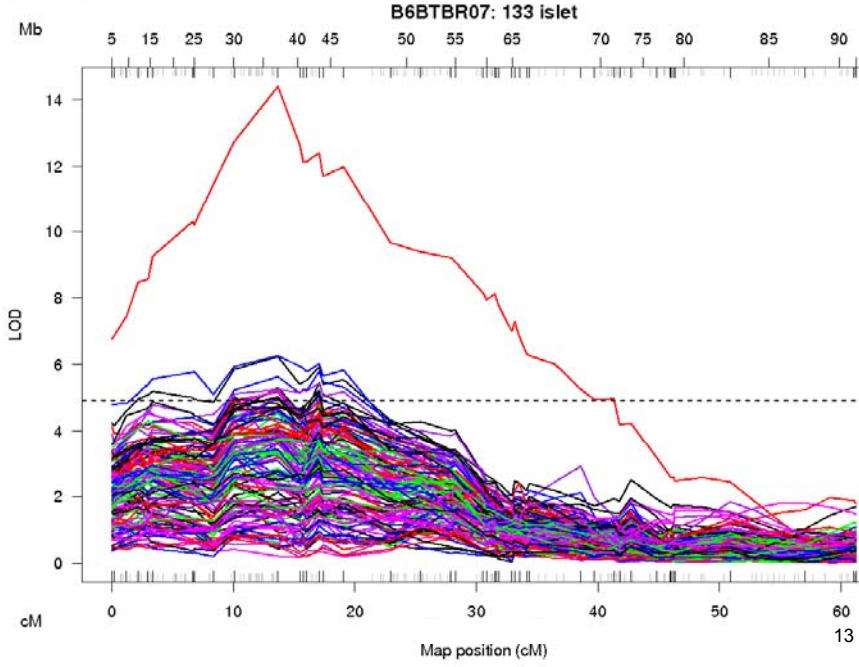
→ linked segregating QTL

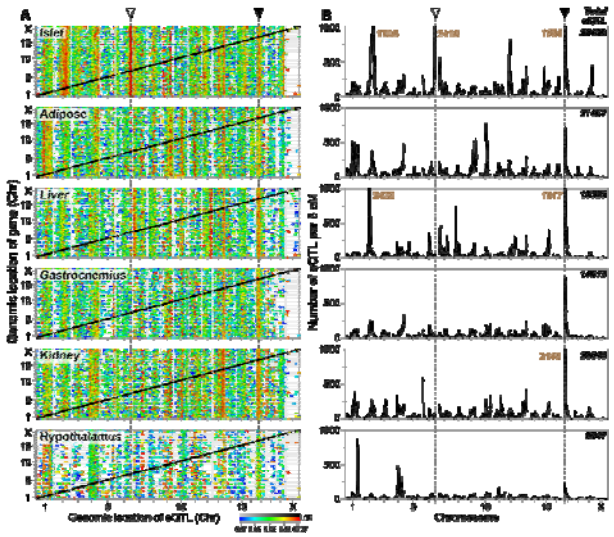
(after Gary Churchill)



Making sense of multiple traits

- Aligning QTL mapping results
- Mapping correlated traits
- Inferring hot spots where many traits map
- Organizing traits into correlated sets
  - Function, clustering, QTL alignment
- Inferring (causal) networks





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).*

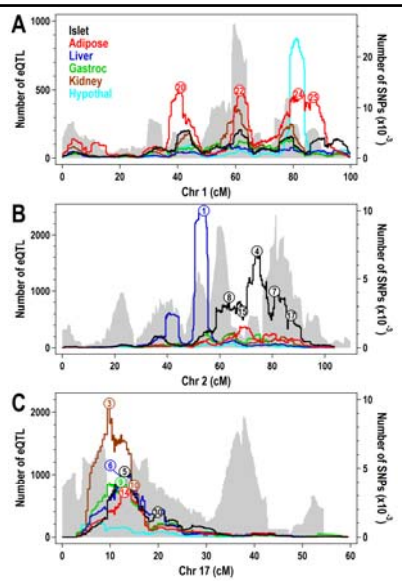
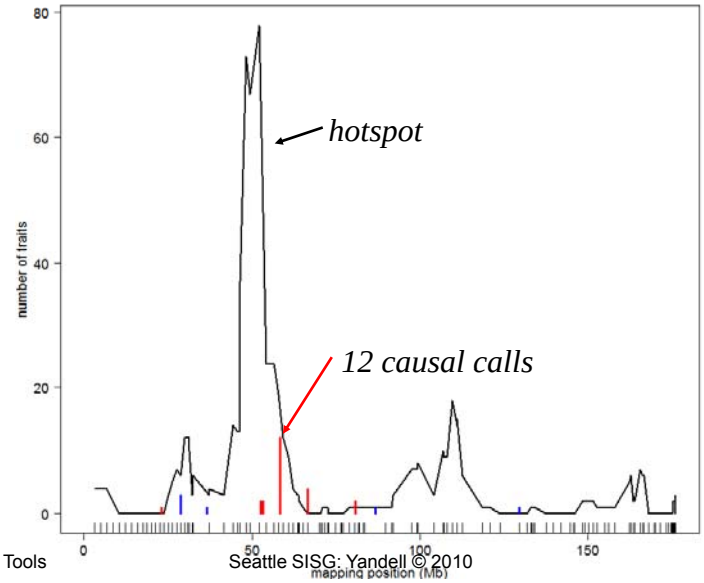


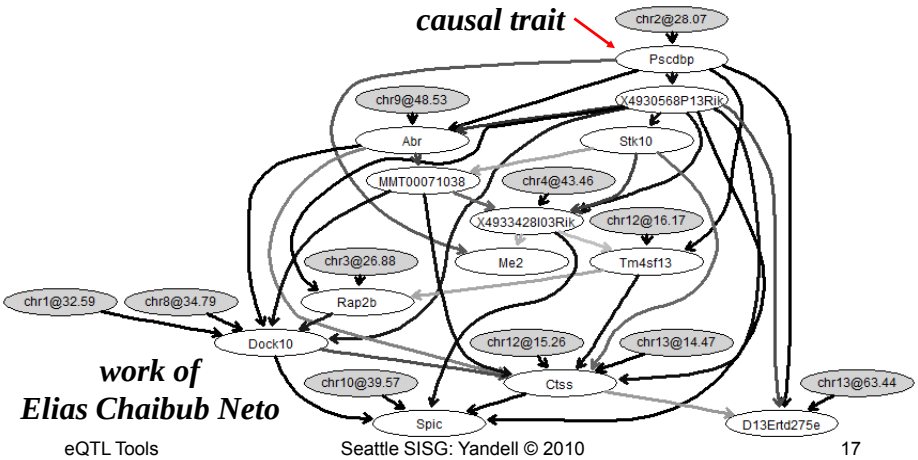
Figure 4 *Tissue-specific hotspots with eQTL and SNP architecture for Chrs 1, 2 and 17.*

*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.*

BxH ApoE-/- chr 2: causal architecture



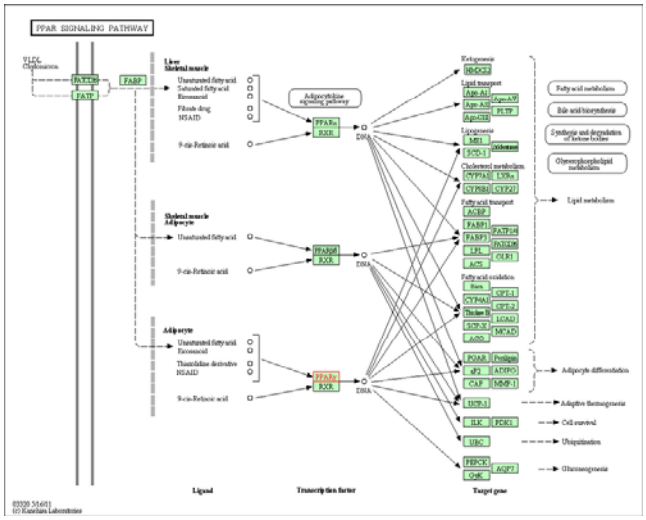
BxH ApoE-/- causal network for transcription factor Pscdbp



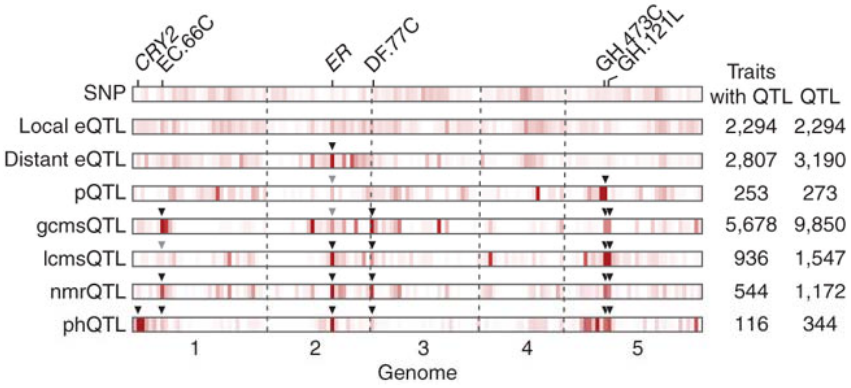
Connecting to biochemical pathways

- Gene ontology (GO)
  - Functional groups
  - Gene enrichment tests
- KO, PPI, TF, interactome databases
  - Networks built from databases
  - Hybrid networks using QTL and databases
- Proof of concept experiments
  - Do findings apply to your organisms?

KEGG pathway: pparg in mouse



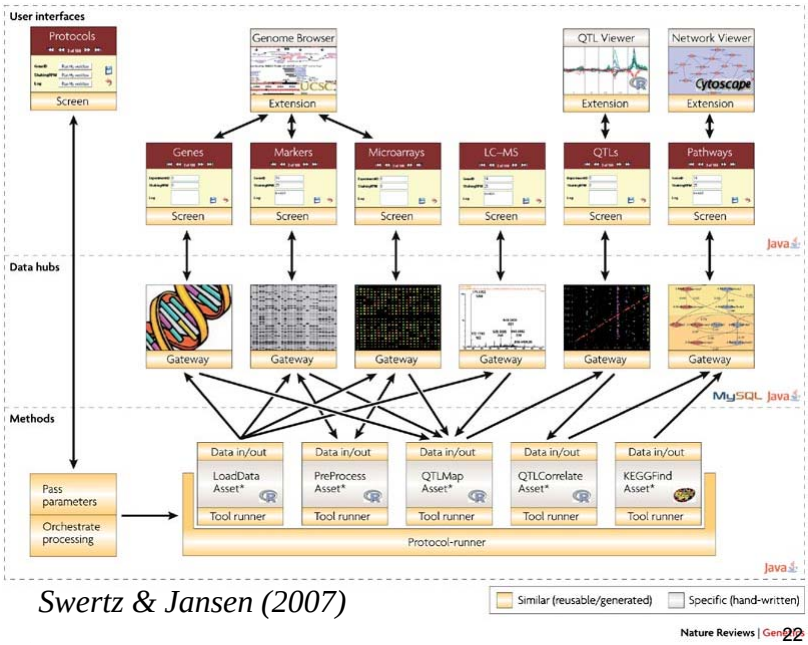
phenotypic buffering of molecular QTL



Fu et al. Jansen (2009 *Nature Genetics*)

Putting it all together: workflows

- Ideally have all tools & data connected
  - Reduce duplication of copies, effort
  - Reduce errors, save time
- Make tools more broadly available
  - User-friendly interfaces
  - Documentation & examples
- Enable comparison of methods
  - Reduce start-up time & translation errors



Swertz & Jansen (2007)

Nature Reviews | Gen

what is the goal of QTL study?

- uncover underlying biochemistry
  - identify how networks function, break down
  - find useful candidates for (medical) intervention
  - epistasis may play key role
  - statistical goal: maximize number of correctly identified QTL
- basic science/evolution
  - how is the genome organized?
  - identify units of natural selection
  - additive effects may be most important (Wright/Fisher debate)
  - statistical goal: maximize number of correctly identified QTL
- select "elite" individuals
  - predict phenotype (breeding value) using suite of characteristics (phenotypes) translated into a few QTL
  - statistical goal: mimimize prediction error

problems of single QTL approach

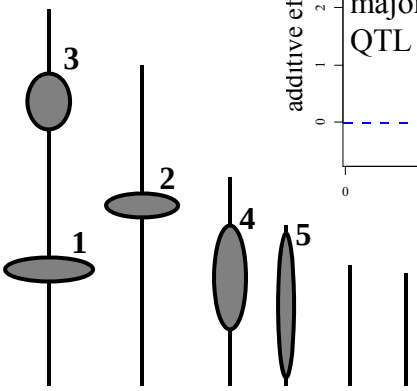
- wrong model: biased view
  - fool yourself: bad guess at locations, effects
  - detect ghost QTL between linked loci
  - miss epistasis completely
- low power
- bad science
  - use best tools for the job
  - maximize scarce research resources
  - leverage already big investment in experiment

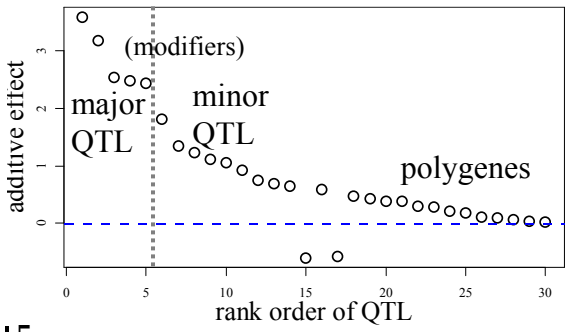
advantages of multiple QTL approach

- improve statistical power, precision
  - increase number of QTL detected
  - better estimates of loci: less bias, smaller intervals
- improve inference of complex genetic architecture
  - patterns and individual elements of epistasis
  - appropriate estimates of means, variances, covariances
    - asymptotically unbiased, efficient
  - assess relative contributions of different QTL
- improve estimates of genotypic values
  - less bias (more accurate) and smaller variance (more precise)
  - mean squared error = MSE = $(\text{bias})^2$ + variance

Pareto diagram of QTL effects

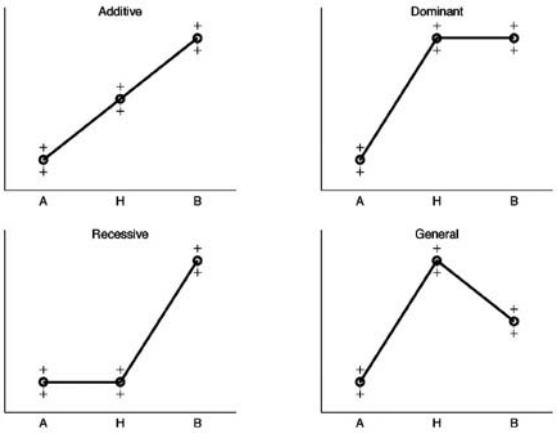
major QTL on linkage map





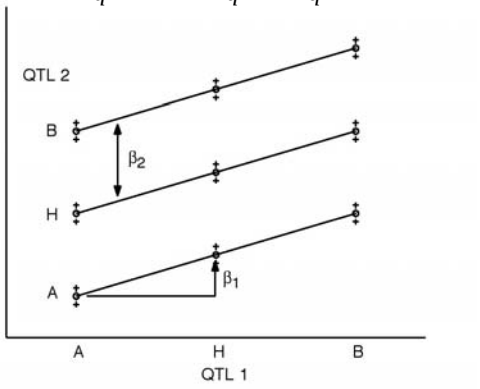
Gene Action and Epistasis

additive, dominant, recessive, general effects of a single QTL (Gary Churchill)



additive effects of two QTL (Gary Churchill)

$$\mu_q = \mu + \beta_{q1} + \beta_{q2}$$



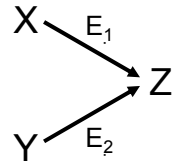
Epistasis (Gary Churchill)

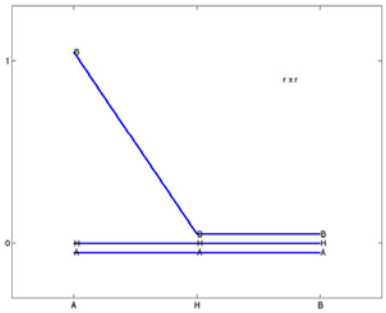
The allelic state at one locus can mask or uncover the effects of allelic variation at another.

- W. Bateson, 1907.

epistasis in parallel pathways (GAC)

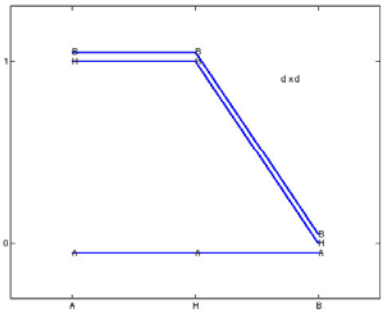
- Z keeps trait value low
- neither E_1 nor E_2 is rate limiting
- loss of function alleles are segregating from parent A at E_1 and from parent B at E_2





epistasis in a serial pathway (GAC)

- Z keeps trait value high
- either E_1 or E_2 is rate limiting
- loss of function alleles are segregating from parent B at E_1 or from parent A at E_2



3. Bayesian vs. classical QTL study

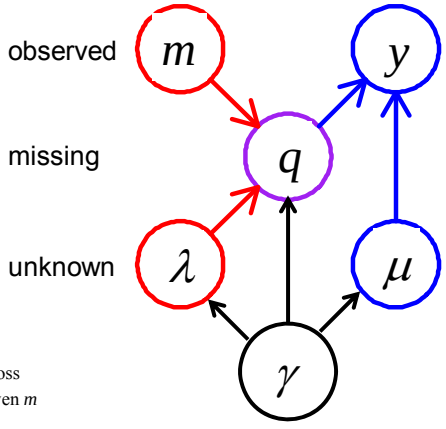
- classical study
  - ***maximize*** over unknown effects
  - ***test*** for detection of QTL at loci
  - model selection in stepwise fashion
- Bayesian study
  - ***average*** over unknown effects
  - ***estimate*** chance of detecting QTL
  - sample all possible models
- both approaches
  - average over missing QTL genotypes
  - scan over possible loci

Bayesian idea

- Reverend Thomas Bayes (1702-1761)
  - part-time mathematician
  - buried in Bunhill Cemetary, Moongate, London
  - famous paper in 1763 *Phil Trans Roy Soc London*
  - was Bayes the first with this idea? (Laplace?)
- basic idea (from Bayes' original example)
  - two billiard balls tossed at random (uniform) on table
  - where is first ball if the second is to its left?
    - prior: anywhere on the table
    - posterior: more likely toward right end of table

QTL model selection: key players

- observed measurements
  - y = phenotypic trait
  - m = markers & linkage map
  - i = individual index (1,…,n)
- missing data
  - missing marker data
  - q = QT genotypes
    - alleles QQ, Qq, or qq at locus
- unknown quantities
  - λ = QT locus (or loci)
  - μ = phenotype model parameters
  - γ = QTL model/genetic architecture
- pr($q|m,\lambda,\gamma$) genotype model
  - grounded by linkage map, experimental cross
  - recombination yields multinomial for q given m
- pr($y|q,\mu,\gamma$) phenotype model
  - distribution shape (assumed normal here)
  - unknown parameters μ (could be non-parametric)



after
Sen Churchill (2001)

Bayes posterior vs. maximum likelihood

- LOD: classical Log ODds
  - maximize likelihood over effects μ
  - `R/qtl scanone/scantwo: method = "em"`
- *LPD*: Bayesian *L*og *P*osterior *D*ensity
  - average posterior over effects μ
  - `R/qtl scanone/scantwo: method = "imp"`

$$\mathrm{LOD}(\lambda) = \log_{10}\{\max_{\mu} \mathrm{pr}(y \mid m, \mu, \lambda)\} + c$$

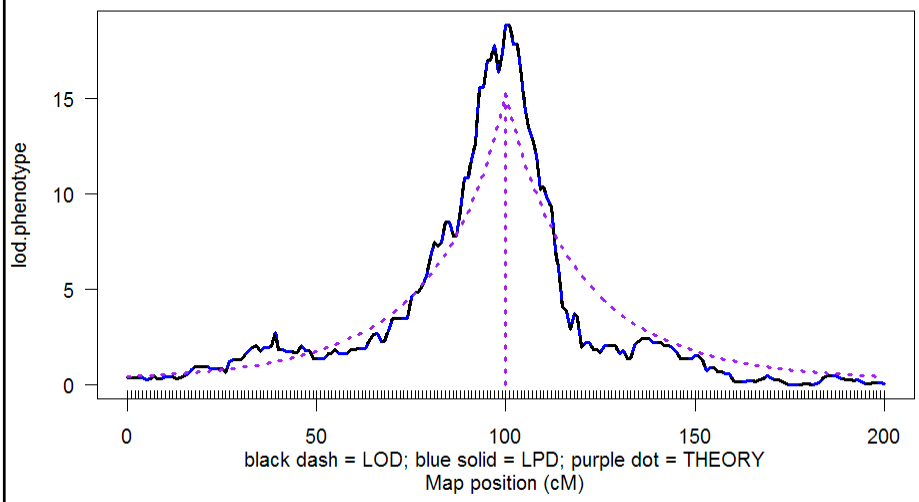
$$\mathrm{LPD}(\lambda) = \log_{10}\{\mathrm{pr}(\lambda \mid m)\int \mathrm{pr}(y \mid m, \mu, \lambda)\mathrm{pr}(\mu)d\mu\} + C$$

likelihood mixes over missing QTL genotypes:

$$\mathrm{pr}(y \mid m, \mu, \lambda) = \sum_q \mathrm{pr}(y \mid q, \mu)\mathrm{pr}(q \mid m, \lambda)$$

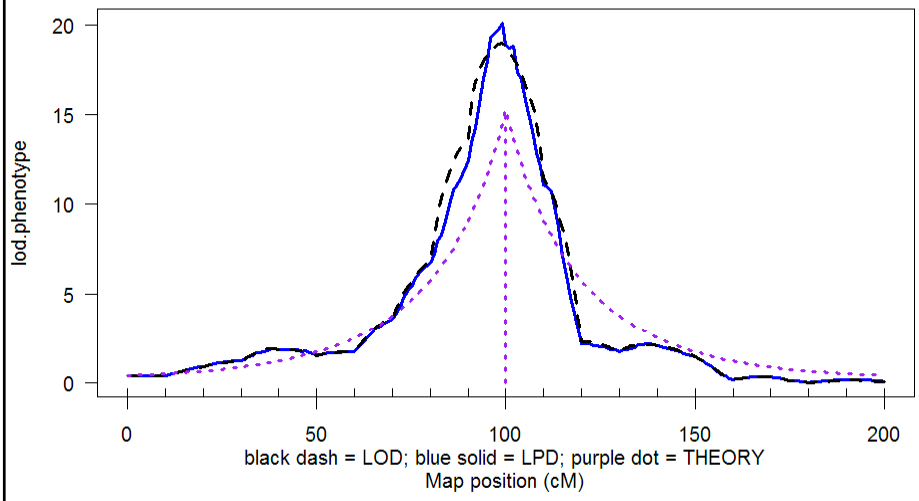
LOD & LPD: 1 QTL

n.ind = 100, 1 cM marker spacing



LOD & LPD: 1 QTL

n.ind = 100, 10 cM marker spacing



marginal LOD or LPD

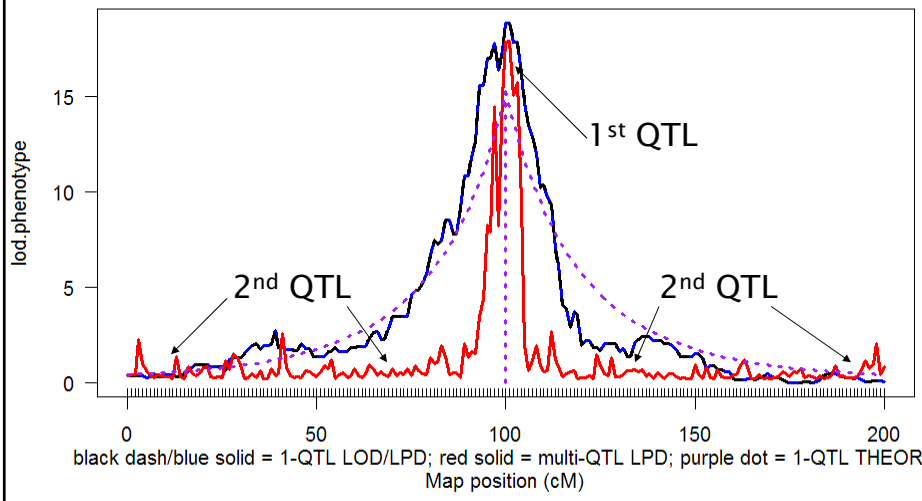
- compare two genetic architectures (γ_2,γ_1) at each locus
  - with (γ_2) or without (γ_1) another QTL at locus λ
    - preserve model hierarchy (e.g. drop any epistasis with QTL at λ)
  - with (γ_2) or without (γ_1) epistasis with QTL at locus λ
  - γ_2 contains γ_1 as a sub-architecture
- allow for multiple QTL besides locus being scanned
  - architectures γ_1 and γ_2 may have QTL at several other loci
  - use marginal LOD, LPD or other diagnostic
  - posterior, Bayes factor, heritability

$$\mathrm{LOD}(\lambda \mid \gamma_2) - \mathrm{LOD}(\lambda \mid \gamma_1)$$

$$\mathrm{LPD}(\lambda \mid \gamma_2) - \mathrm{LPD}(\lambda \mid \gamma_1)$$

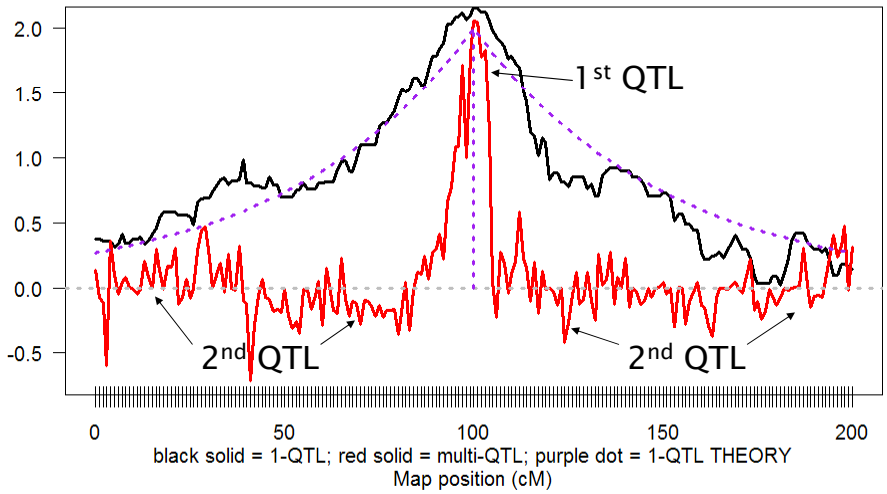
LPD: 1 QTL vs. multi-QTL

marginal contribution to LPD from QTL at λ



substitution effect: 1 QTL vs. multi-QTL

single QTL effect vs. marginal effect from QTL at λ



why use a Bayesian approach?

- first, do *both* classical and Bayesian
  - always nice to have a separate validation
  - each approach has its strengths and weaknesses
- classical approach works quite well
  - selects large effect QTL easily
  - directly builds on regression ideas for model selection
- Bayesian approach is comprehensive
  - samples most probable genetic architectures
  - formalizes model selection within one framework
  - readily (!) extends to more complicated problems

comparing models

- balance model fit against model complexity
  - want to fit data well (maximum likelihood)
  - without getting too complicated a model

	smaller model	**bigger model**
fit model	miss key features	fits better
estimate phenotype	may be biased	no bias
predict new data	may be biased	no bias
interpret model	easier	more complicated
estimate effects	low variance	high variance

QTL software options

- methods
  - approximate QTL by markers
  - exact multiple QTL interval mapping
- software platforms
  - MapMaker/QTL (obsolete)
  - QTLCart (statgen.ncsu.edu/qtlcart**)**
  - R/qtl (www.rqtl.org)
  - R/qtlbim (www.qtlbim.org)
  - Yandell, Bradbury (2007) book chapter

QTL software platforms

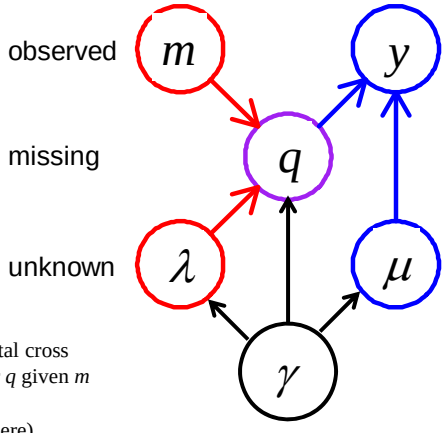
- QTLCart (statgen.ncsu.edu/qtlcart**)**
  - includes features of original MapMaker/QTL
    - not designed for building a linkage map
  - easy to use Windows version WinQTLCart
  - based on Lander-Botstein maximum likelihood LOD
    - extended to marker cofactors (CIM) and multiple QTL (MIM)
    - epistasis, some covariates (GxE)
    - stepwise model selection using information criteria
  - some multiple trait options
  - OK graphics
- R/qtl (www.rqtl.org)
  - includes functionality of classical interval mapping
  - many useful tools to check genotype data, build linkage maps
  - excellent graphics
  - several methods for 1-QTL and 2-QTL mapping
    - epistasis, covariates (GxE)
  - tools available for multiple QTL model selection

QTL Model Selection

1. Bayesian strategy
2. Markov chain sampling
3. sampling genetic architectures
4. criteria for model selection

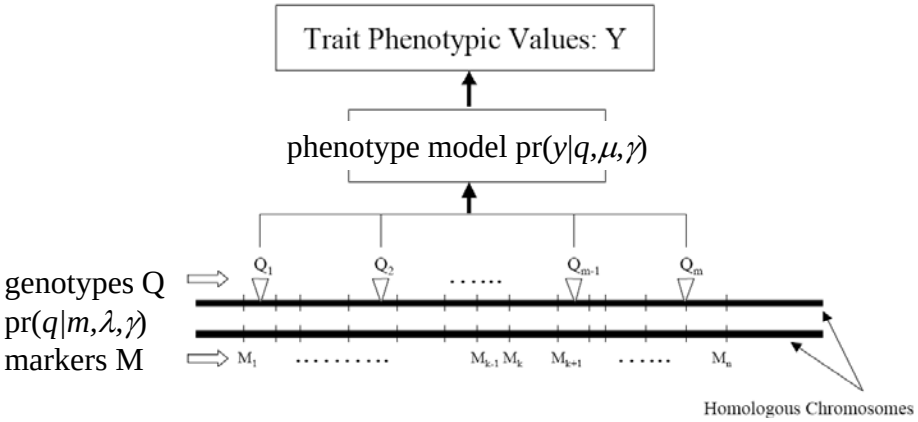
QTL model selection: key players

- observed measurements
  - y = phenotypic trait
  - m = markers & linkage map
  - i = individual index (1,…,n)
- missing data
  - missing marker data
  - q = QT genotypes
    - alleles QQ, Qq, or qq at locus
- unknown quantities
  - λ = QT locus (or loci)
  - μ = phenotype model parameters
  - γ = QTL model/genetic architecture
- pr($q|m,\lambda,\gamma$) genotype model
  - grounded by linkage map, experimental cross
  - recombination yields multinomial for q given m
- pr($y|q,\mu,\gamma$) phenotype model
  - distribution shape (assumed normal here)
  - unknown parameters μ (could be non-parametric)



after
Sen Churchill (2001)

QTL mapping (from ZB Zeng)



classical likelihood approach

- genotype model pr($q|m,\lambda,\gamma$)
  - missing genotypes q depend on observed markers m across genome
- phenotype model pr($y|q,\mu,\gamma$)
  - link phenotypes y to genotypes q

$$\text{LOD}(\lambda) = \log_{10}\{\max_{\mu} \text{pr}(y \mid m, \mu, \lambda)\} + c$$

likelihood mixes over missing QTL genotypes:

$$\text{pr}(y \mid m, \mu, \lambda) = \sum_q \text{pr}(y \mid q, \mu)\text{pr}(q \mid m, \lambda)$$

EM approach

- Iterate E and M steps
  - expectation (E): geno prob's pr($q|m,\lambda,\gamma$)
  - maximization (M): pheno model parameters
    - mean, effects, variance
  - careful attention when many QTL present
    - Multiple papers by Zhao-Bang Zeng and others
  - Start with simple initial model
    - Add QTL, epistatic effects sequentially

classic model search

- initial model from single QTL analysis
- search for additional QTL
- search for epistasis between pairs of QTL
  - Both in model? One in model? Neither?
- Refine model
  - Update QTL positions
  - Check if existing QTL can be dropped
- Analogous to stepwise regression

comparing models (details later)

- balance model fit against model complexity
  - want to fit data well (maximum likelihood)
  - without getting too complicated a model

	smaller model	**bigger model**
fit model	miss key features	fits better
estimate phenotype	may be biased	no bias
predict new data	may be biased	no bias
interpret model	easier	more complicated
estimate effects	low variance	high variance

1. Bayesian strategy for QTL study

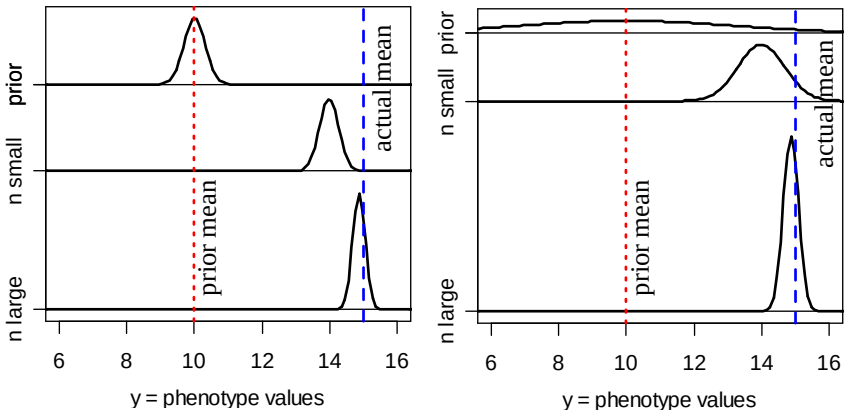
- augment data (y,m) with missing genotypes q
- study unknowns (μ,λ,γ) given augmented data (y,m,q)
  - find better genetic architectures γ
  - find most likely genomic regions = QTL = λ
  - estimate phenotype parameters = genotype means = μ
- sample from posterior in some clever way
  - multiple imputation (Sen Churchill 2002)
  - Markov chain Monte Carlo (MCMC)
    - (Satagopan et al. 1996; Yi et al. 2005, 2007)

$$\text{posterior} = \frac{\text{likelihood} * \text{prior}}{\text{constant}}$$

$$\text{posterior for } q,\mu,\lambda,\gamma = \frac{\text{phenotype likelihood} * [\text{prior for } q,\mu,\lambda,\gamma]}{\text{constant}}$$

$$\text{pr}(q,\mu,\lambda,\gamma \mid y,m) = \frac{\text{pr}(y \mid q,\mu,\gamma) * [\text{pr}(q \mid m,\lambda,\gamma)\text{pr}(\mu \mid \gamma)\text{pr}(\lambda \mid m,\gamma)\text{pr}(\gamma)]}{\text{pr}(y \mid m)}$$

Bayes posterior for normal data

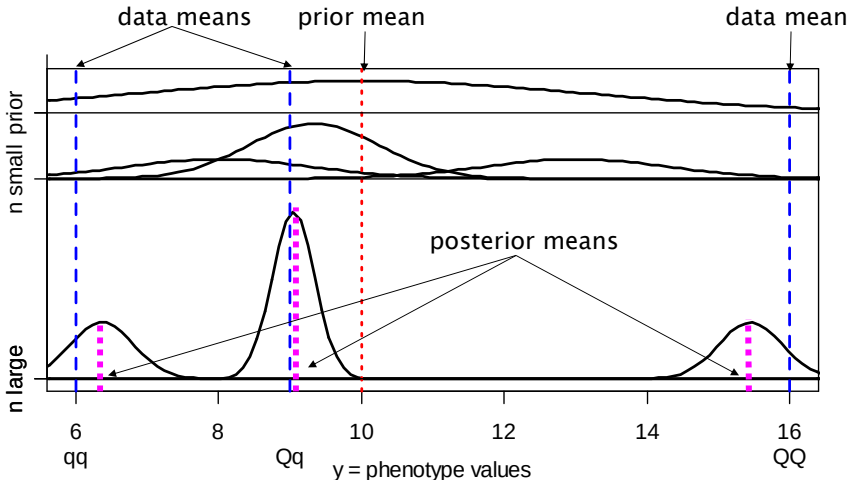


small prior variance

large prior variance

Posterior on genotypic means?

phenotype model pr($y|q,\mu$)



Bayes posterior QTL means

posterior centered on sample genotypic mean
but shrunken slightly toward overall mean

phenotype mean: $E(y \mid q) = \mu_q \qquad V(y \mid q) = \sigma^2$

genotypic prior: $E(\mu_q) = \bar{y}_{\bullet} \qquad V(\mu_q) = \kappa\sigma^2$

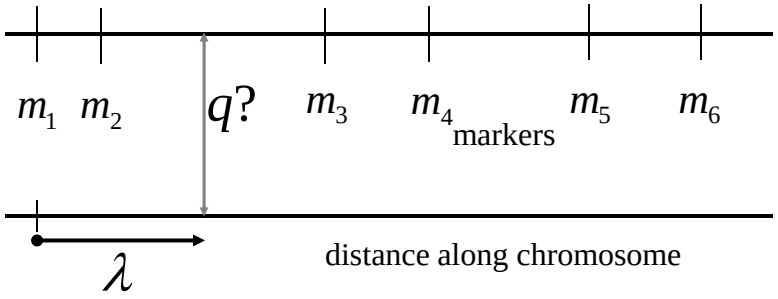
posterior: $E(\mu_q \mid y) = b_q \bar{y}_q + (1 - b_q)\bar{y}_{\bullet} \qquad V(\mu_q \mid y) = b_q \sigma^2 / n_q$

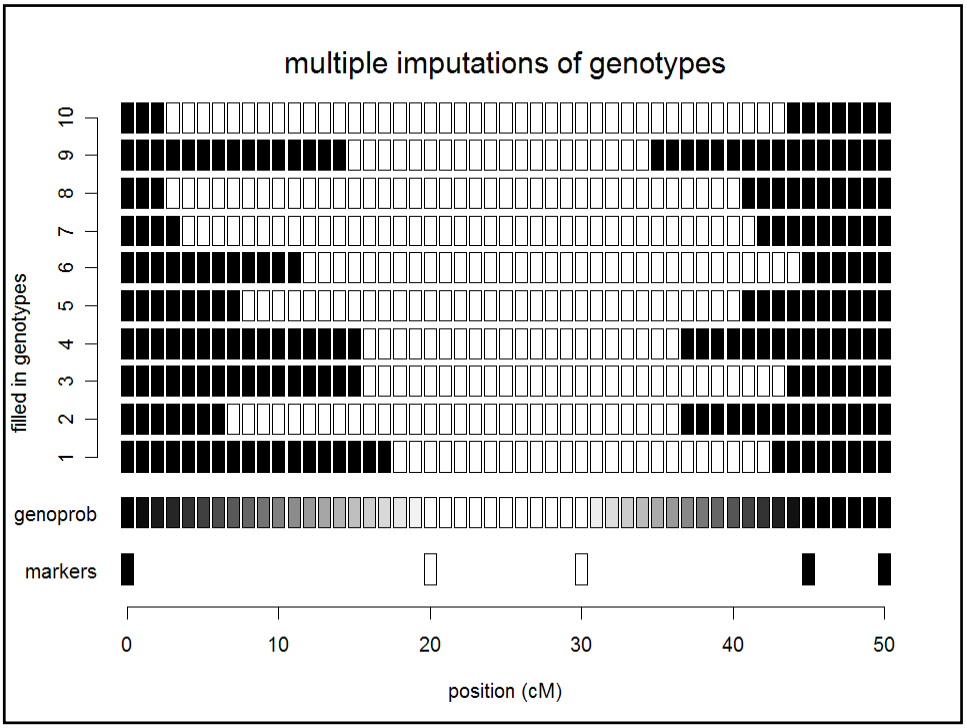
$n_q = \text{count}\{q_i = q\} \qquad \bar{y}_q = \underset{\{q_i = q\}}{\text{sum}} y_i / n_q$

shrinkage: $b_q = \dfrac{\kappa n_q}{\kappa n_q + 1} \to 1$

pr($q|m,\lambda$) recombination model

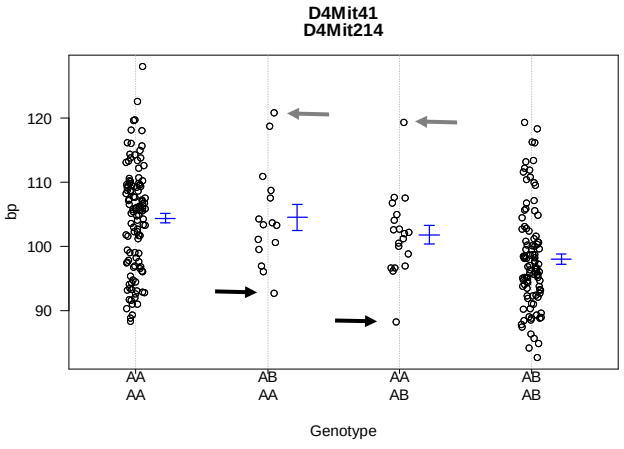
pr($q|m,\lambda$) = pr(geno | map, locus) ≈
pr(geno | flanking markers, locus)





what are likely QTL genotypes *q?*

how does phenotype *y* improve guess?



what are probabilities for genotype *q* between markers?

recombinants AA:AB

all 1:1 if ignore *y* and if we use *y*?

posterior on QTL genotypes q

- full conditional of q given data, parameters
  - proportional to prior pr(q | m, λ)
    - weight toward q that agrees with flanking markers
  - proportional to likelihood pr(y | q, μ)
    - weight toward q with similar phenotype values
  - posterior recombination model balances these two
- this *is* the E-step of EM computations

$$\text{pr}(q \mid y, m, \mu, \lambda) = \frac{\text{pr}(y \mid q, \mu) * \text{pr}(q \mid m, \lambda)}{\text{pr}(y \mid m, \mu, \lambda)}$$

Where are the loci λ on the genome?

- prior over genome for QTL positions
  - flat prior = no prior idea of loci
  - or use prior studies to give more weight to some regions
- posterior depends on QTL genotypes q

  pr(λ | m,q) = pr(λ) pr(q | m,λ) / constant
  - constant determined by averaging
    - over all possible genotypes q
    - over all possible loci λ on entire map
- no easy way to write down posterior

what is the genetic architecture γ?

- which positions correspond to QTLs?
  - priors on loci (previous slide)
- which QTL have main effects?
  - priors for presence/absence of main effects
    - same prior for all QTL
    - can put prior on each d.f. (1 for BC, 2 for F2)
- which pairs of QTL have epistatic interactions?
  - prior for presence/absence of epistatic pairs
    - depends on whether 0,1,2 QTL have main effects
    - epistatic effects less probable than main effects

γ = genetic architecture:

loci:

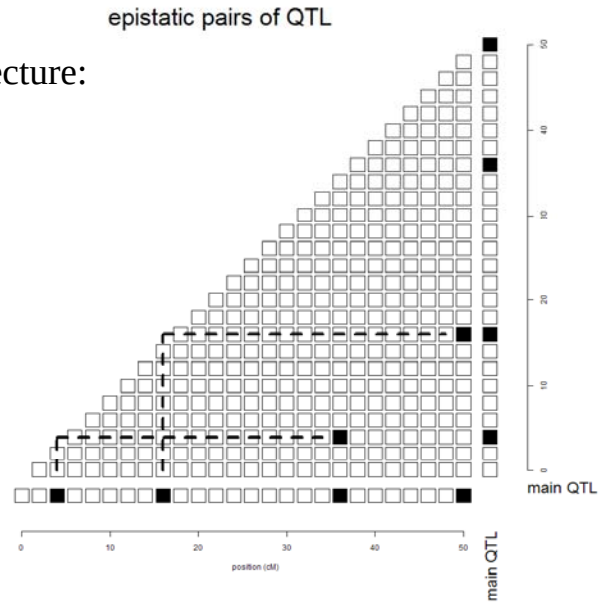
main QTL

epistatic pairs

effects:

add, dom

aa, ad, dd



Bayesian priors & posteriors

- augmenting with missing genotypes q
  - prior is recombination model
  - posterior is (formally) E step of EM algorithm
- sampling phenotype model parameters μ
  - prior is "flat" normal at grand mean (no information)
  - posterior shrinks genotypic means toward grand mean
  - (details for unexplained variance omitted here)
- sampling QTL loci λ
  - prior is flat across genome (all loci equally likely)
- sampling QTL genetic architecture model γ
  - number of QTL
    - prior is Poisson with mean from previous IM study
  - genetic architecture of main effects and epistatic interactions
    - priors on epistasis depend on presence/absence of main effects

2. Markov chain sampling

- construct Markov chain around posterior
  - want posterior as stable distribution of Markov chain
  - in practice, the chain tends toward stable distribution
    - initial values may have low posterior probability
    - burn-in period to get chain mixing well
- sample QTL model components from full conditionals
  - sample locus λ given q,γ (using Metropolis-Hastings step)
  - sample genotypes q given λ,μ,y,γ (using Gibbs sampler)
  - sample effects μ given q,y,γ (using Gibbs sampler)
  - sample QTL model γ given λ,μ,y,q (using Gibbs or M-H)

$$(\lambda, q, \mu, \gamma) \sim \text{pr}(\lambda, q, \mu, \gamma \mid y, m)$$

$$(\lambda, q, \mu, \gamma)_1 \to (\lambda, q, \mu, \gamma)_2 \to \cdots \to (\lambda, q, \mu, \gamma)_N$$

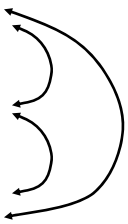
MCMC sampling of unknowns (q,μ,λ)
for given genetic architecture γ

- Gibbs sampler
  - genotypes q
  - effects μ
  - *not* loci λ

$$q \sim \text{pr}(q \mid y_i, m_i, \mu, \lambda)$$

$$\mu \sim \frac{\text{pr}(y \mid q, \mu)\text{pr}(\mu)}{\text{pr}(y \mid q)}$$

$$\lambda \sim \frac{\text{pr}(q \mid m, \lambda)\text{pr}(\lambda \mid m)}{\text{pr}(q \mid m)}$$



- Metropolis-Hastings sampler
  - extension of Gibbs sampler
  - does not require normalization
    - pr(q | m) = sum$_\lambda$ pr(q | m, λ) pr(λ)

Gibbs sampler
for two genotypic means

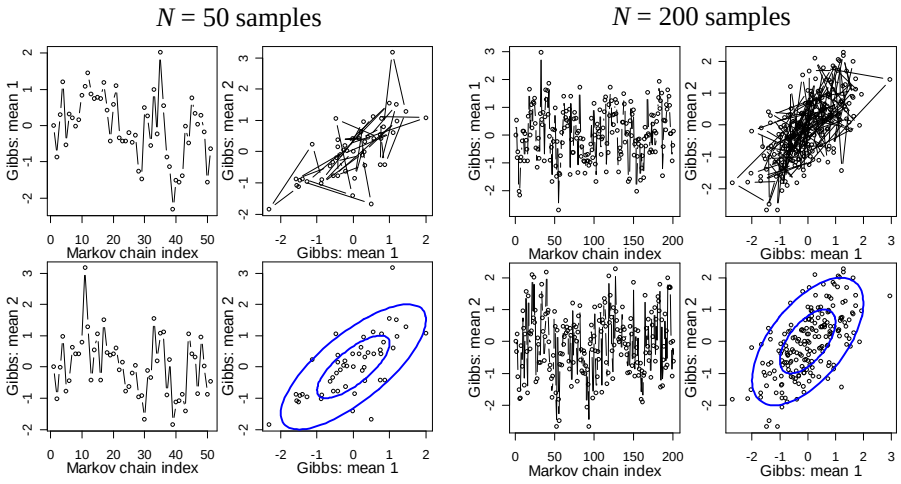
- want to study two correlated effects
  - could sample directly from their bivariate distribution
  - assume correlation ρ is known
- instead use Gibbs sampler:
  - sample each effect from its full conditional given the other
  - pick order of sampling at random
  - repeat many times

$$\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

$$\mu_1 \sim N\left(\rho\mu_2, 1-\rho^2\right)$$

$$\mu_2 \sim N\left(\rho\mu_1, 1-\rho^2\right)$$

Gibbs sampler samples: $\rho = 0.6$



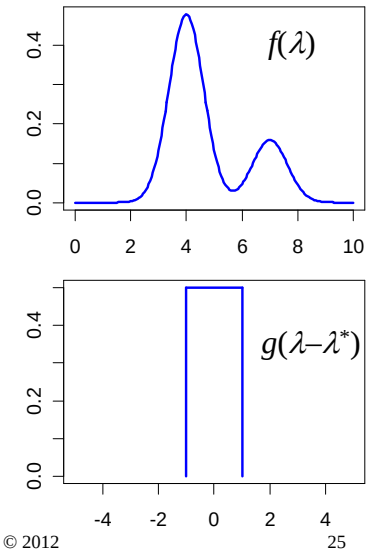
full conditional for locus

- cannot easily sample from locus full conditional
  $\text{pr}(\lambda \mid y,m,\mu,q) = \text{pr}(\lambda \mid m,q)$
  $= \text{pr}(q \mid m, \lambda)\ \text{pr}(\lambda) / \text{constant}$
- constant is very difficult to compute explicitly
  - must average over all possible loci λ over genome
  - must do this for every possible genotype q
- Gibbs sampler will not work in general
  - but can use method based on ratios of probabilities
  - Metropolis-Hastings is extension of Gibbs sampler

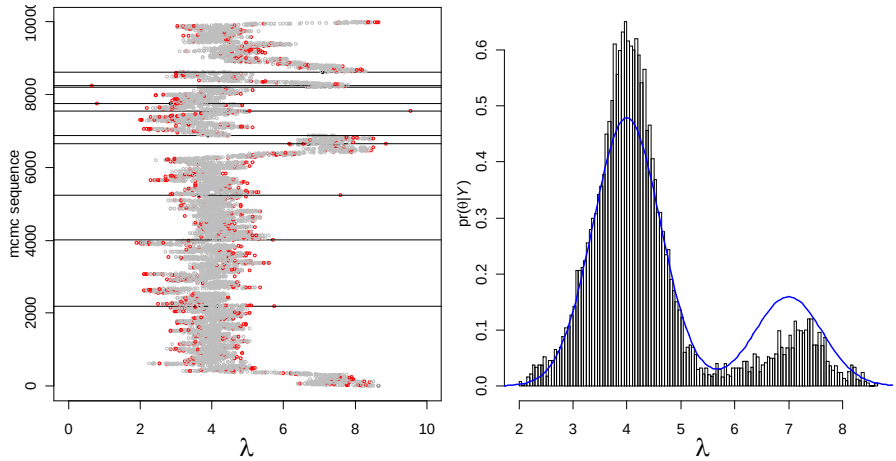
Metropolis-Hastings idea

- want to study distribution $f(\lambda)$
  - take Monte Carlo samples
    - unless too complicated
  - take samples using ratios of f
- Metropolis-Hastings samples:
  - propose new value λ^*
    - near (?) current value λ
    - from some distribution g
  - accept new value with prob a
    - Gibbs sampler: $a = 1$ always

$$a = \min\left(1, \frac{f(\lambda^*)g(\lambda^* - \lambda)}{f(\lambda)g(\lambda - \lambda^*)}\right)$$

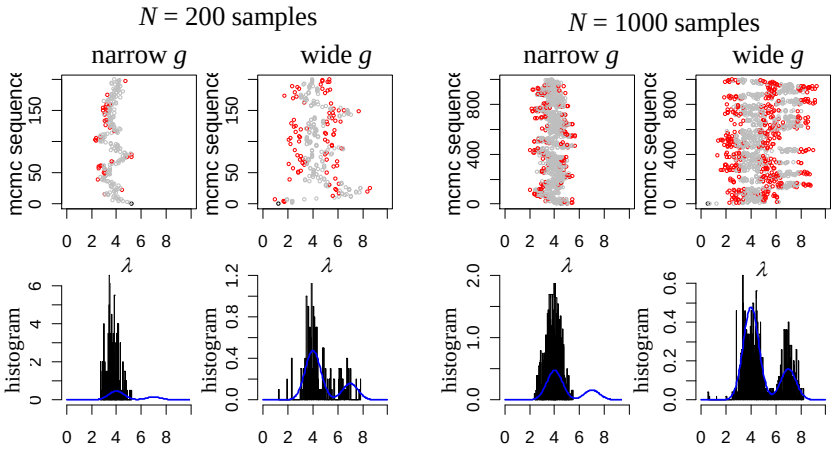


Metropolis-Hastings for locus λ



added twist: occasionally propose from entire genome

Metropolis-Hastings samples



3. sampling genetic architectures

- search across genetic architectures γ of various sizes
  - allow change in number of QTL
  - allow change in types of epistatic interactions
- methods for search
  - reversible jump MCMC
  - Gibbs sampler with loci indicators
- complexity of epistasis
  - Fisher-Cockerham effects model
  - general multi-QTL interaction & limits of inference

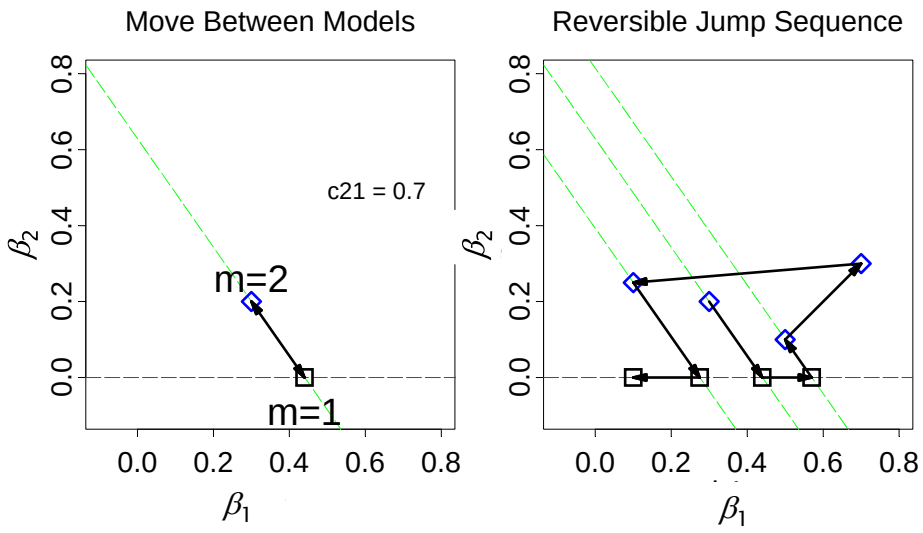
reversible jump MCMC

- consider known genotypes q at 2 known loci λ
  - models with 1 or 2 QTL
- M-H step between 1-QTL and 2-QTL models
  - model changes dimension (via careful bookkeeping)
  - consider mixture over QTL models H

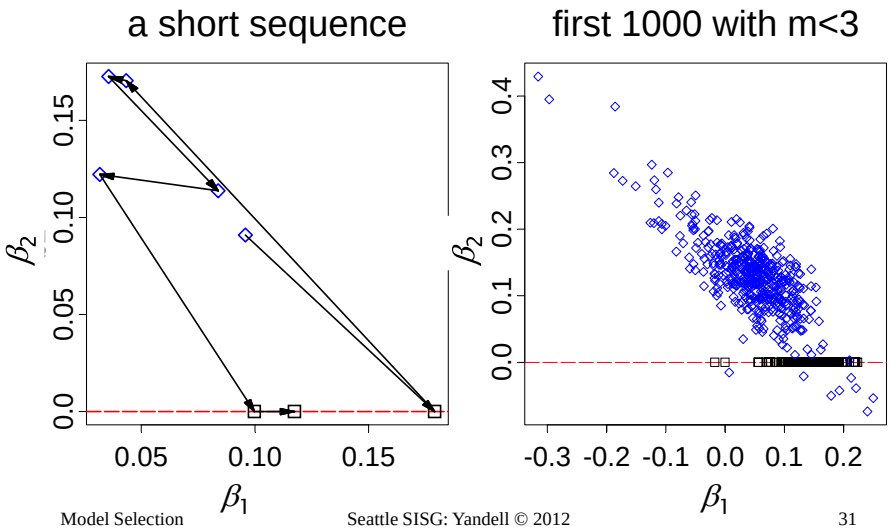
$$\gamma = 1\,\text{QTL} : Y = \beta_0 + \beta(q_1) + e$$

$$\gamma = 2\,\text{QTL} : Y = \beta_0 + \beta_1(q_1) + \beta_2(q_2) + e$$

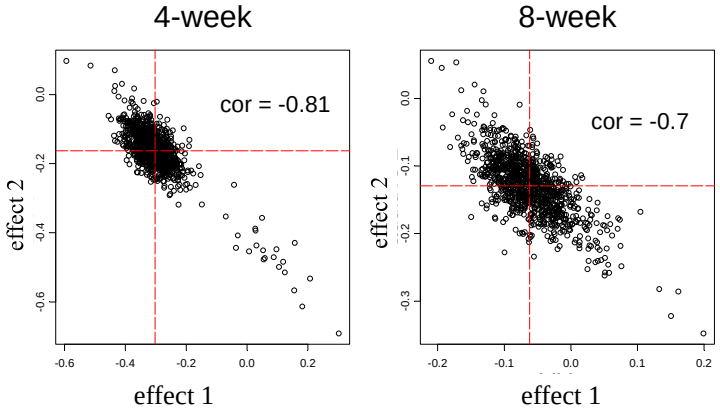
geometry of reversible jump



geometry allowing q and λ to change

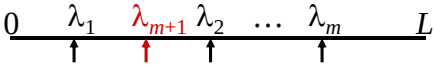


collinear QTL = correlated effects



- linked QTL = collinear genotypes
  - correlated estimates of effects (negative if in coupling phase)
  - sum of linked effects usually fairly constant

sampling across QTL models γ



action steps: draw one of three choices

- update QTL model γ with probability 1-$b(\gamma)$-$d(\gamma)$
  - update current model using full conditionals
  - sample QTL loci, effects, and genotypes
- add a locus with probability $b(\gamma)$
  - propose a new locus along genome
  - innovate new genotypes at locus and phenotype effect
  - decide whether to accept the "birth" of new locus
- drop a locus with probability $d(\gamma)$
  - propose dropping one of existing loci
  - decide whether to accept the "death" of locus

Gibbs sampler with loci indicators

- consider only QTL at pseudomarkers
  - every 1-2 cM
  - modest approximation with little bias
- use loci indicators in each pseudomarker
  - $\gamma = 1$ if QTL present
  - $\gamma = 0$ if no QTL present
- Gibbs sampler on loci indicators γ
  - relatively easy to incorporate epistasis
  - Yi, Yandell, Churchill, Allison, Eisen, Pomp (2005 *Genetics*)
    - (see earlier work of Nengjun Yi and Ina Hoeschele)

$$\mu_q = \mu + \gamma_1 \beta_1(q_1) + \gamma_2 \beta_2(q_2), \quad \gamma_k = 0,1$$

Bayesian shrinkage estimation

- soft loci indicators
  - strength of evidence for λ_j depends on γ
  - $0 \le \gamma \le 1$ (grey scale)
  - shrink most γs to zero
- Wang et al. (2005 *Genetics*)
  - Shizhong Xu group at U CA Riverside

$$\mu_q = \beta_0 + \gamma_1 \beta_1(q_1) + \gamma_2 \beta_2(q_1), \quad 0 \le \gamma_k \le 1$$

other model selection approaches

- include all potential loci in model
- assume "true" model is "sparse" in some sense
- Sparse partial least squares
  - Chun, Keles (2009 *Genetics*; 2010 *JRSSB*)
- LASSO model selection
  - Foster (2006); Foster Verbyla Pitchford (2007 *JABES*)
  - Xu (2007 *Biometrics*); Yi Xu (2007 *Genetics*)
  - Shi Wahba Wright Klein Klein (2008 *Stat & Infer*)

4. criteria for model selection

balance fit against complexity

- classical information criteria
  - penalize likelihood L by model size $|\gamma|$
  - IC = $-2 \log L(\gamma \mid y)$ + penalty(γ)
  - maximize over unknowns
- Bayes factors
  - marginal posteriors pr($y \mid \gamma$)
  - average over unknowns

classical information criteria

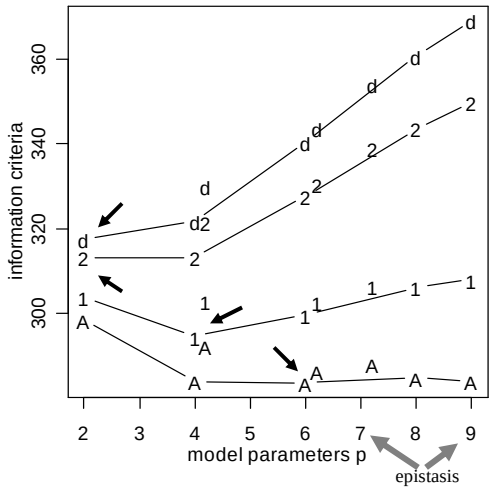
- start with likelihood $L(\gamma \mid y, m)$
  - measures fit of architecture (γ) to phenotype (y)
    - given marker data (m)
  - genetic architecture (γ) depends on parameters
    - have to estimate loci (μ) and effects (λ)
- complexity related to number of parameters
  - $|\gamma|$ = size of genetic architecture
    - BC: $|\gamma| = 1 + n.qtl + n.qtl(n.qtl - 1) = 1 + 4 + 12 = 17$
    - F2: $|\gamma| = 1 + 2n.qtl + 4n.qtl(n.qtl - 1) = 1 + 8 + 48 = 57$

classical information criteria

- construct information criteria
  - balance fit to complexity
  - Akaike AIC $= -2 \log(L) + 2 |\gamma|$
  - Bayes/Schwartz BIC $= -2 \log(L) + |\gamma| \log(n)$
  - Broman $\text{BIC}_\delta = -2 \log(L) + \delta |\gamma| \log(n)$
  - general form: IC $= -2 \log(L) + |\gamma| D(n)$
- compare models
  - hypothesis testing: designed for one comparison
    - $2 \log[LR(\gamma_1, \gamma_2)] = L(y|m, \gamma_2) - L(y|m, \gamma_1)$
  - model selection: penalize complexity
    - $\text{IC}(\gamma_1, \gamma_2) = 2 \log[LR(\gamma_1, \gamma_2)] + (|\gamma_2| - |\gamma_1|) D(n)$

information criteria vs. model size

- WinQTL 2.0
- SCD data on F2
- A=AIC
- 1=BIC(1)
- 2=BIC(2)
- d=BIC(δ)
- models
  - 1,2,3,4 QTL
    - 2+5+9+2
  - epistasis
    - 2:2 AD



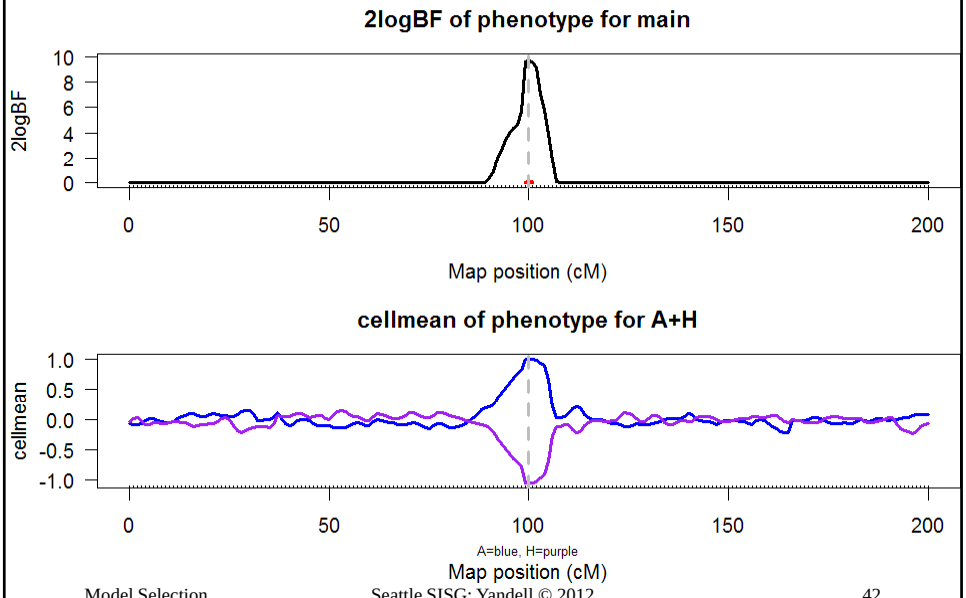
Bayes factors

- ratio of model likelihoods
  - ratio of posterior to prior odds for architectures
  - averaged over unknowns

$$B_{12} = \frac{\text{pr}(\gamma_1 \mid y, m) / \text{pr}(\gamma_2 \mid y, m)}{\text{pr}(\gamma_1) / \text{pr}(\gamma_2)} = \frac{\text{pr}(y \mid m, \gamma_1)}{\text{pr}(y \mid m, \gamma_2)}$$

- roughly equivalent to BIC
  - BIC maximizes over unknowns
  - BF averages over unknowns
  - $-2\log(B_{12}) = -2\log(LR) - (|\gamma_2| - |\gamma_1|)\log(n)$

scan of marginal Bayes factor & effect



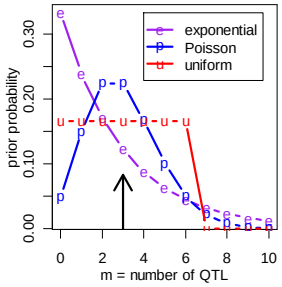
issues in computing Bayes factors

- *BF* insensitive to shape of prior on γ
  - geometric, Poisson, uniform
  - precision improves when prior mimics posterior
- *BF* sensitivity to prior variance on effects θ
  - prior variance should reflect data variability
  - resolved by using hyper-priors
    - automatic algorithm; no need for user tuning
- easy to compute Bayes factors from samples
  - sample posterior using MCMC
  - posterior pr($\gamma \mid y, m$) is marginal histogram

Bayes factors & genetic architecture γ

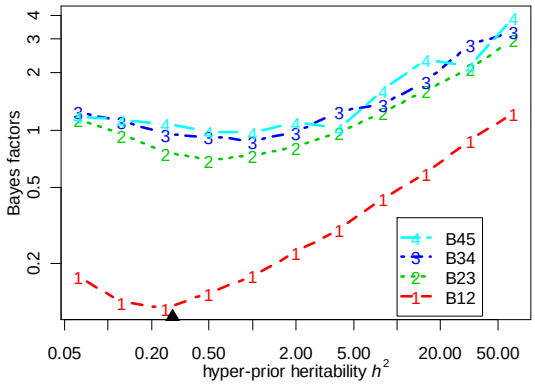
- $|\gamma|$ = number of QTL
  - prior pr(γ) chosen by user
  - posterior pr($\gamma \mid y,m$)
    - sampled marginal histogram
    - shape affected by prior pr(*A*)

$$BF_{\gamma_1,\gamma_2} = \frac{\Pr(\gamma_1|y,m)/\Pr(\gamma_1)}{\Pr(\gamma_2|y,m)/\Pr(\gamma_2)}$$



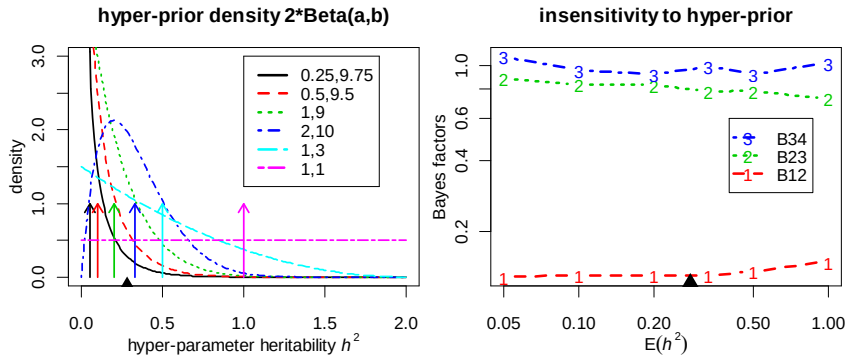
- pattern of QTL across genome
- gene action and epistasis

BF sensitivity to fixed prior for effects



$$\beta_{qj} \sim \mathrm{N}\left(0, \sigma_G^2 / m\right), \sigma_G^2 = h^2 \sigma_{\text{total}}^2, h^2 \text{ fixed}$$

BF insensitivity to random effects prior



$$\beta_{qj} \sim \mathrm{N}\left(0, \sigma_G^2 / m\right), \sigma_G^2 = h^2 \sigma_{\text{total}}^2, \tfrac{1}{2} h^2 \sim \mathrm{Beta}(a, b)$$

Multiple Correlated Traits

- Pleiotropy vs. close linkage
- Analysis of covariance
  - Regress one trait on another before QTL search
- Classic GxE analysis
- Formal joint mapping (MTM)
- Seemingly unrelated regression (SUR)
- Reducing many traits to one
  - Principle components for *similar* traits

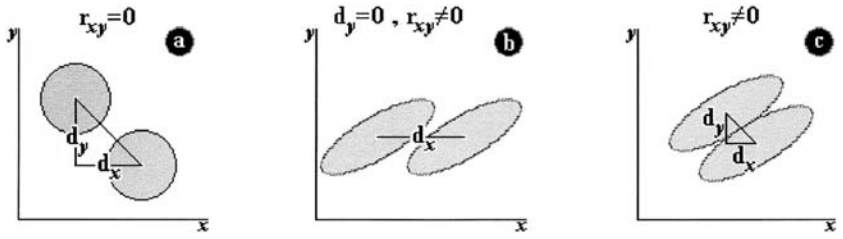
co-mapping multiple traits

- avoid reductionist approach to biology
  - address physiological/biochemical mechanisms
  - Schmalhausen (1942); Falconer (1952)
- separate close linkage from pleiotropy
  - 1 locus or 2 linked loci?
- identify epistatic interaction or canalization
  - influence of genetic background
- establish QTL x environment interactions
- decompose genetic correlation among traits
- increase power to detect QTL

Two types of data

- Design I: multiple traits on same individual
  - Related measurements, say of shape or size
  - Same measurement taken over time
  - Correlation within an individual
- Design II: multiple traits on different individuals
  - Same measurement in two crosses
  - Male vs. female differences
  - Different individuals in different locations
  - No correlation between individuals

interplay of pleiotropy & correlation



pleiotropy only *correlation only* *both*

Korol et al. (2001)

Brassica napus: 2 correlated traits

- 4-week & 8-week vernalization effect
  - log(days to flower)
- genetic cross of
  - Stellar (annual canola)
  - Major (biennial rapeseed)
- 105 F1-derived double haploid (DH) lines
  - homozygous at every locus (*QQ* or *qq*)
- 10 molecular markers (RFLPs) on LG9
  - two QTLs inferred on LG9 (now chromosome N2)
  - corroborated by Butruille (1998)
  - exploiting synteny with *Arabidopsis thaliana*

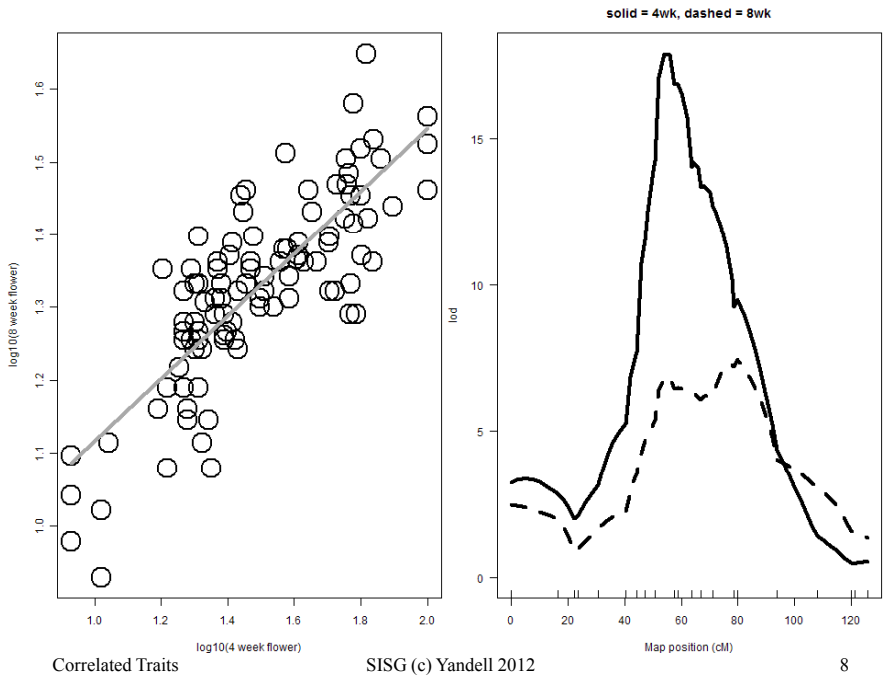
QTL with GxE or Covariates

- adjust phenotype by covariate
  - covariate(s) = environment(s) or other trait(s)
- additive covariate
  - covariate adjustment same across genotypes
  - "usual" analysis of covariance (ANCOVA)
- interacting covariate
  - address GxE
  - capture genotype-specific relationship among traits
- another way to think of multiple trait analysis
  - examine single phenotype adjusted for others

R/qtl & covariates

- additive and/or interacting covariates
- test for QTL after adjusting for covariates

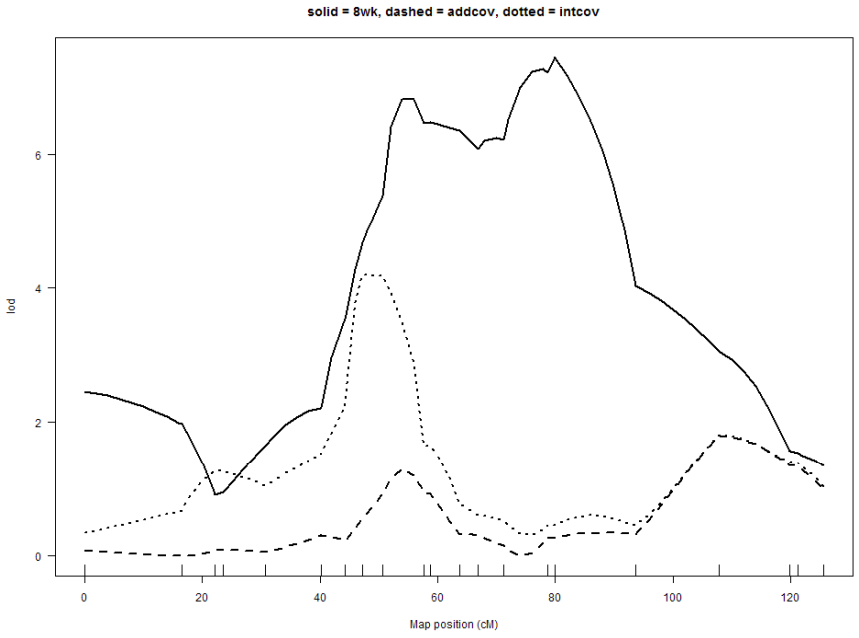
```
## Get Brassica data.
library(qtlbim)
data(Bnapus)
Bnapus <- calc.genoprob(Bnapus, step = 2, error = 0.01)

## Scatterplot of two phenotypes: 4wk & 8wk flower time.
plot(Bnapus$pheno$log10flower4,Bnapus$pheno$log10flower8)

## Unadjusted IM scans of each phenotype.
fl8 <- scanone(Bnapus,, find.pheno(Bnapus, "log10flower8"))
fl4 <- scanone(Bnapus,, find.pheno(Bnapus, "log10flower4"))
plot(fl4, fl8, chr = "N2", col = rep(1,2), lty = 1:2,
  main = "solid = 4wk, dashed = 8wk", lwd = 4)
```

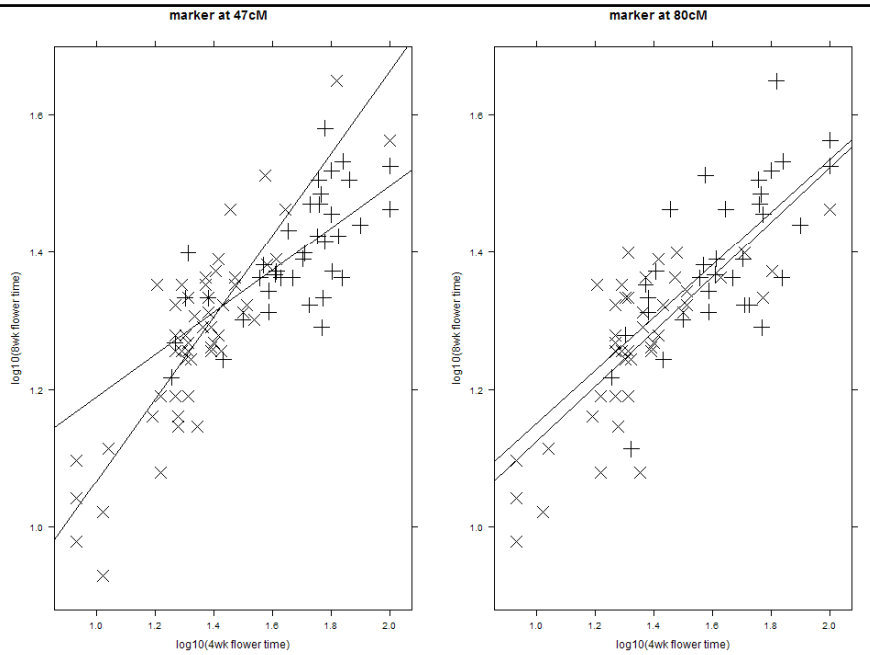



R/qtl & covariates

- additive and/or interacting covariates
- test for QTL after adjusting for covariates

```
## IM scan of 8wk adjusted for 4wk.
## Adjustment independent of genotype
fl8.4 <- scanone(Bnapus,, find.pheno(Bnapus, "log10flower8"),
  addcov = Bnapus$pheno$log10flower4)

## IM scan of 8wk adjusted for 4wk.
## Adjustment changes with genotype.
fl8.4 <- scanone(Bnapus,, find.pheno(Bnapus, "log10flower8"),
  intcov = Bnapus$pheno$log10flower4)

plot(fl8, fl8.4a, fl8.4, chr = "N2",
  main = "solid = 8wk, dashed = addcov, dotted = intcov")
```



scatterplot adjusted for covariate

```
## Set up data frame with peak markers, traits.
markers <- c("E38M50.133","ec2e5a","wg7f3a")
tmpdata <- data.frame(pull.geno(Bnapus)[,markers])
tmpdata$fl4 <- Bnapus$pheno$log10flower4
tmpdata$fl8 <- Bnapus$pheno$log10flower8

## Scatterplots grouped by marker.
library(lattice)
xyplot(fl8 ~ fl4, tmpdata, group = wg7f3a,
  col = "black", pch = 3:4, cex = 2, type = c("p","r"),
  xlab = "log10(4wk flower time)",
  ylab = "log10(8wk flower time)",
  main = "marker at 47cM")
xyplot(fl8 ~ fl4, tmpdata, group = E38M50.133,
  col = "black", pch = 3:4, cex = 2, type = c("p","r"),
  xlab = "log10(4wk flower time)",
  ylab = "log10(8wk flower time)",
  main = "marker at 80cM")
```

Multiple trait mapping

- Joint mapping of QTL
  - testing and estimating QTL affecting multiple traits
- Testing pleiotropy vs. close linkage
  - One QTL or two closely linked QTLs
- Testing QTL x environment interaction
- Comprehensive model of multiple traits
  - Separate genetic & environmental correlation

Formal Tests: 2 traits

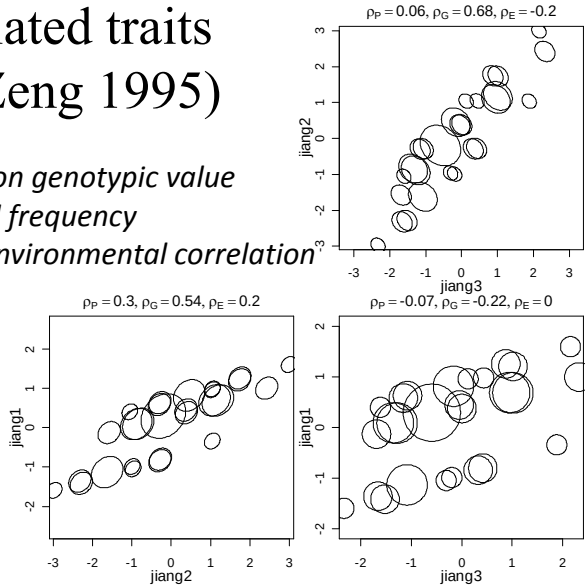
$y_1 \sim N(\mu_{q1}, \sigma^2)$ for group 1 with QTL at location λ_1
$y_2 \sim N(\mu_{q2}, \sigma^2)$ for group 2 with QTL at location λ_2

- Pleiotropy vs. close linkage
  - test QTL at same location: $\lambda_1 = \lambda_2$
  - likelihood ratio test (LOD): null forces same location
- if pleiotropic ($\lambda_1 = \lambda_2$)
  - test for same mean: $\mu_{q1} = \mu_{q2}$
  - Likelihood ratio test (LOD)
    - null forces same mean, location
    - alternative forces same location
  - only make sense if traits are on same scale
  - test sex or location effect

3 correlated traits (Jiang Zeng 1995)

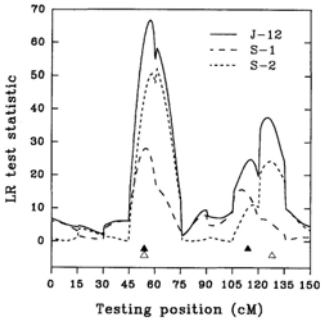
ellipses centered on genotypic value
width for nominal frequency
main axis angle environmental correlation
3 QTL, F2
27 genotypes

note signs of genetic and environmental correlation



pleiotropy or close linkage?

2 traits, 2 qtl/trait
pleiotropy @ 54cM
linkage @ 114,128cM
Jiang Zeng (1995)



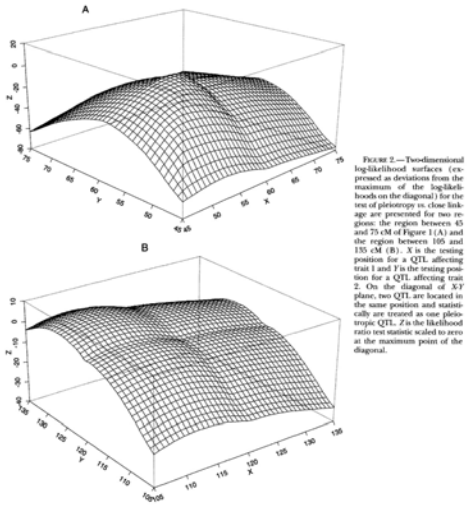


FIGURE 2.—Two-dimensional log-likelihood surfaces (expressed as deviations from the maximum of the log-likelihoods on the diagonal) for the test of pleiotropy vs. close linkage are presented for two regions: the region between 45 and 75 cM of Figure 1(A) and the region between 105 and 135 cM (B). X is the testing position for a QTL affecting trait 1 and Y is the testing position for a QTL affecting trait 2. On the diagonal of X/Y plane, two QTL are located in the same position and statistically are treated as one pleiotropic QTL. Z is the likelihood ratio test statistic scaled to zero at the maximum point of the diagonal.

More detail for 2 traits

$y_1 \sim N(\mu_{q1}, \sigma^2)$ for group 1

$y_2 \sim N(\mu_{q2}, \sigma^2)$ for group 2

- two possible QTLs at locations λ_1 and λ_2
- effect β_{kj} in group k for QTL at location λ_j

$\mu_{q1} = \mu_1 + \beta_{11}(q_1) + \beta_{12}(q_2)$

$\mu_{q2} = \mu_2 + \beta_{21}(q_1) + \beta_{22}(q_2)$

- classical: test $\beta_{kj} = 0$ for various combinations

seemingly unrelated regression (SUR)

$\mu_{q1} = \mu_1 + \gamma_{11}\beta_{q11} + \gamma_{12}\beta_{q12}$

$\mu_{q2} = \mu_2 + \gamma_{21}\beta_{q21} + \gamma_{22}\beta_{q22}$

indicators γ_{kj} are 0 (no QTL) or 1 (QTL)

- include γs in formal model selection

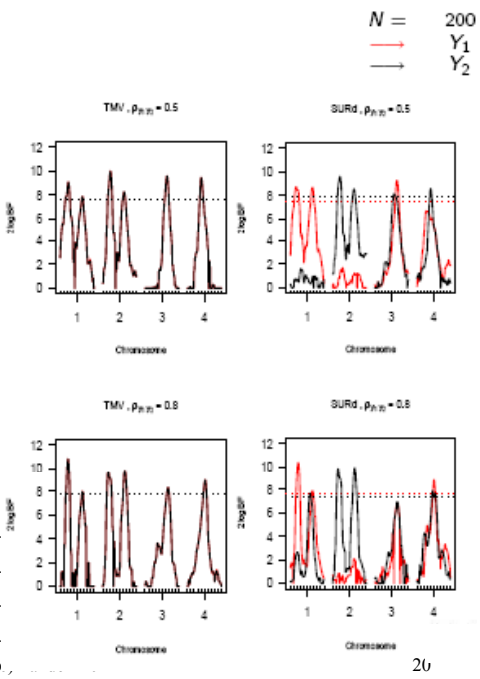
SUR for multiple loci across genome

- consider only QTL at pseudomarkers (lecture 2)
- use loci indicators γ_j (=0 or 1) for each pseudomarker
- use SUR indicators γ_{kj} (=0 or 1) for each trait
- Gibbs sampler on both indicators
  - Banerjee, Yandell, Yi (2008 *Genetics*)

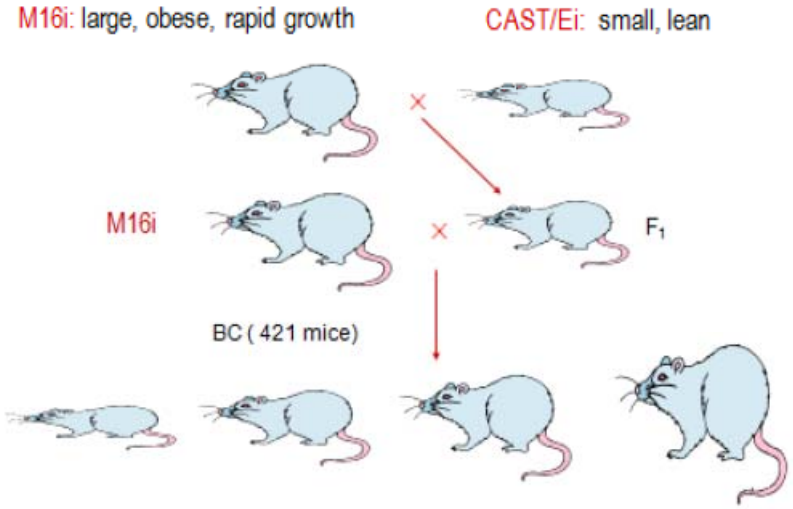
$$\mu_{q1} = \mu_1 + \gamma_1\gamma_{11}\beta_{11}(q_1) + \gamma_2\gamma_{12}\beta_{12}(q_2) + \ldots$$

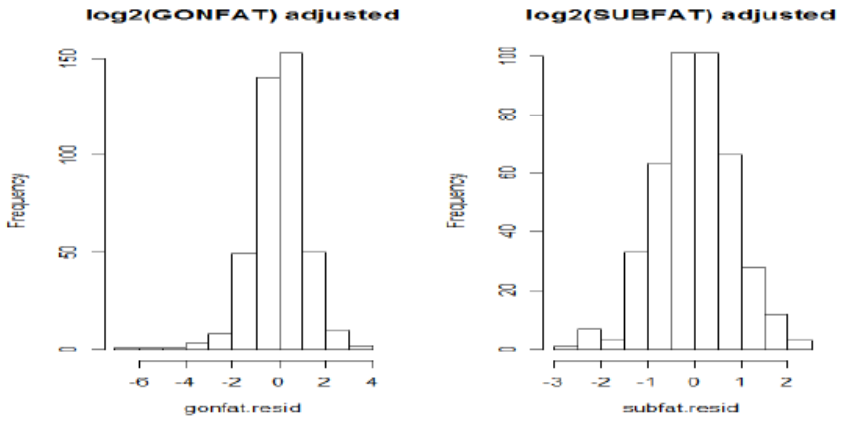
$$\mu_{q2} = \mu_2 + \gamma_1\gamma_{21}\beta_{21}(q_1) + \gamma_2\gamma_{22}\beta_{22}(q_2) + \ldots$$

Simulation
5 QTL
2 traits
n=200
TMV vs. SUR

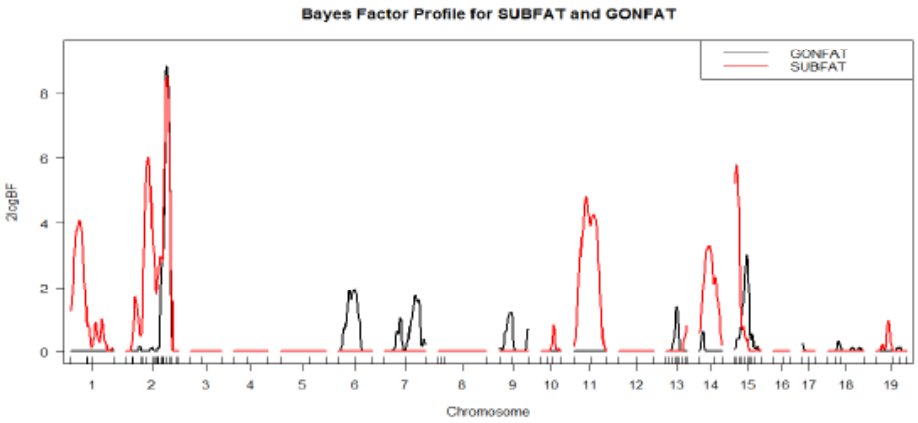


	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6
Chr	1	1	2	2	3	4
Pos(cM)	22	55	22	65	65	45
y_1	0.8	0.6	0	0	0.8	0.6
y_2	0	0	-0.8	-0.6	0.8	0.6
y_1	8.8%	4.9%	0	0	8.8%	4.9%
y_2	0	0	9.3%	5.2%	9.3%	5.2%





- GONFAT → Right Gonadal fat pad
- SUBFAT → Subcutaneous fat pad



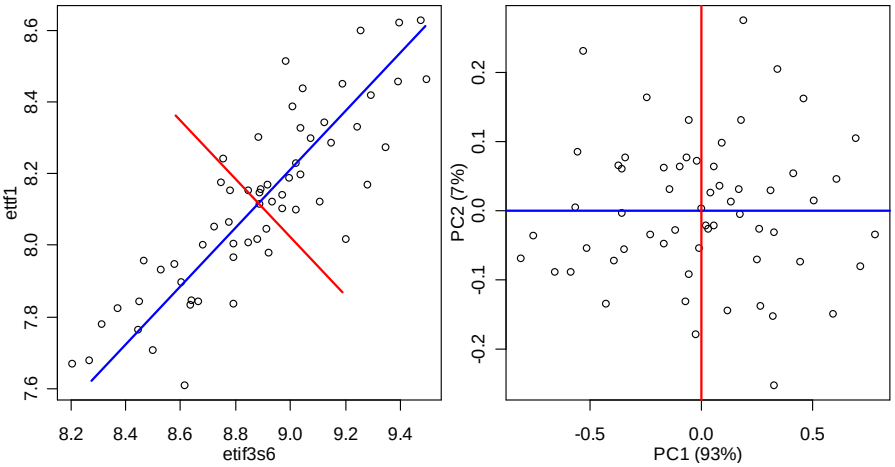
R/qtlbim and GxE

- similar idea to R/qtl
  - fixed and random additive covariates
  - GxE with fixed covariate
- multiple trait analysis tools coming soon
  - theory & code mostly in place
  - properties under study
  - expect in R/qtlbim later this year
  - Samprit Banerjee (N Yi, advisor)

reducing many phenotypes to 1

- *Drosophila mauritiana* x *D. simulans*
  - reciprocal backcrosses, ~500 per bc
- response is "shape" of reproductive piece
  - trace edge, convert to Fourier series
  - reduce dimension: first principal component
- many linked loci
  - brief comparison of CIM, MIM, BIM

PC for two correlated phenotypes



shape phenotype via PC

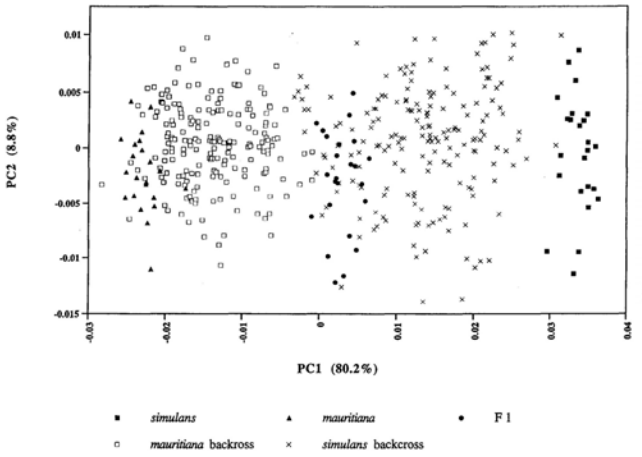


FIGURE 5.—A plot of the first two principal components of the Fourier coefficients from posterior lobe outlines. Many individuals from each of five genotypic classes are represented. Each point represents an average of scores from the left and right sides of an individual (with a few exceptions for which the score is from one side only). The percentage of variation in the Fourier coefficients accounted for by each principal component is given in parentheses. *Liu et al. (1996) Genetics*

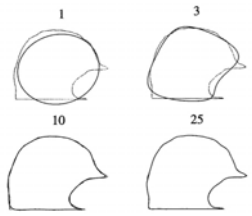


FIGURE 2.—The effect of harmonic number on the accuracy of reconstruction of a posterior lobe outline by elliptical Fourier analysis.

shape phenotype in BC study indexed by PC1

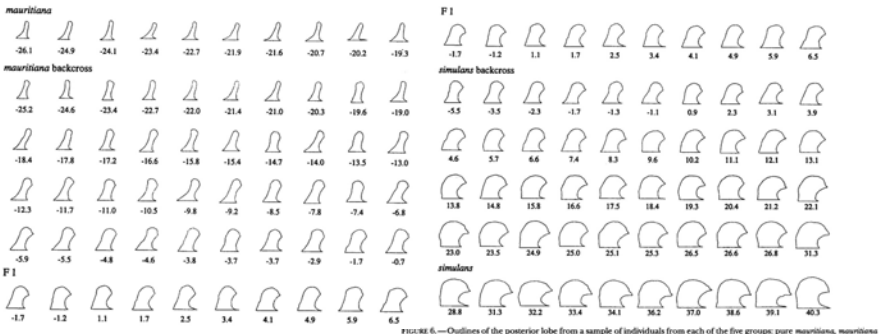


FIGURE 6.—Outlines of the posterior lobe from a sample of individuals from each of the five groups: pure *mauritiana*, *mauritiana* backcross, F_1, *simulans* backcross, and pure *simulans*. Within each group, the outlines are presented in order of their PC1 score (sampled at even intervals from the range of variation). The number below each specimen is its PC1 score. The outlines are drawn to scale with the origin at the centroid of each outline and with all baselines parallel.

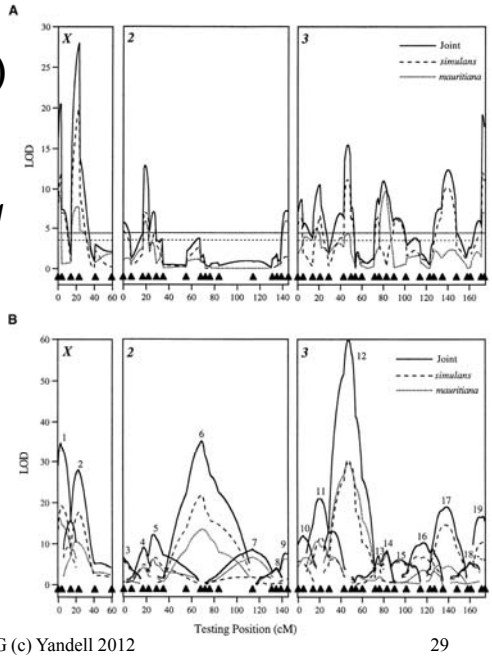
Liu et al. (1996) Genetics

Zeng et al. (2000) CIM vs. MIM

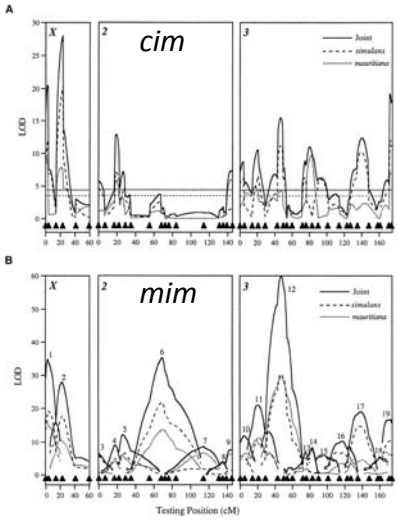
composite interval mapping
(Liu et al. 1996)
narrow peaks
miss some QTL

multiple interval mapping
(Zeng et al. 2000)
triangular peaks

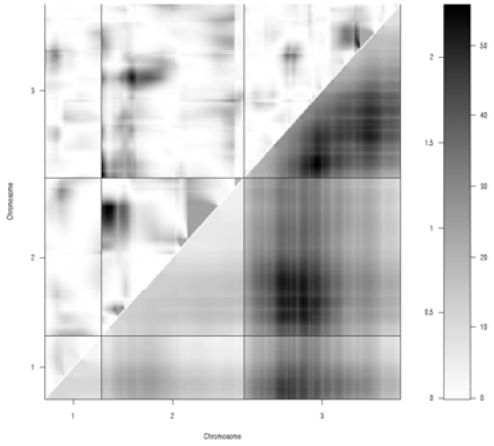
both conditional 1-D scans
fixing all other "QTL"



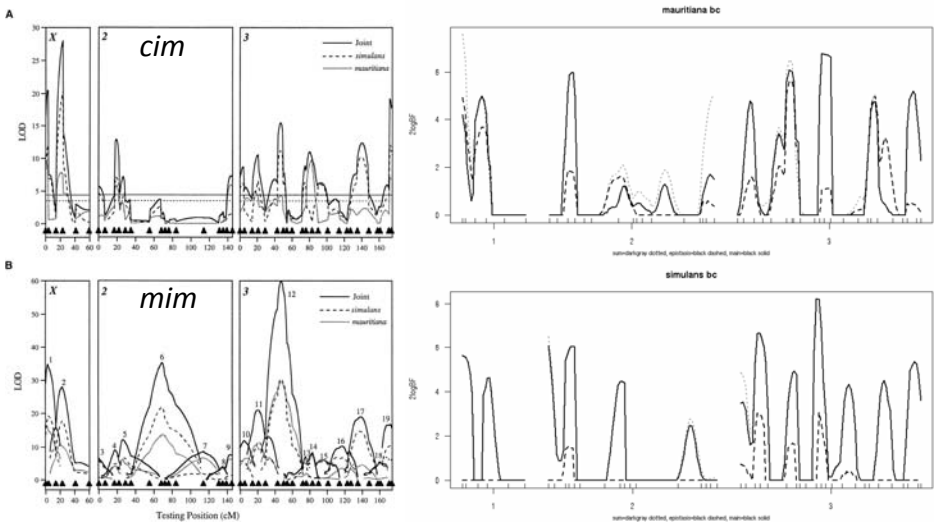
CIM, MIM and IM pairscan



2-D im



multiple QTL: CIM, MIM and BIM



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)

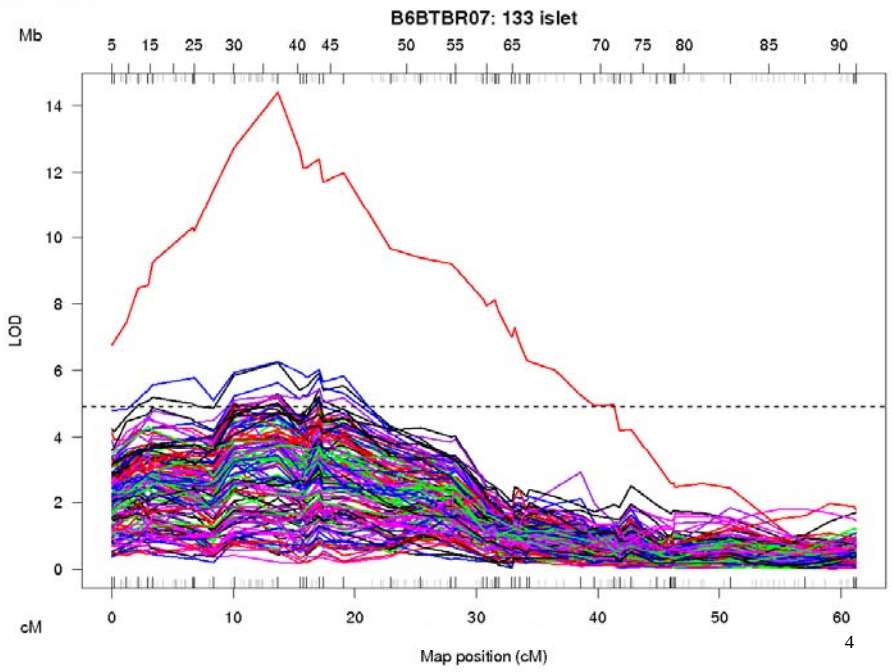
QTL hotspots = genomic locations affecting many traits

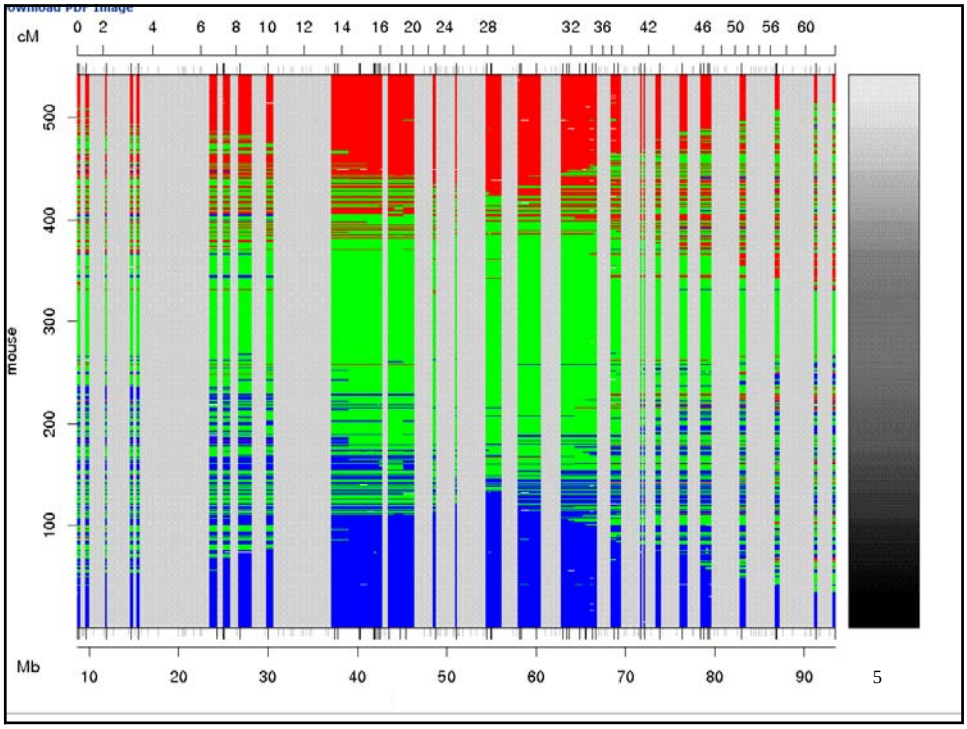
common feature in genetical genomics studies

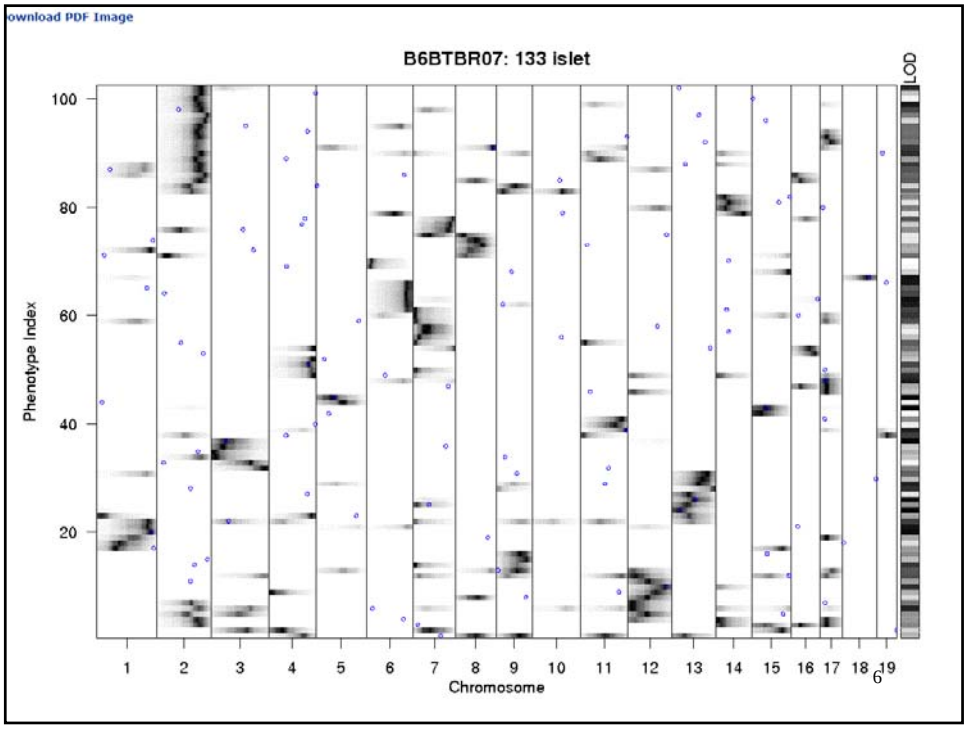
biologically interesting--may harbor critical regulators

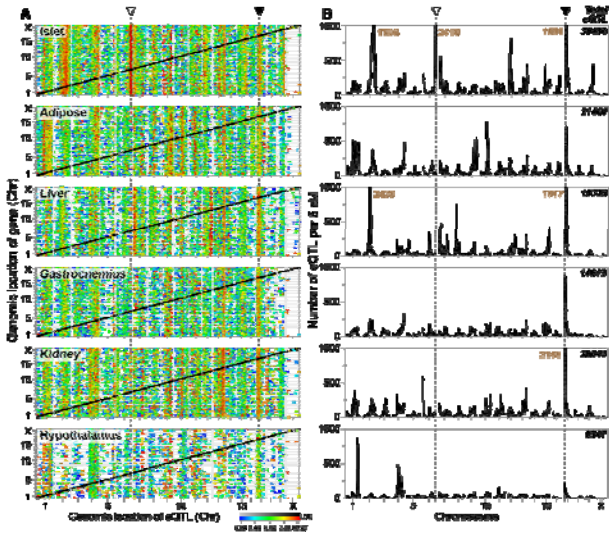
But are these hotspots real? Or are they spurious or random?

non-genetic correlation from other environmental factors




cM
0 2 4 6 8 10 12 14 16 20 24 28 32 36 42 46 50 56 60
mouse
500
400
300
200
100
Mb
10 20 30 40 50 60 70 80 90
5


ownload PDF Image
B6BTBR07: 133 islet
LOD
Phenotype Index
100
80
60
40
20
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
Chromosome



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).*

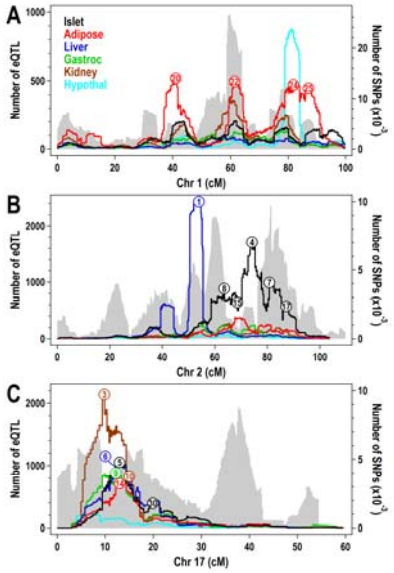


Figure 4 *Tissue-specific hotspots with eQTL and SNP architecture for Chrs 1, 2 and 17.*

*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.*

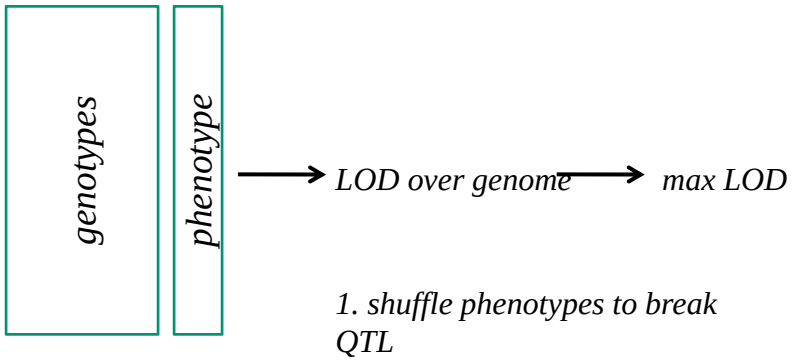
How large a hotspot is large?

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

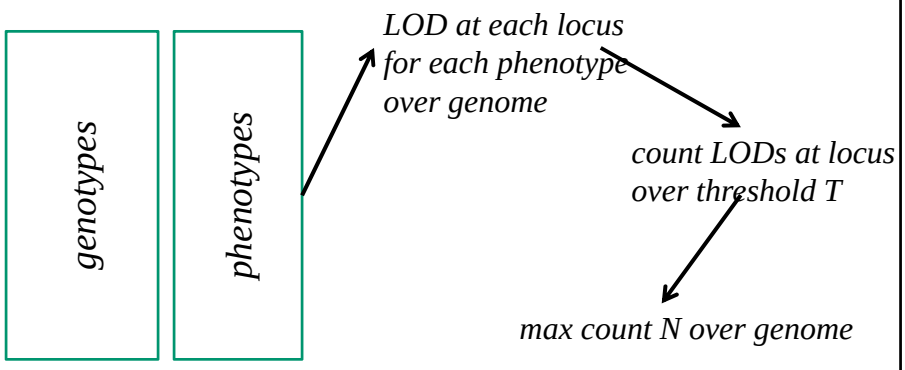


1. shuffle phenotypes to break QTL
2. repeat 1000 times and summarize

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

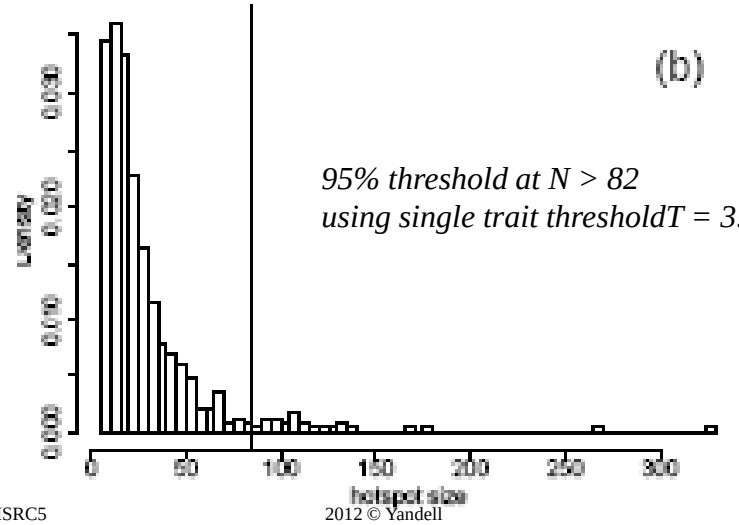


1. shuffle phenotypes by row to break QTL, keep correlation

2. repeat 1000 times and summarize

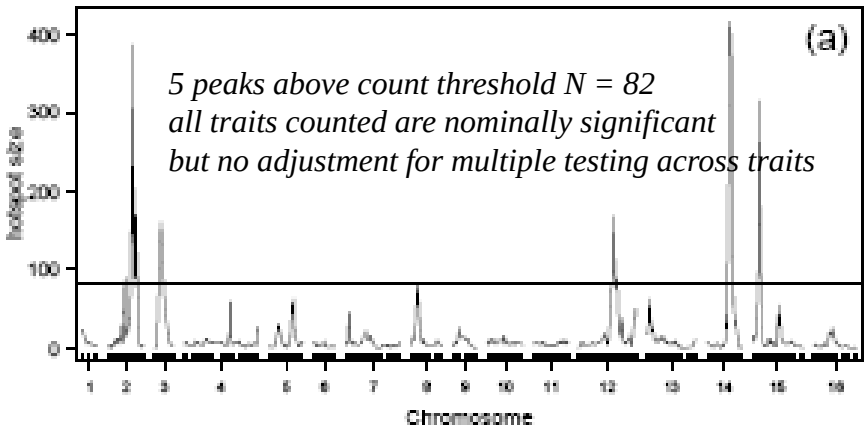
spurious hotspot permutation histogram

for hotspot size above 1 trait threshold



95% threshold at N > 82
using single trait thresholdT = 3.41

Hotspot sizes based on count of LODs above single-trait



hotspot permutation test

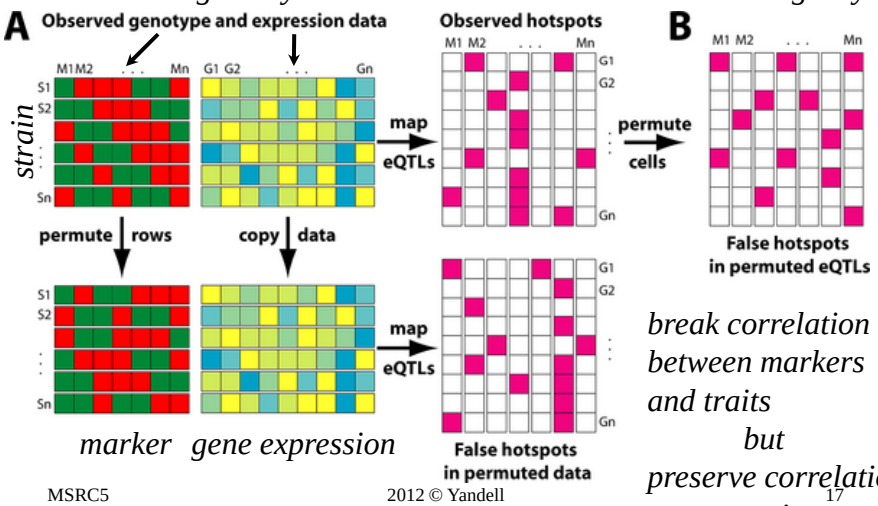
(Breitling et al. Jansen 2008 *PLoS Genetics*)

- 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*)

right way *wrong way*



break correlation between markers and traits

but

preserve correlation among traits

quality vs. quantity in hotspots

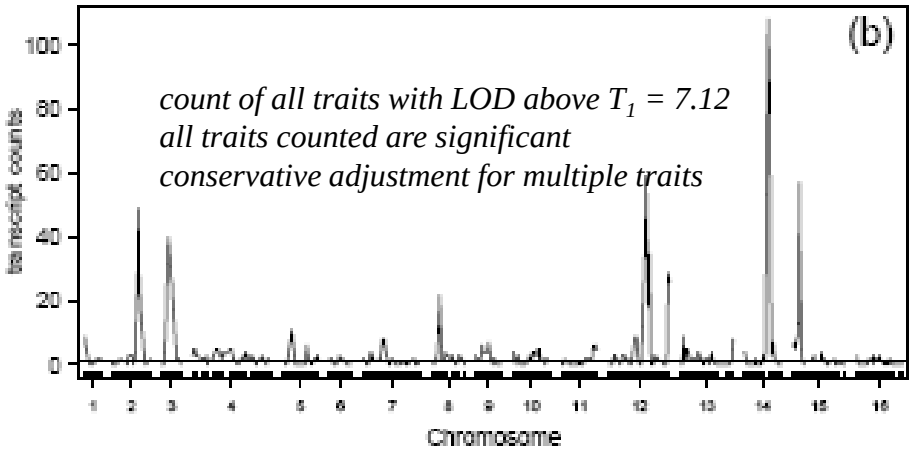
(Chaibub Neto et al. in review)

- 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

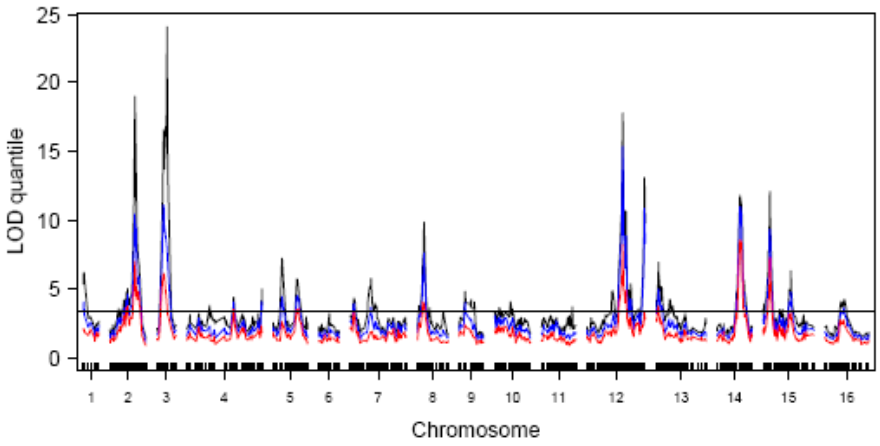
Rethinking the approach

- 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

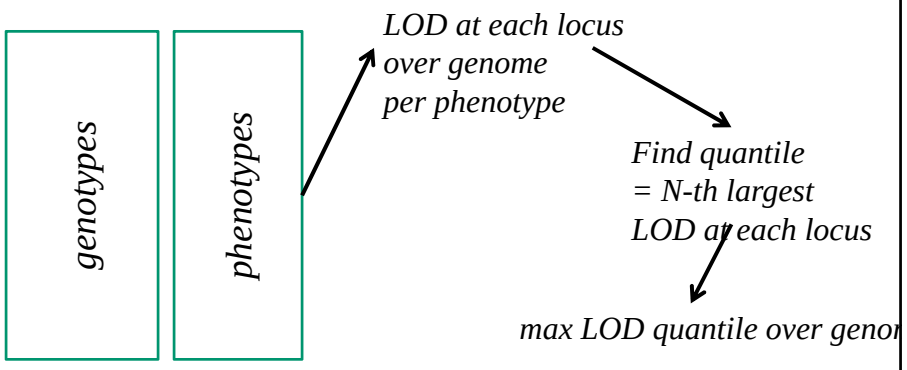


locus-specific LOD quantiles

- 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



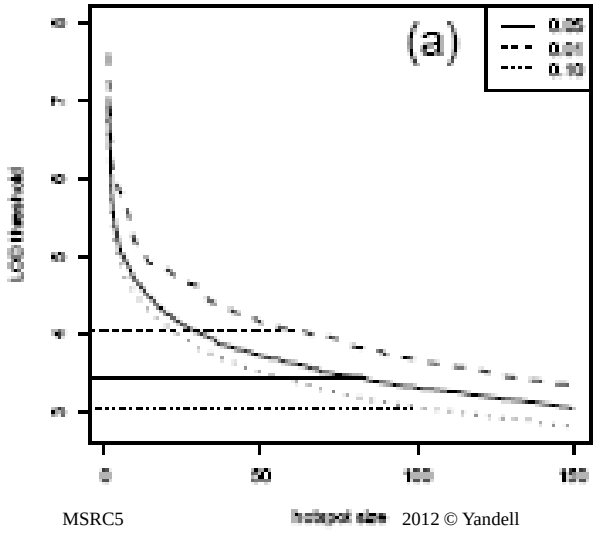
1. shuffle phenotypes by row to break QTL, keep correlation

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

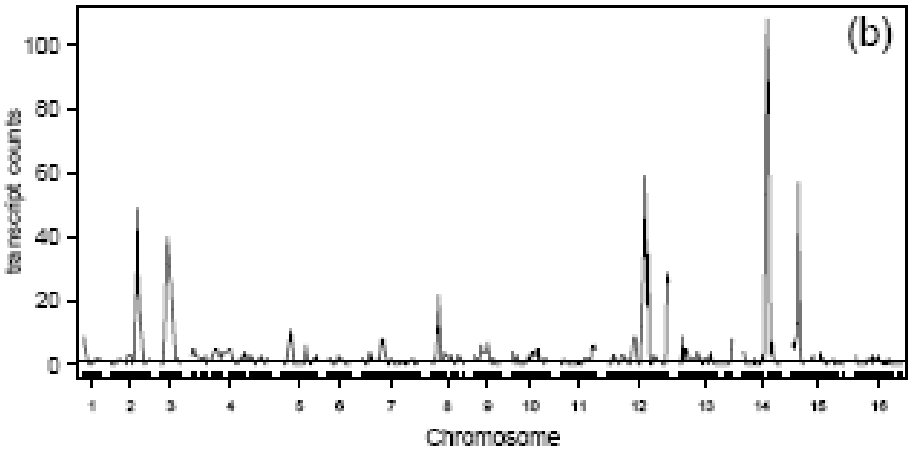


size

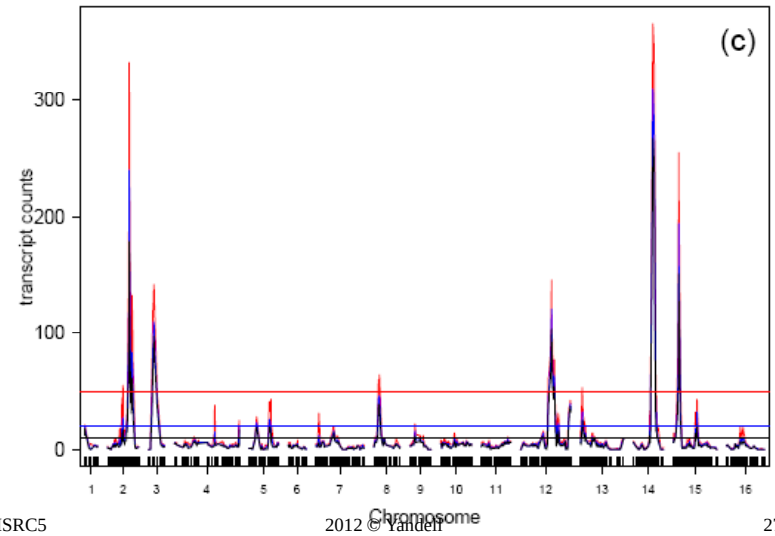
smaller spurious hotspots have higher LOD thresholds

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

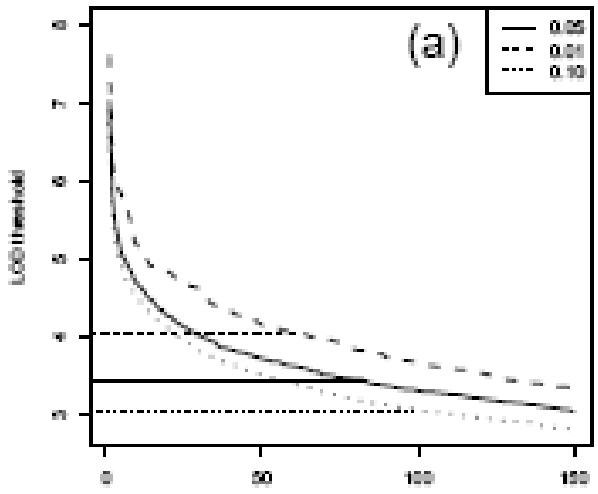
Hotspot architecture using multiple trait GWER threshold (*T* =7.12)



hotspot architectures using LOD thresholds



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



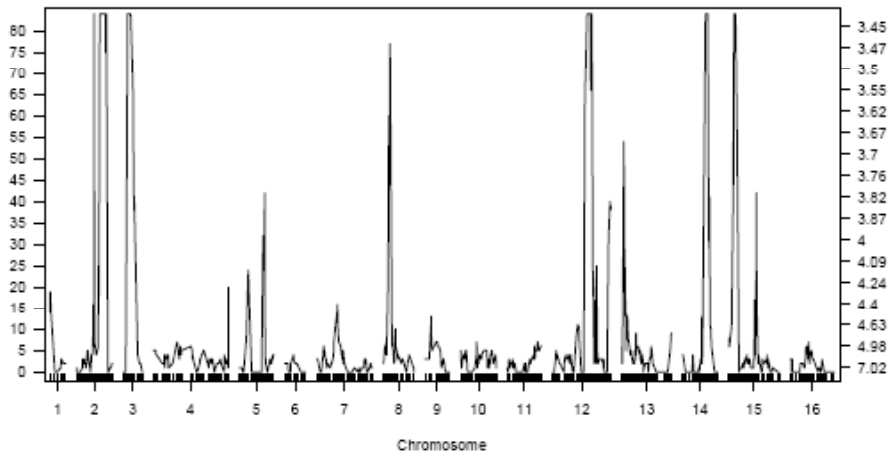
T_1=7.12 controls GWER across all traits

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



Yeast study

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

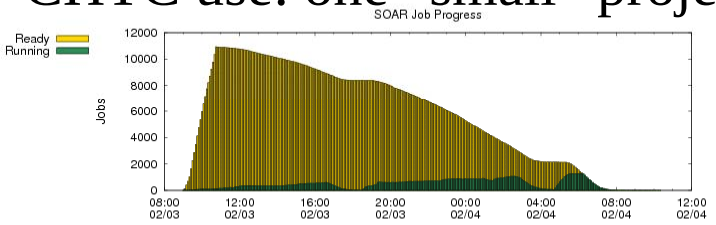
Mouse study

- 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?

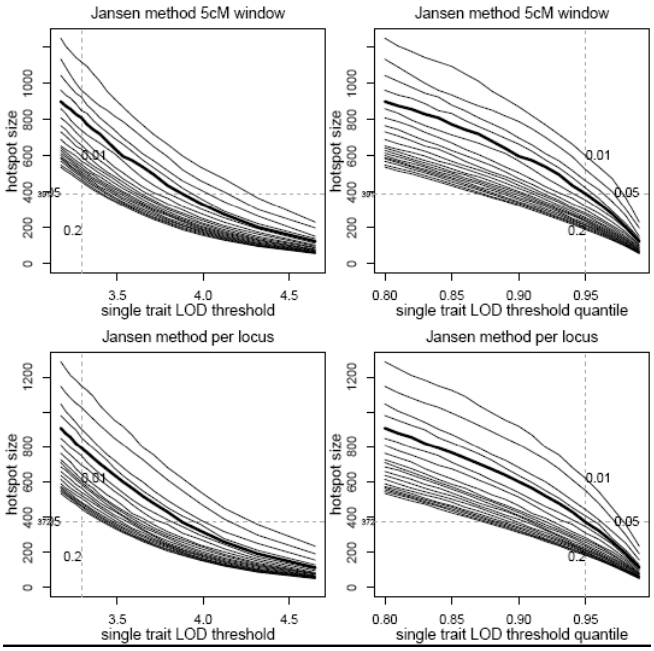
CHTC use: one “small” project

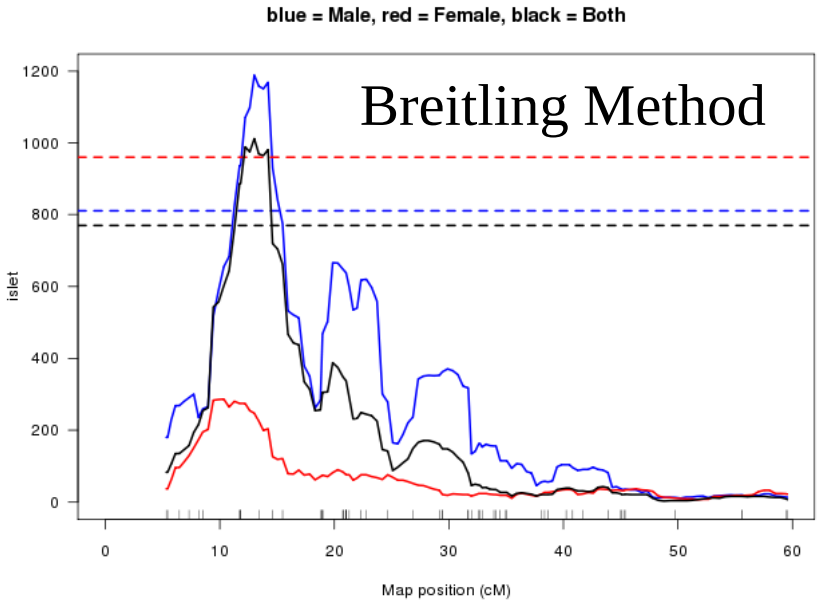


Open Science Grid Glidein Usage (4 feb 2012)

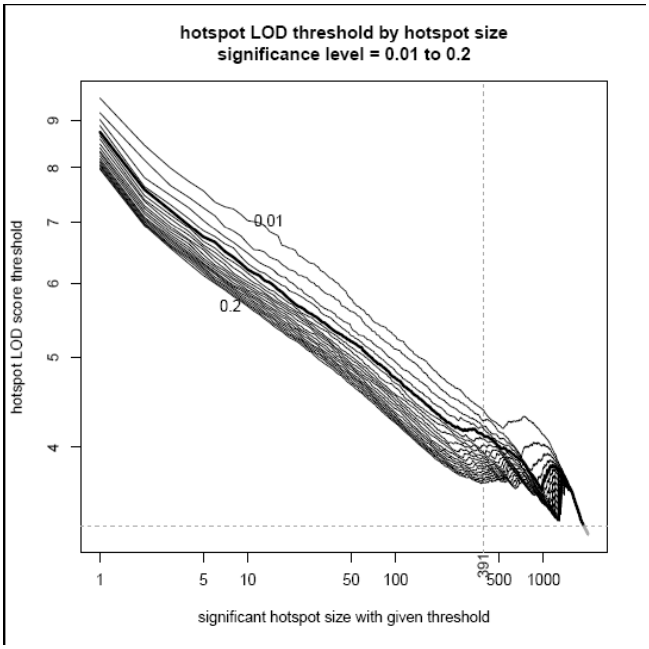
	group	hours	percent
1	BMRB	10710.3	73.49%
2	Biochem_Attie	3660.2	25.11%
3	Statistics_Wahba	178.5	1.22%

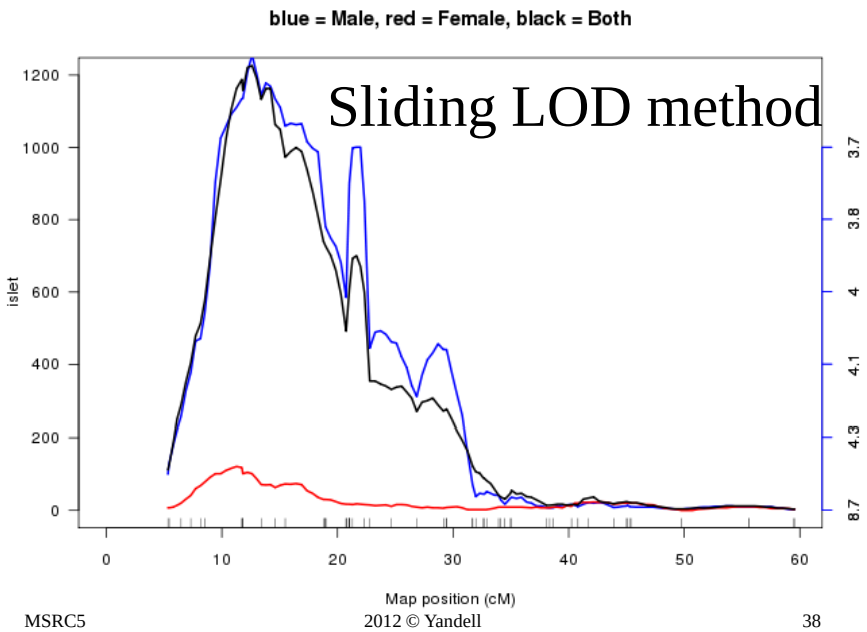
Brietling et al (2008) hotspot size thresholds from permutations





Chaibub Neto sliding LOD thresholds





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.

Causal Graphical Models

Elias Chaibub Neto and Brian S Yandell

SISG 2012

July 12, 2012

Correlation and Causation

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.

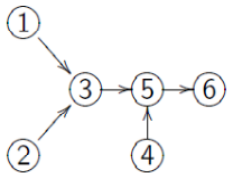
Sewall Wright (1921)

Basic concepts

Directed graphical models

A graphical model is a multivariate probabilistic model whose conditional independence relations are represented by a graph.

We will focus on directed acyclic graph (DAG) models (aka Bayes nets),



Assuming the Markov property, the joint distribution factors according to the conditional independence relations:

$$P(1,2,3,4,5,6) = P(6 \mid 5)\, P(5 \mid 3,4)\, P(4)\, P(3 \mid 1,2)\, P(2)\, P(1)$$

$$6 \perp\!\!\!\perp \{1,2,3,4\} \mid 5, \quad 5 \perp\!\!\!\perp \{1,2,3\} \mid 4, \quad \text{and so on}$$

i.e., each node is independent of its non-descendants given its parents.

Standard Bayesian networks and causality

Even though the direct edges in a Bayes net are often interpreted as causal relations, in reality they only represent conditional dependencies.

Different phenotype networks, for instance,

$$Y_1 \rightarrow Y_2 \rightarrow Y_3 \ , \quad Y_1 \leftarrow Y_2 \rightarrow Y_3 \ , \quad Y_1 \leftarrow Y_2 \leftarrow Y_3 \ ,$$

can represent the same set of conditional independence relations ($Y_1 \perp\!\!\!\perp Y_3 \mid Y_2$, in this example). When that is the case, we say the nets are *Markov equivalent*.

In general (although it is not always true), Markov equivalent networks will have equivalent likelihood functions, so that model selection criteria cannot distinguish between them. The best we can do is to learn *equivalent classes* of *likelihood equivalent* phenotype networks from the data.

Genetics as a mean to reduce the size of equivalence classes

The incorporation of genetic information can help distinguish between likelihood equivalent nets two distinct ways:

1. By creating priors for the network structures, using the results of causality tests (Zhu et al. 2007).
2. By augmenting the phenotype network with QTL nodes, creating new sets of conditional independence relations (Chaibub Neto et al. 2008, 2010).

Genetic priors

Consider the networks

$$G_Y^1 : \; Y_1 \rightarrow Y_2 \rightarrow Y_3 \; , \quad G_Y^2 : \; Y_1 \leftarrow Y_2 \leftarrow Y_3 \; .$$

These Markov equivalent networks have the same likelihood, i.e.,

$$P(Y \mid G_Y^1) = P(Y \mid G_Y^2) \; .$$

If the phenotypes are associated with QTLs, we can use the results of the causality tests to compute prior probabilities for the network structures. If

$$\frac{P(G_Y^1)}{P(G_Y^2)} \neq 1 \; , \quad \text{then} \quad \frac{P(G_Y^1 \mid Y)}{P(G_Y^2 \mid Y)} = \frac{P(G_Y^1)}{P(G_Y^2)} \neq 1 \; ,$$

and we can use the posterior probability ratio to distinguish between the networks.

Augmenting the phenotype network with QTL nodes

By augmenting the phenotype network with a QTL node,

$$G^1 : \; Q \rightarrow Y_1 \rightarrow Y_2 \rightarrow Y_3 \; , \quad G^2 : \; Q \rightarrow Y_1 \leftarrow Y_2 \leftarrow Y_3 \; ,$$

we have that G^1 and G^2 have distinct sets of conditional independence relations:

$$Y_2 \perp\!\!\!\perp Q \mid Y_1 \; , \quad \text{on} \quad G^1$$
$$Y_2 \not\perp\!\!\!\perp Q \mid Y_1 \; , \quad \text{on} \quad G^2$$

Hence, G^1 and G^2 are no longer likelihood equivalent.

In the inferential approaches we address here we adopt this augmentation approach.

d-separation

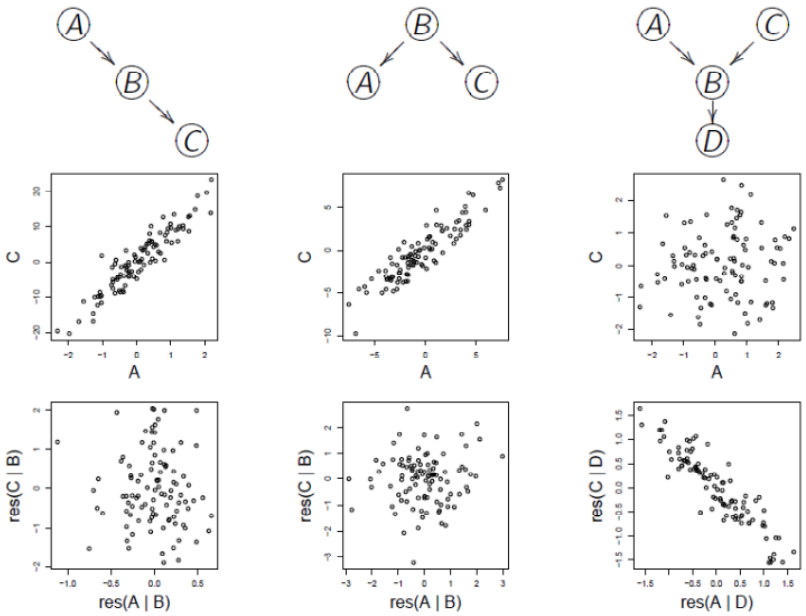
Graphical criterion to read out conditional independence relations from a DAG.

Definition (d-separation): A path p is said to be d-separated (or blocked) by a set of nodes Z if and only if

1. p contains a chain $i \to m \to j$ or a fork $i \leftarrow m \to j$ such that the middle node m is in Z, or
2. p contains an inverted fork (or collider) $i \to m \leftarrow j$ such that the middle node m is not in Z and such that no descendant of m is in Z.

A set Z is said to d-separate X from Y if and only if Z blocks every path from a node in X to a node in Y. X and Y are d-connected if they are not d-separated (Pearl, 1988, 2000).

d-separation

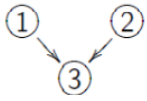


Simple graphical criterium to detect Markov equivalence

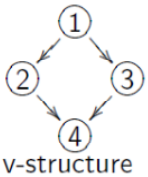
Detecting Markov equivalence: Two DAGs are Markov equivalent if and only if they have the same skeletons and the same set of v-structures. (Verma and Pearl 1990).

The **skeleton** of a causal graph is the undirected graph obtained by replacing its arrows by undirected edges.

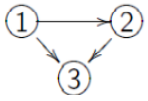
A **v-structure** is composed by two converging arrows whose tails are not connected by an arrow.



v-structure



v-structure



not a v-structure

Simple graphical criterium to detect Markov equivalence

DAG structures	skeletons	v-structures
$Y_1 \to Y_2 \to Y_3$	$Y_1 - Y_2 - Y_3$	$\emptyset$
$Y_1 \to Y_2 \leftarrow Y_3$	$Y_1 - Y_2 - Y_3$	$Y_1 \to Y_2 \leftarrow Y_3$
$Y_1 \leftarrow Y_2 \to Y_3$	$Y_1 - Y_2 - Y_3$	$\emptyset$

Extended DAG structures	skeletons	v-structures
$Q_1 \to Y_1 \to Y_2 \to Y_3$	$Q - Y_1 - Y_2 - Y_3$	$\emptyset$
$Q_1 \to Y_1 \leftarrow Y_2 \to Y_3$	$Q - Y_1 - Y_2 - Y_3$	$Q \to Y_1 \leftarrow Y_2$

Faithfulness assumption

Given a graph and a probability distribution associated with it, all the conditional independence relations spanned by a probability distribution must match the d-separation relations predicted from the graph structure (Spirtes et al. 2000).

Unfaithfulness example:

$$Y_1 = \epsilon_1 \,, \quad Y_2 = \beta_{21}\, Y_1 + \epsilon_2 \,, \quad Y_3 = \beta_{31}\, Y_1 + \beta_{32}\, Y_2 + \epsilon_3$$

$$\epsilon_k \sim N(0, \sigma_k^2) \,, \quad Cov(Y_1\,, Y_3) = (\beta_{31} + \beta_{32}\,\beta_{21})\,\sigma_1^2$$

$$\text{If } \beta_{31} = -\beta_{32}\beta_{21} \text{ then } Cov(Y_1\,, Y_3) = 0.$$

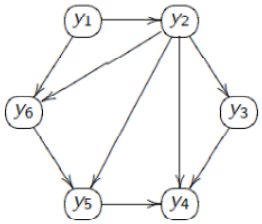
Although the data is generated from a, its probability distribution is faithful to b.

The PC skeleton algorithm

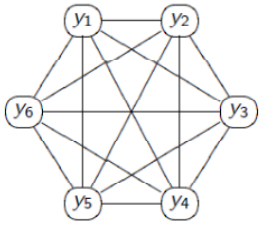
Infers the skeleton of the causal model (Spirtes et al. 1993).

PC skeleton algorithm

Suppose the true network describing the causal relationships between six transcripts is



The PC-algorithm starts with the complete undirected graph



and progressively eliminates edges based on conditional independence tests.

PC skeleton algorithm

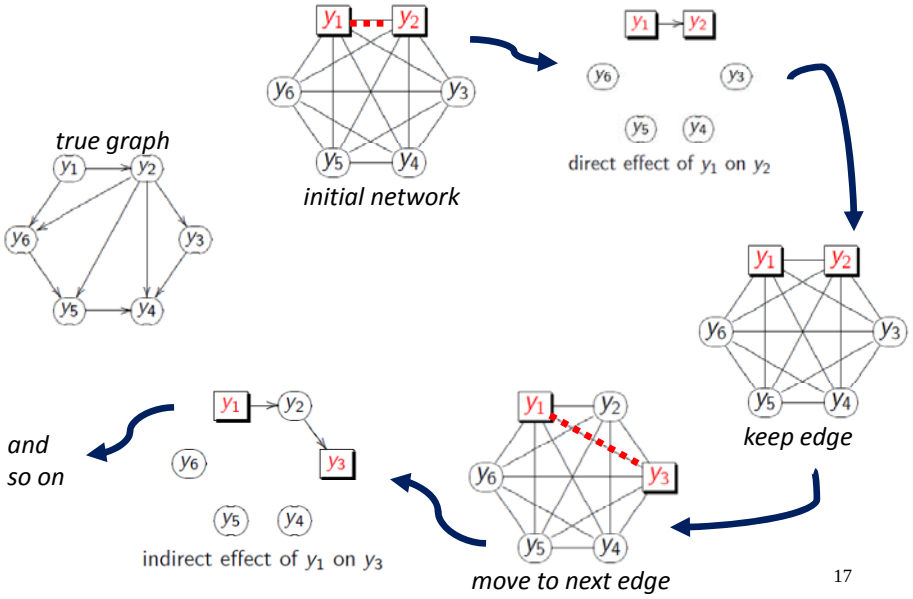
The algorithm performs several rounds of conditional independence tests of increasing order.

It starts with all zero order tests, then performs all first order, second order, and so on.

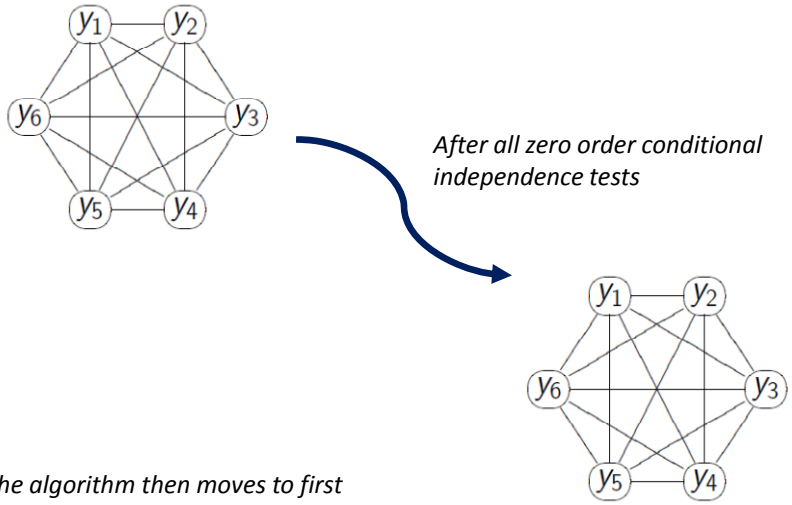
- Remark: in the Gaussian case zero partial correlation implies conditional independence, thus

$$i \perp\!\!\!\perp j \mid k \quad \Leftrightarrow \quad cor(i, j \mid k) = 0 \quad \Rightarrow \quad \text{drop } (i, j) \text{ edge}$$

PC algorithm - zero order



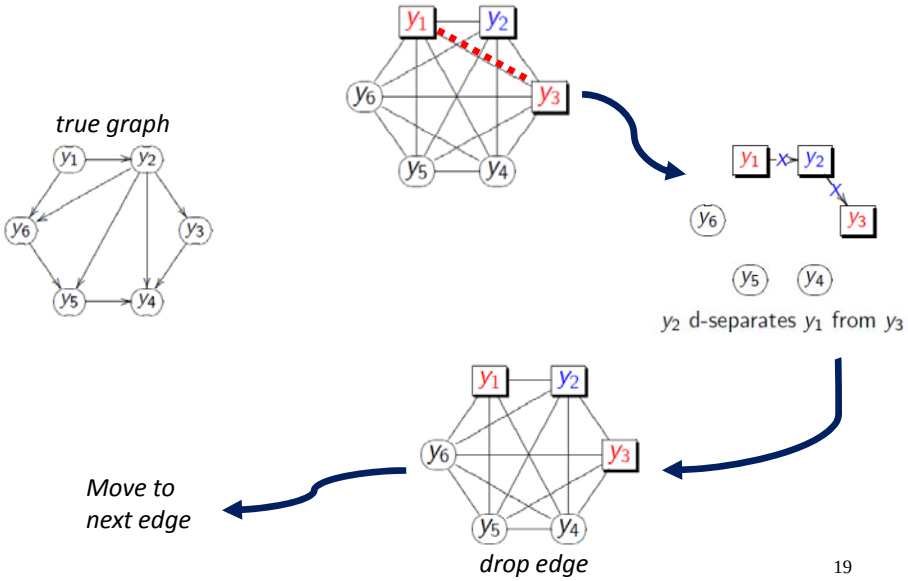
PC algorithm - zero order



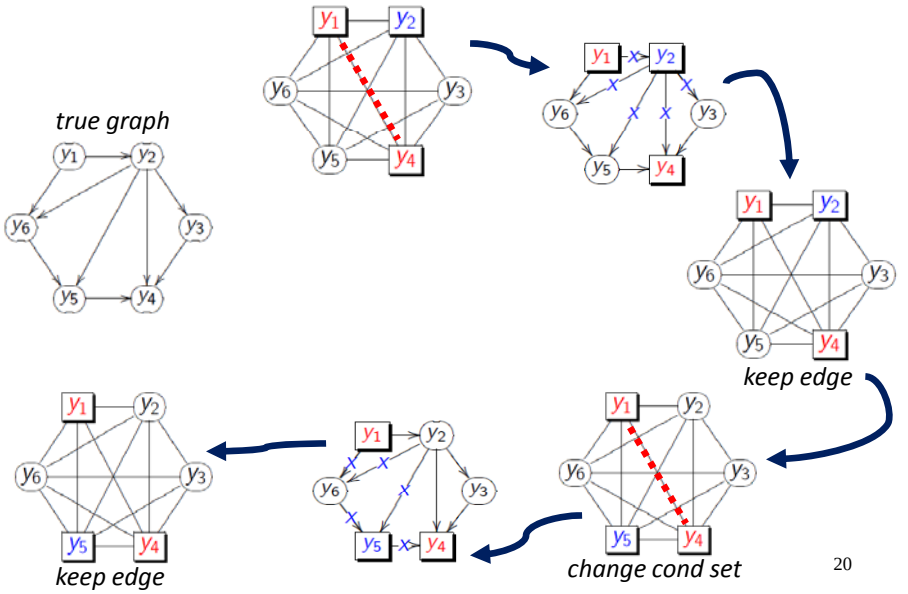
After all zero order conditional independence tests

The algorithm then moves to first order conditional independence tests.

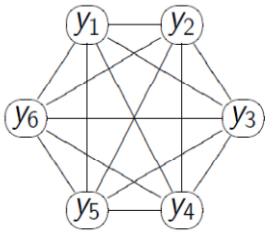
PC algorithm - first order



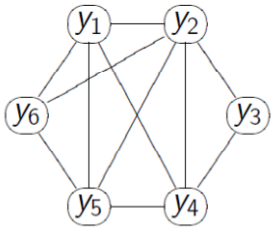
PC algorithm - first order



PC algorithm - first order



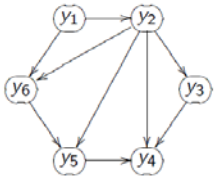
After all first order conditional independence tests.

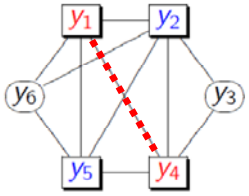


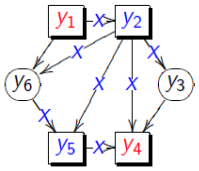
The algorithm then moves to second order conditional independence tests.

PC algorithm - second order

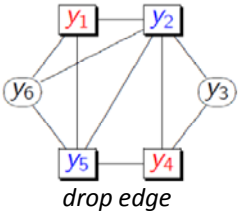
true graph







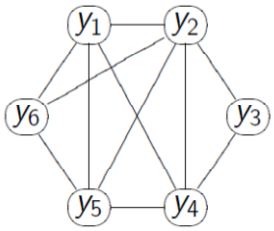
(y_2, y_5) d-separate y_1 from y_4



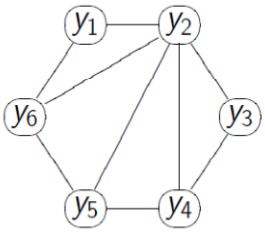
drop edge

move to next edge

PC algorithm - second order



After all second order conditional independence tests



Then the algorithm moves to third order, fourth order …

Edge orientarion with the QDG algorithm

Edge orientation

We perform model selection using a direction LOD score

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

where $f()$ represents the predictive density, that is, the sampling model with parameters replaced by the corresponding maximum likelihood estimates.

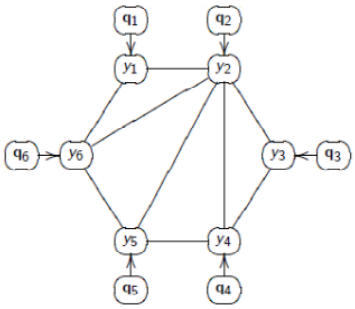
QDG algorithm

The QTL-driven Dependency Graph algorithm is composed of 7 steps:

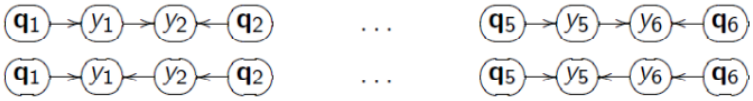
1. Get the causal skeleton (with the PC skeleton algorithm).
2. Use QTLs to orient the edges in the skeleton.
3. Choose a random ordering of edges, and
4. Recompute orientations incorporating causal phenotypes in the models (update the causal model according to changes in directions).
5. Repeat 4 iteratively until no more edges change direction (the resulting graph is one solution).
6. Repeat steps 3, 4, and 5 many times and store all different solutions.
7. Score all solutions and select the graph with best score.

QDG algorithm - step 2

Now suppose that for each transcript we have a set of e-QTLs

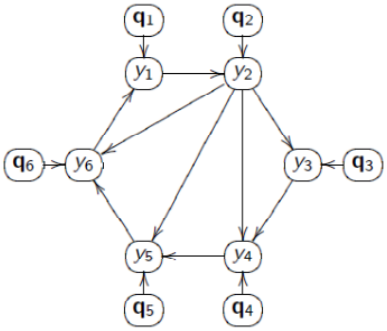


Given the QTLs we can distinguish causal direction:

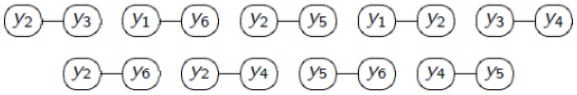


QDG algorithm - steps 2 and 3

First estimate of the causal model, DG_0, (using only QTLs to infer causal direction)

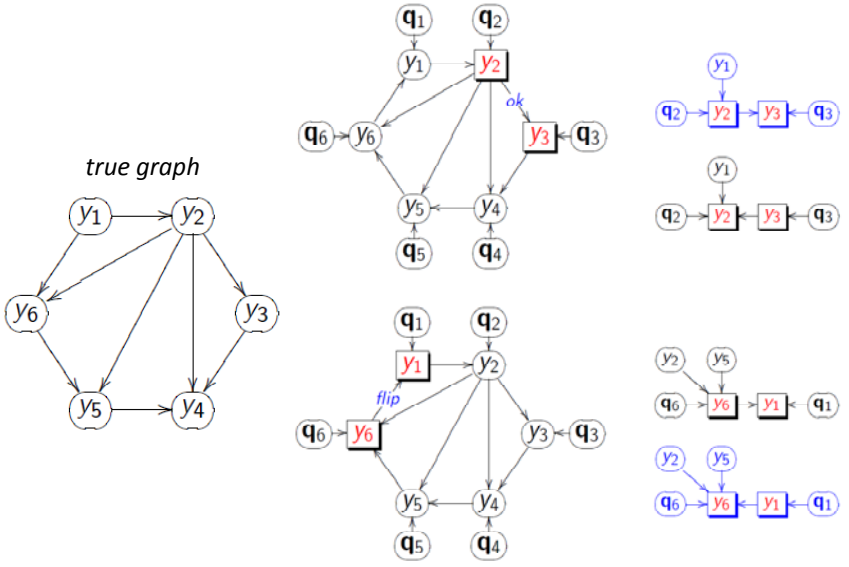


In step 3 we randomly choose an ordering of all edges in DG_0. Say,



In step 4 we recompute the directions including other transcripts as covariates in the models (following the above ordering).

QDG algorithm - step 4



QDG algorithm - steps 5, 6, and 7

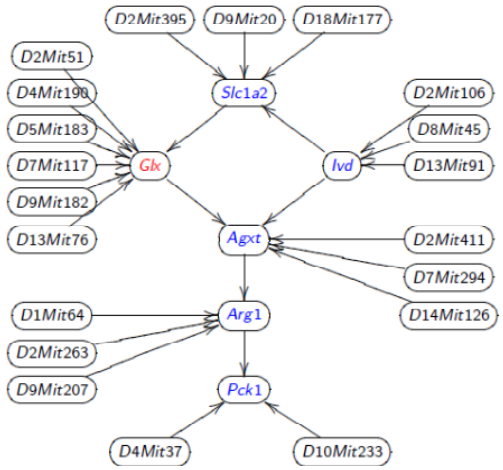
Step 5: repeat 4 iteratively until no more edges change direction (the resulting graph is one solution).

Step 6: repeat the process starting from different random orderings several times, and store all different solutions.

Step 7: score all solutions and select the graph with best score.

Real data example

Network of metabolites and transcripts involved in liver metabolism.



Four out of six predictions were validated experimentally (Ferrara et al. 2008).

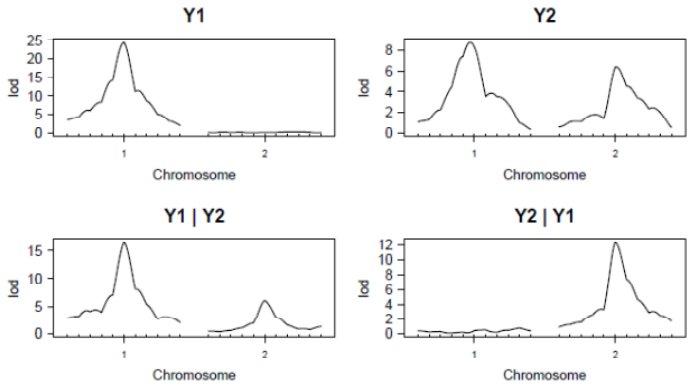
QTLnet algorithm

QTLnet algorithm

- Perform joint inference of the causal phenotype network and the associated genetic architecture.
- The genetic architecture is inferred conditional on the phenotype network.
- Because the phenotype network structure is itself unknown, the algorithm iterates between updating the network structure and genetic architecture using a Markov chain Monte Carlo (MCMC) approach.
- QTLnet corresponds to a mixed Bayesian network with continuous and discrete nodes representing phenotypes and QTLs, respectively.

QTL mapping conditional on the pheno net structure

We simulated data from the model $Q_1 \rightarrow Y_1 \rightarrow Y_2 \leftarrow Q_2$ with Q_1 located on chr 1, and Q_2 on chr 2.



- Y_2 maps indirectly to Q_1 (top right), but Y_1 d-separates Y_2 and Q_1 (bottom right).
- Y_1 is marginally independent from Q_2 (top left), but conditional on Y_2 became associated (bottom left).

QTLnet algorithm - MCMC steps

1. Propose a new phenotype network, $\mathcal{M}_{new}$, by adding, deleting or reversing (with parent orphaning) an edge.
2. Recompute the genetic architecture (only for the phenotypes y_t whose parent set, $pa(y_t)$, has changed).
3. Compute the marginal likelihood $p(\mathbf{y} \mid \mathbf{q}, \mathcal{M}_{new})$.
4. Accept or reject the new phenotype network and QTLs according to the Metropolis-Hastings acceptance probability:

$$\alpha = min\left\{1, \frac{p(\mathbf{y} \mid \mathbf{q}, \mathcal{M}_{new})\, p(\mathcal{M}_{new})}{p(\mathbf{y} \mid \mathbf{q}, \mathcal{M}_{old})\, p(\mathcal{M}_{old})} \frac{q(\mathcal{M}_{old} \mid \mathcal{M}_{new})}{q(\mathcal{M}_{new} \mid \mathcal{M}_{old})}\right\} .$$

QTLnet algorithm

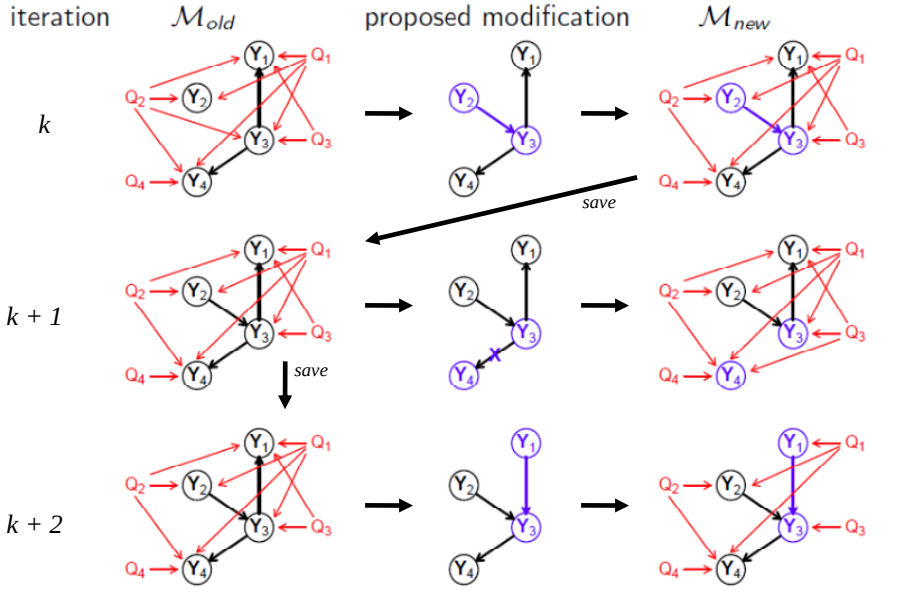
We approximate the Bayes factor comparing old and new models by

$$\frac{p(\mathbf{y} \mid \mathbf{q}, \mathcal{M}_{new})}{p(\mathbf{y} \mid \mathbf{q}, \mathcal{M}_{old})} \approx \exp\left\{-\frac{1}{2}(BIC_{\mathcal{M}_{new}} - BIC_{\mathcal{M}_{old}})\right\} ,$$

and adopt $p(\mathcal{M}_{new})/p(\mathcal{M}_{old}) = 1$. The proposal distribution ratio is computed as

$$\frac{q(\mathcal{M}_{old} \mid \mathcal{M}_{new})}{q(\mathcal{M}_{new} \mid \mathcal{M}_{old})} = \frac{\#\text{ of DAGs that can be reached from } \mathcal{M}_{old}}{\#\text{ of DAGs that can be reached from } \mathcal{M}_{new}} .$$

QTLnet algorithm

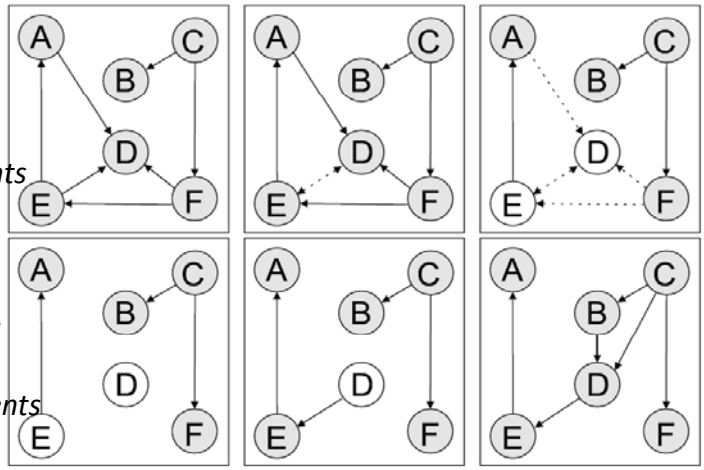


37

Neighborhood edge reversal

select edge
drop edge
identify parents

orphan nodes
reverse edge
find new parents

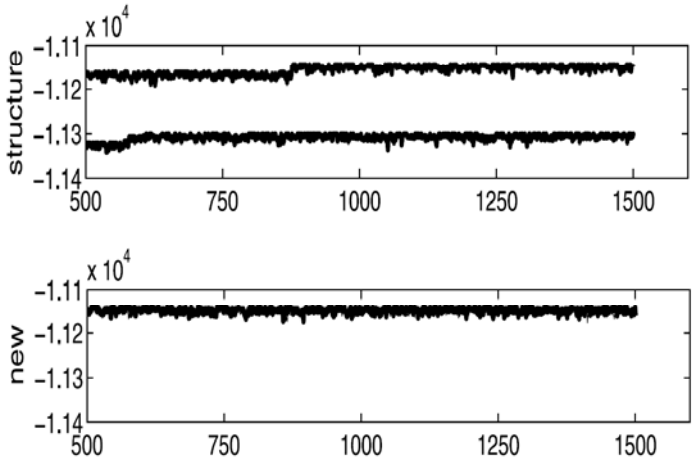


from Grzegorczyk and Husmier (2008)

38

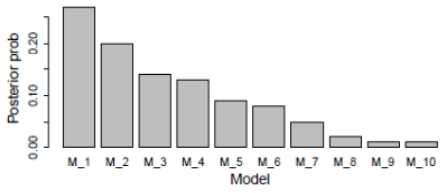
Neighborhood edge reversal

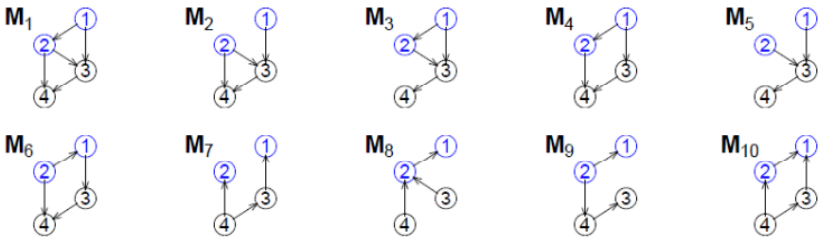
Trace plots of the logarithmic scores of the DAGs after the burn-in phase.



from Grzegorczyk and Husmier (2008)

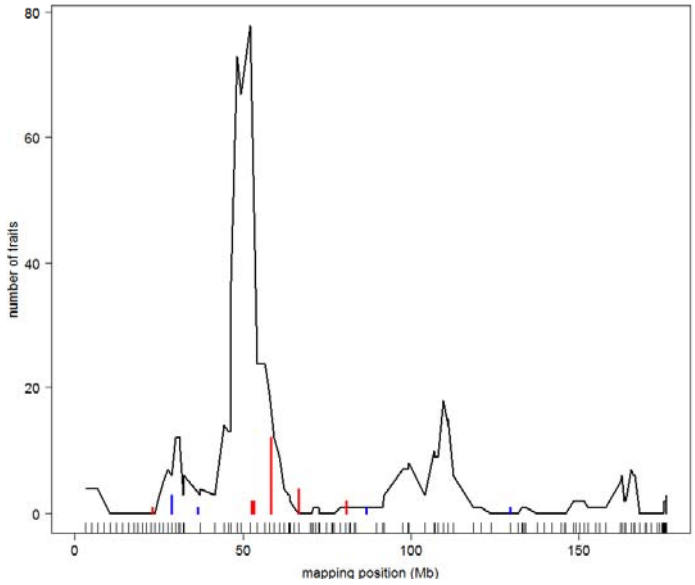
Bayesian model averaging





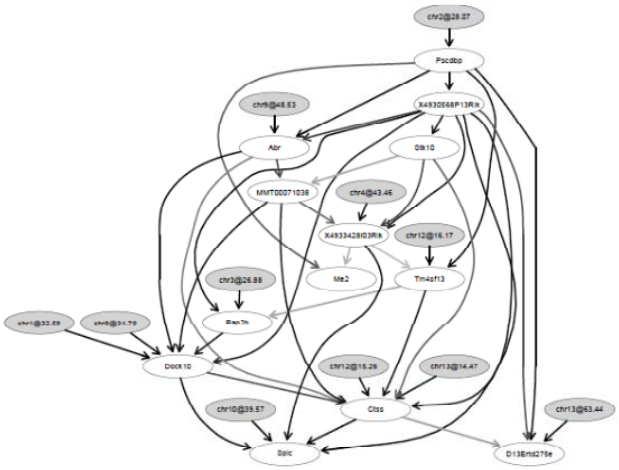
$Pr(Y_1 \rightarrow Y_2) = Pr(M_1) + Pr(M_3) + Pr(M_4) = 0.54$
$Pr(Y_1 \ldots Y_2) = Pr(M_2) + Pr(M_5) + Pr(M_7) = 0.34$
$Pr(Y_1 \leftarrow Y_2) = Pr(M_6) + Pr(M_8) + Pr(M_9) + Pr(M_{10}) = 0.12$

BxH ApoE-/- chr 2: causal architecture



41

BxH ApoE-/- chr 2: causal network for transcription factor Pscdbp



42

Scaling up to larger networks

- Reduce complexity of graphs
  - restrict number of causal edges into each node

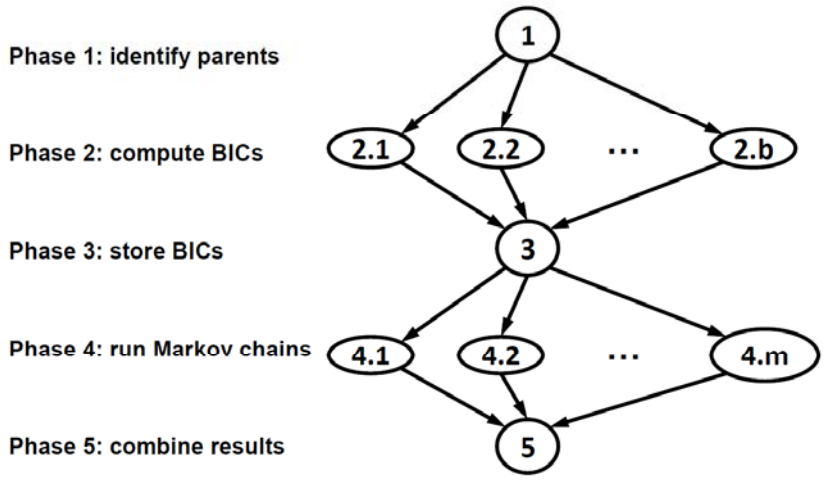
BIC computations by maximum number of parents

#	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

(limit complexity by allowing only 3-4 parents)

- make task parallel: run on many machines
  - pre-compute BIC scores
  - run multiple parallel Markov chains

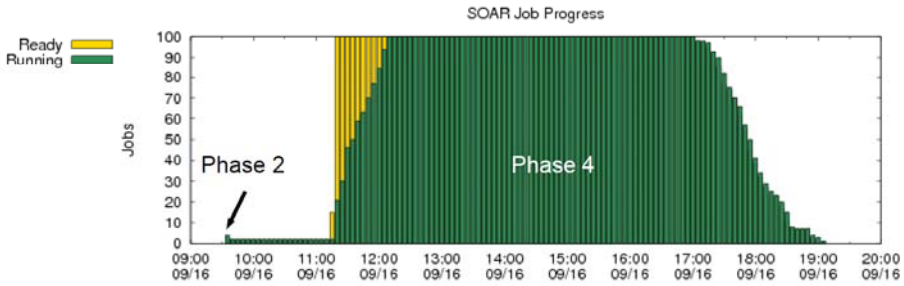
Parallel phases for larger projects



Parallel implementation

R/qtlnet available at CRAN

- 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



Final remarks

Potential issues

- Steady state (static) measures may not reflect dynamic processes (Przytycha and Kim 2010).
- Population-based estimates (from a sample of individuals) may not reflect processes within an individual.

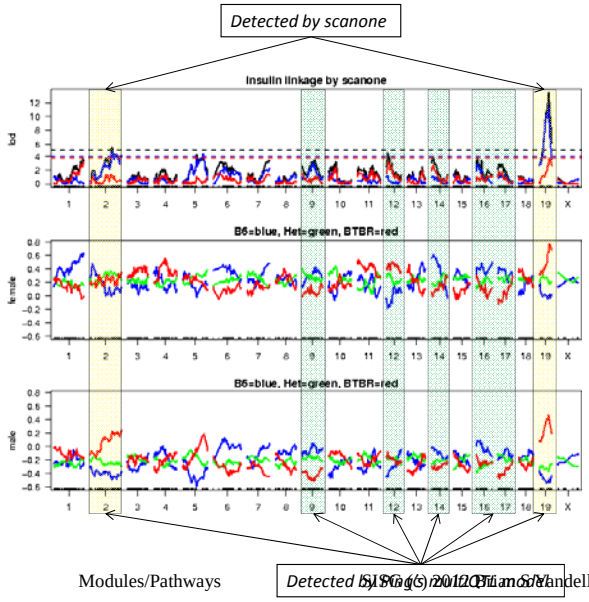
References

1. Chaibub Neto et al. (2008) *Genetics* **179**: 1089-1100.
2. Chaibub Neto et al. (2010) *Annals of Applied Statistics* **4**: 320-339.
3. Ferrara et al. (2008) *Plos Genetics* **4**: e1000034.
4. Grzegorczyk and Husmier (2008) *Machine Learning* **71**: 265-305.
5. Pearl (1988) *Probabilistic reasoning in intelligent systems* Morgan Kauffman.
6. Pearl (2000) *Causality: models, reasoning and inference* Cambridge U Press
7. Przytycha and Kim (2010) *BMC Biology* **8**: 48.
8. Spirtes et al. (2000) *Causation, prediction and search* MIT Press.
9. Wright (1921) *Journal of Agricultural Research* **20**: 557-585.
10. Verma and Pearl (1990) In *Readings in uncertain reasoning* Morgan Kauffmann
11. Zhu et al. (2007) *Plos Computational Biology* **3**: e69.

Expression Modules

Brian S. Yandell (with slides from
Steve Horvath, UCLA, and
Mark Keller, UW-Madison)

Weighted models for insulin

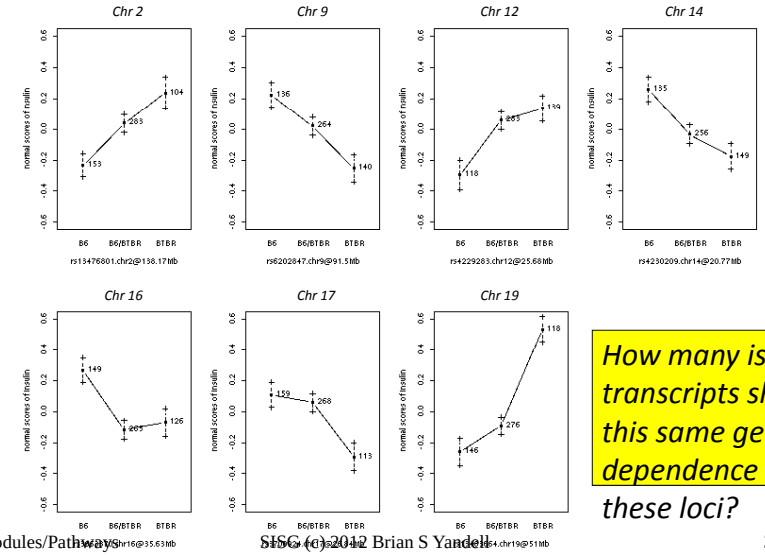


 # transcripts that match weighted insulin model in each of 4 tissues:

tissue	# transcripts
Islet	1984
Adipose	605
Liver	485
Gastroc	404

Ping Wang

insulin main effects



How many islet transcripts show this same genetic dependence at these loci?

Expression Networks
Zhang & Horvath (2005)

www.genetics.ucla/edu/labs/horvath/CoexpressionNetwork

- organize expression traits using correlation

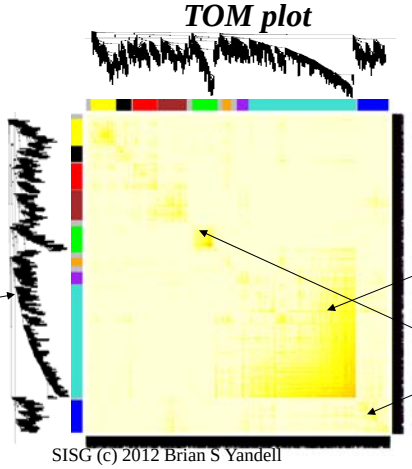
- adjacency $a_{ij} = |cor(x_i, x_j)|^{\beta}, \beta = 6$
- connectivity $k_i = \text{sum}_l(a_{il})$
- topological overlap $TOM_{ij} = \dfrac{a_{ij} + \text{sum}_l(a_{il}a_{jl})}{1 - a_{ij} + \min(k_i, k_j)}$

Using the topological overlap matrix (TOM) to cluster genes

– modules correspond to branches of the dendrogram

Genes correspond to rows and columns

TOM plot



Hierarchical clustering dendrogram

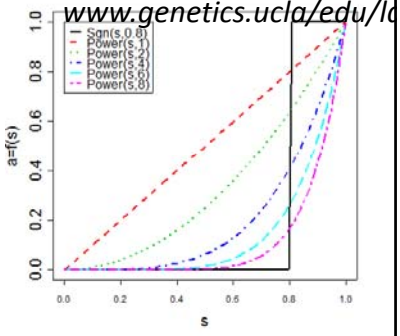
TOM matrix

Module: Correspond to branches

module traits highly correlated

- adjacency attenuates correlation
- can separate positive, nega
- summarize module
  - eigengene
  - weighted average of traits
- relate module
  - to clinical traits
  - map eigengene

www.genetics.ucla/edu/lc

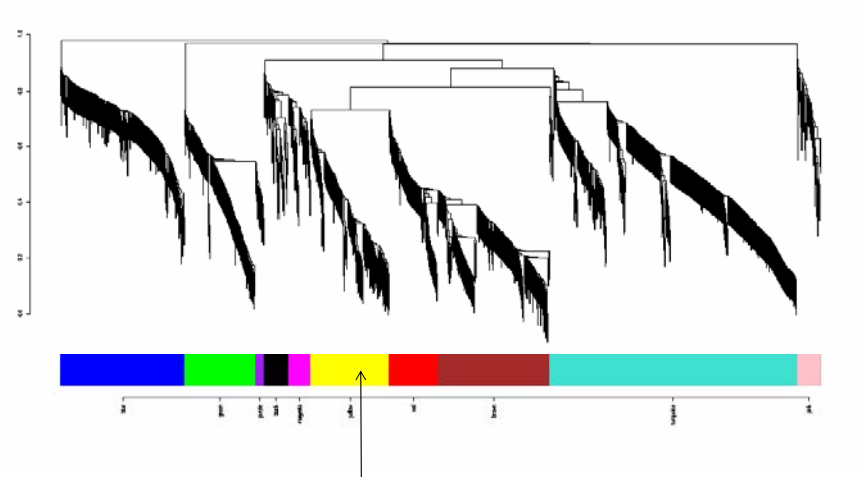


advantages of Horvath modules

- **emphasize modules (pathways) instead of individual genes**
  - Greatly alleviates the problem of multiple comparisons
  - ~20 module comparisons versus 1000s of gene comparisons
- intramodular connectivity k_i finds key drivers (hub genes)
  - quantifies module membership (centrality)
  - highly connected genes have an increased chance of validation
- module definition is based on gene expression data
  - no prior pathway information is used for module definition
  - two modules (eigengenes) can be highly correlated
- unified approach for relating variables
  - compare data sets on same mathematical footing
- scale-free: zoom in and see similar structure

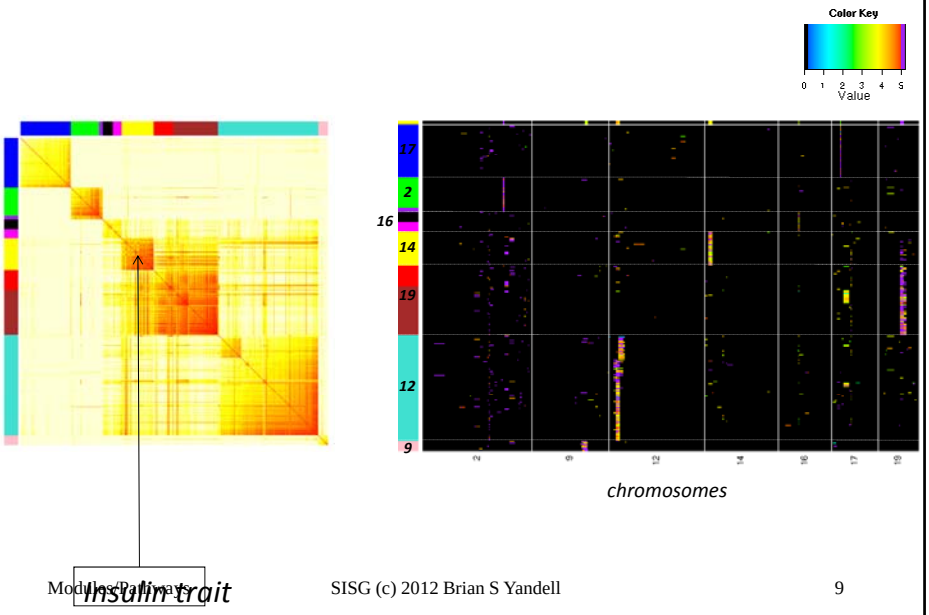
Ping Wang

modules for 1984 transcripts with similar genetic architecture as insulin



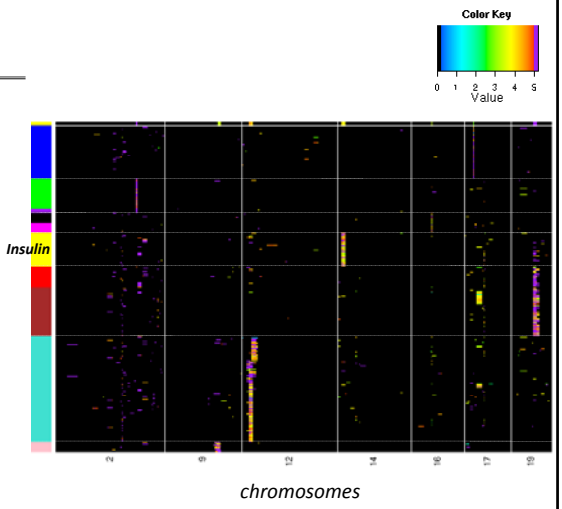
contains the insulin trait

Islet – modules



Islet – enrichment for modules

Module	Pvalue	Qvalue	Count	Size	Term
BLUE	0.0005	0.0463	30	1068	biosynthetic process
	0.0006	0.0470	18	511	cellular lipid metabolic process
	0.0009	0.0507	11	241	lipid biosynthetic process
	0.0012	0.0593	19	590	lipid metabolic process
GREEN	0.0008	0.0457	4	76	phosphate transport
	0.0055	0.0970	2	20	intermediate filament-based process
	0.0056	0.0970	10	707	ion transport
PURPLE	0.0011	0.0165	7	2769	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
BLACK	0.0078	0.0138	2	68	sensory perception of sound
MAGENTA	**2.54E-05**	**0.0011**	**7**	**313**	**cell cycle process**
	0.0001	**0.0040**	**5**	**179**	**microtubule-based process**
	0.0004	**0.0040**	**5**	**225**	**mitotic cell cycle**
	0.0005	**0.0040**	**5**	**228**	**M phase**
	0.0006	**0.0040**	**5**	**239**	**cell division**
	0.0009	**0.0041**	**5**	**266**	**cell cycle phase**
	0.0011	**0.0041**	**4**	**162**	**mitosis**
	0.0012	**0.0041**	**4**	**163**	**M phase of mitotic cell cycle**
YELLOW	0.0026	0.0675	7	281	cell projection organization and biogenesis
	0.0026	0.0675	7	281	cell part morphogenesis
	0.0026	0.0675	7	281	cell projection morphogenesis
RED	0.0017	0.0619	2	13	steroid hormone receptor signaling pathway
	0.0026	0.0619	5	200	reproductive process
BROWN	0.0057	0.1442	4	96	response to pheromone
TURQUOISE	0.0002	0.0830	17	279	enzyme linked receptor protein signaling pathway
	0.0003	0.0830	10	115	morphogenesis of an epithelium
	0.0003	0.0830	7	57	morphogenesis of embryonic epithelium
	0.0004	0.0830	40	1021	anatomical structure morphogenesis
PINK	0.0004	0.0608	2	14	vesicle organization and biogenesis
	0.0092	0.0612	4	384	regulation of apoptosis

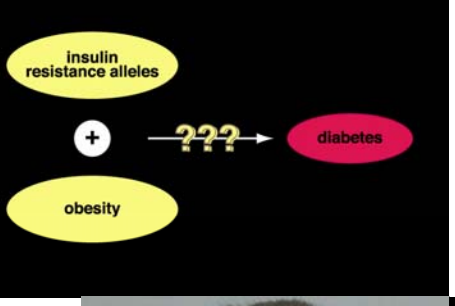


www.geneontology.org

- ontologies
  - Cellular component (GOCC)
  - Biological process (GOBP)
  - Molecular function (GOMF)
- hierarchy of classification
  - general to specific
  - based on extensive literature search, predictions
- prone to errors, historical inaccuracies

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



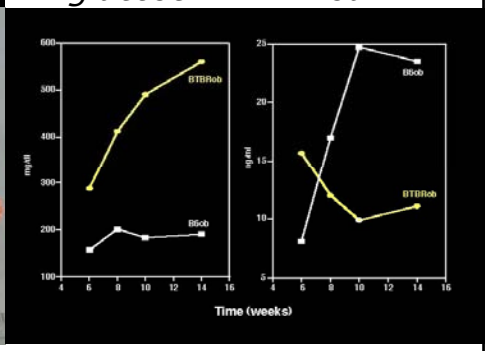
BTBR mouse is insulin resistant

B6 is not

make both obese…

Alan Attie
Biochemistr

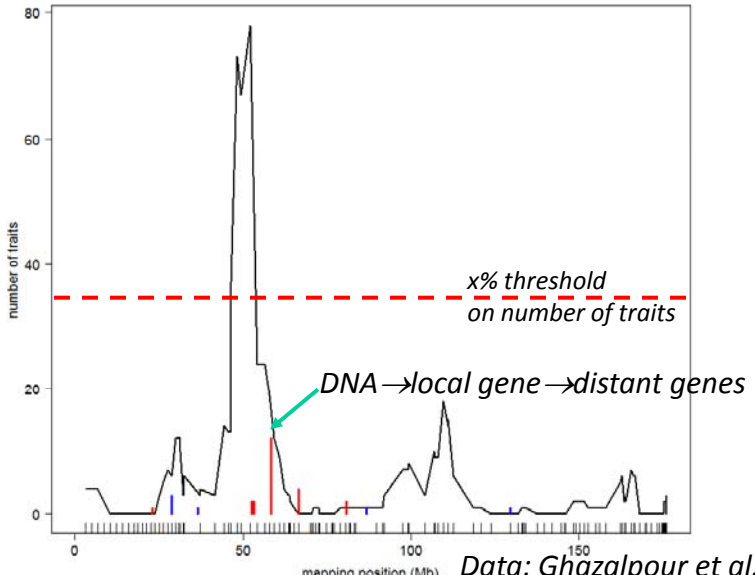
glucose *insulin*



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

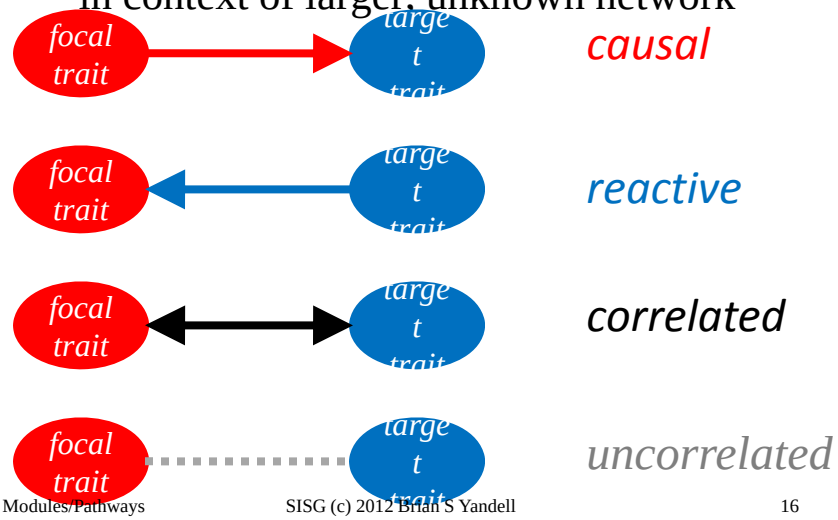
BxH ApoE-/- chr 2: hotspot



Data: Ghazalpour et al.(2006) PLo

causal model selection choices

in context of larger, unknown network



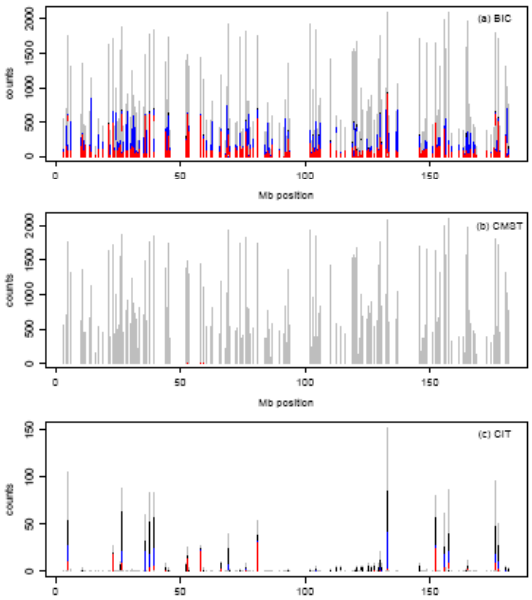
causal architecture references

- 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. (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.

(M_1) (M_2) (M_3)

$Q_1 \rightarrow Y_1 \rightarrow Y_2 \leftarrow Q_{2|1}$ $Q_{1|2} \rightarrow Y_1 \leftarrow Y_2 \leftarrow Q_2$ $Q_1 \rightarrow Y_1 \quad Y_2 \leftarrow Q_2$

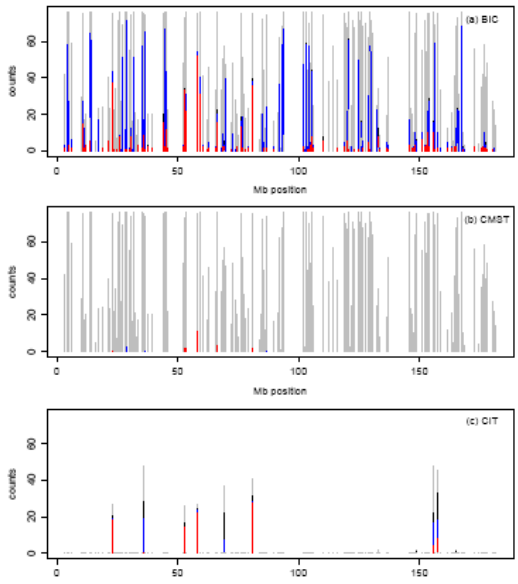


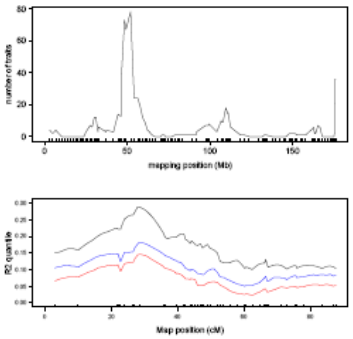
BxH ApoE-/- study
Ghazalpour et al. (2008)
PLoS Genetics

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.

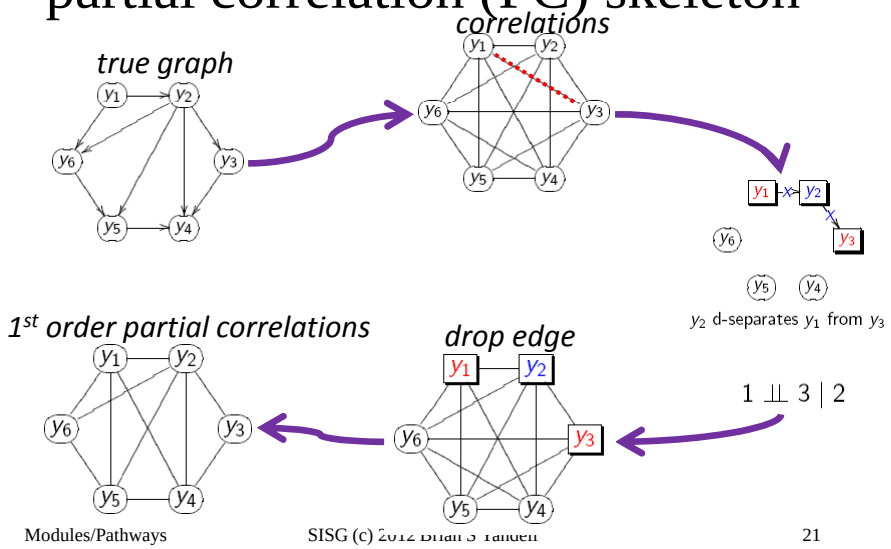
This collection of traits enriches for "immune system process".

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

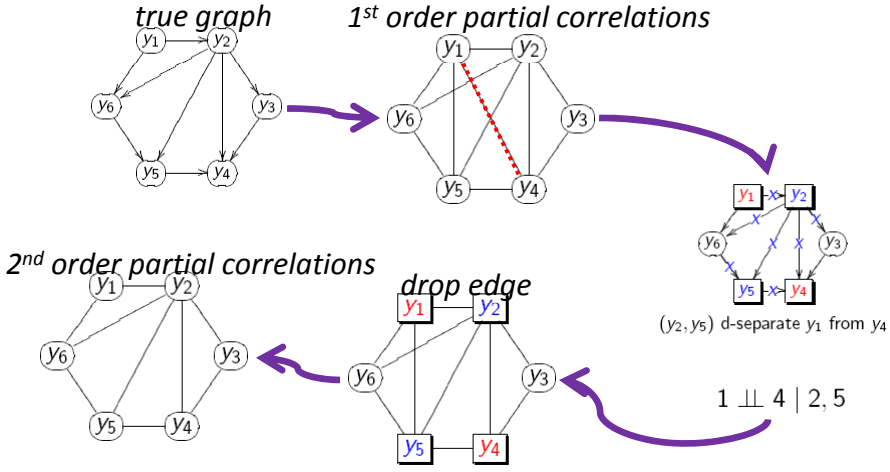
QTL-driven directed graphs

- 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]

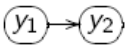
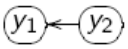
partial correlation (PC) skeleton



partial correlation (PC) skeleton

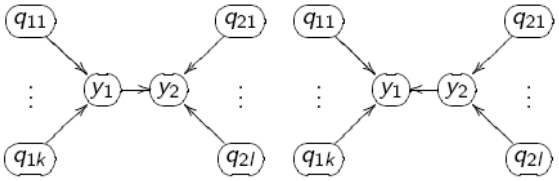


edge direction: which is causal?

M_1 :  $M2$: 

the above models are likelihood equivalent,

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

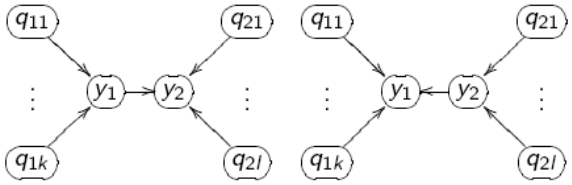


not likelihood equivalent *due to QTL*

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

test edge direction using LOD score

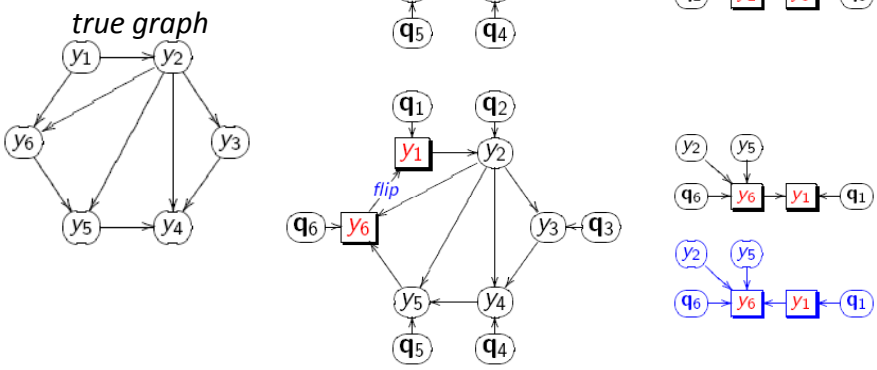
$$LOD = \log_{10}\left\{\frac{\prod_{i=1}^{n} f(y_{1i} \mid \mathbf{q}_{1i})f(y_{2i} \mid y_{1i}, \mathbf{q}_{2i})}{\prod_{i=1}^{n} f(y_{2i} \mid \mathbf{q}_{2i})f(y_{1i} \mid y_{2i}, \mathbf{q}_{1i})}\right\}$$



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



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

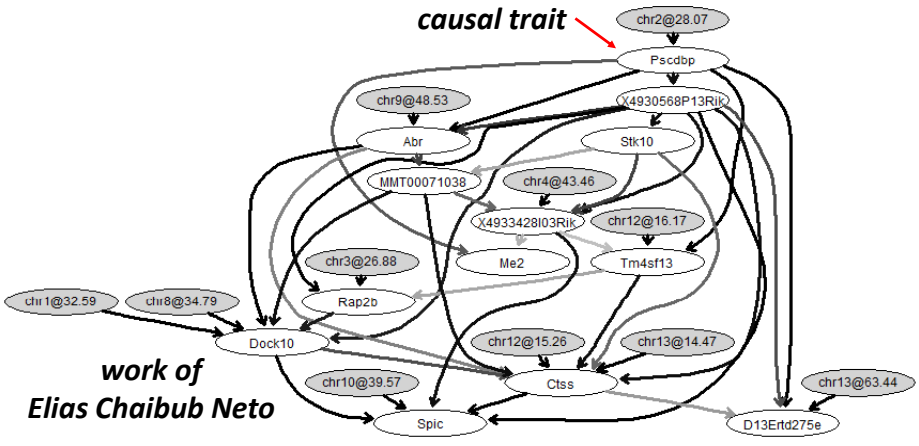
missing data method: MCMC

- 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
  - $\Pr(Y \mid G, Q)$
- accept or reject new network and QTL
  - usual Metropolis-Hastings idea

BxH ApoE-/- causal network for transcription factor Pscdbp

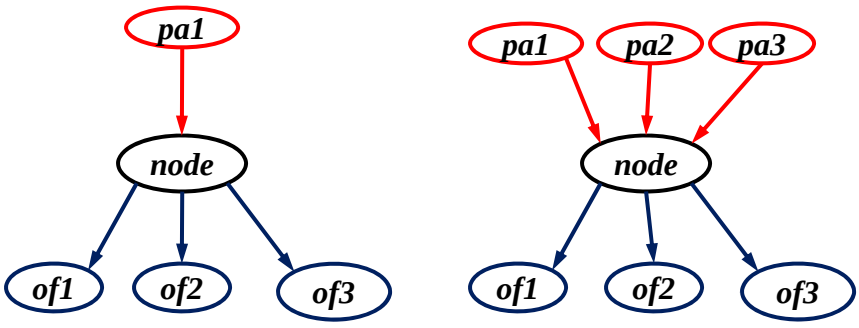


Data: Ghazalpour et al.(2006) PLoS Gen

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

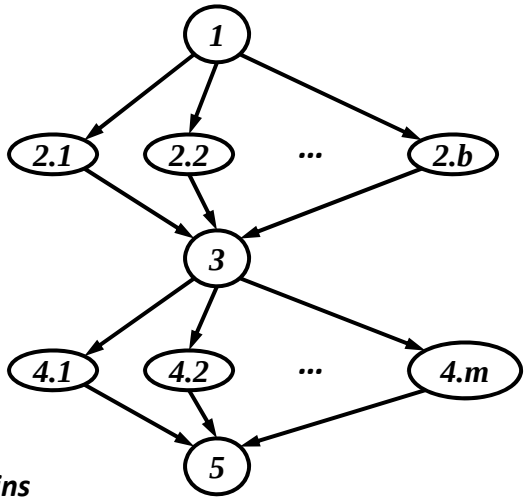
Phase 2: compute BICs

BIC = LOD – penalty

all possible parents to all nodes

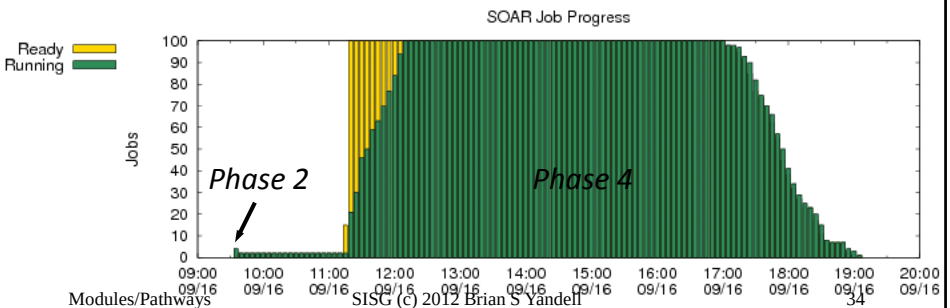
Phase 3: store BICs

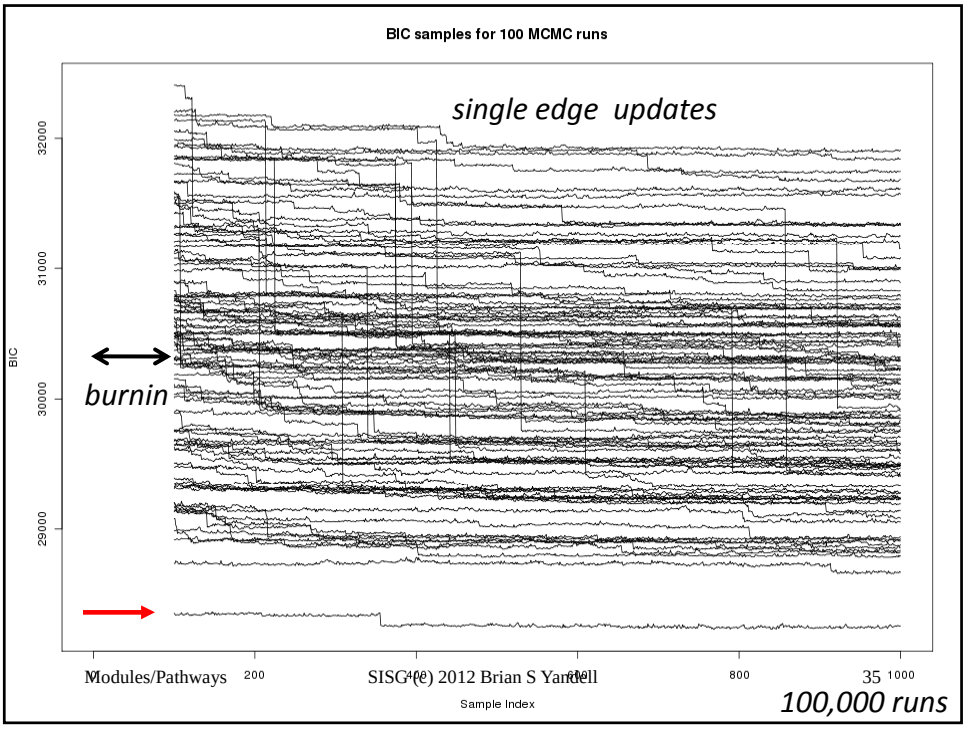
Phase 4: run Markov chains



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

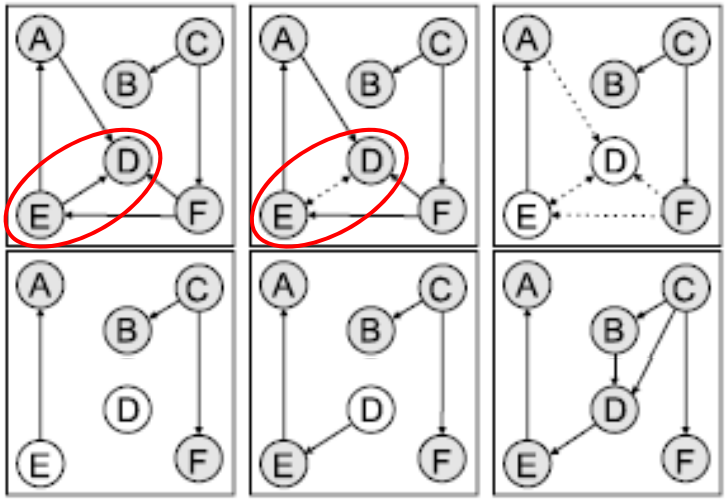




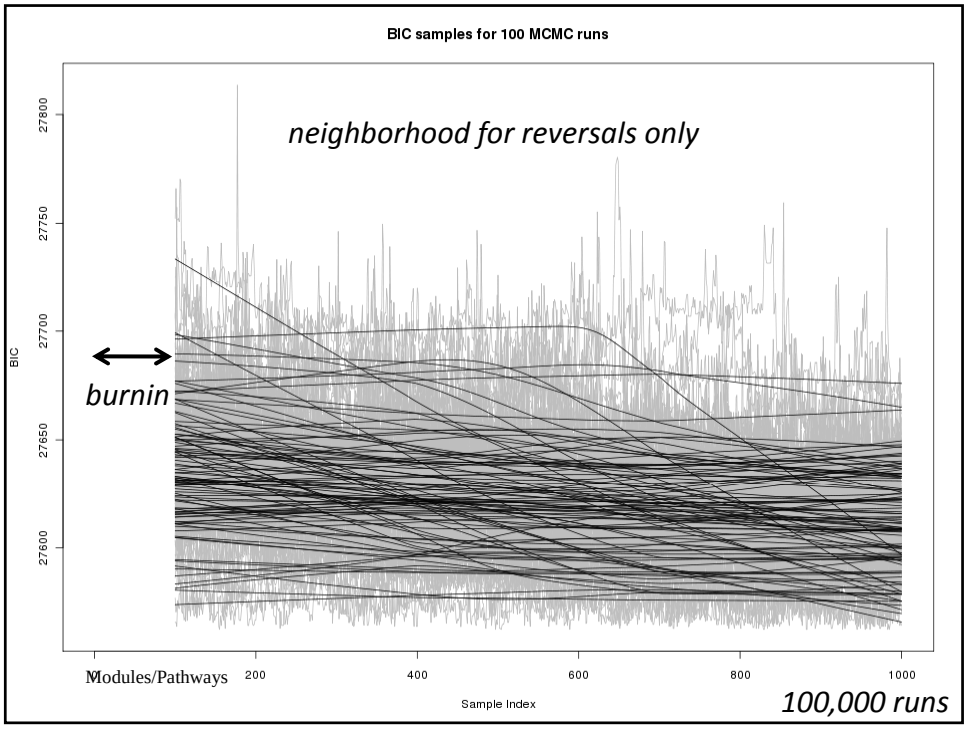
neighborhood edge reversal

select edge
drop edge
identify parents

orphan nodes
reverse edge
find new parents



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



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_Y from biological knowledge B
  - $\Pr(G_Y \mid B, W) = \exp(- W * |B - G_Y|) /$ constant
  - B = prob of edge given TF, PPI, KO database
    - derived using previous experiments, papers, etc.
  - G_Y = 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
  - $\Pr(G|B,W,Q) = \Pr(G_Y|B,W) \Pr(G_{Q \to Y}|Q)$
    - $\Pr(G_Y|B,W)$ relates graph to biological info
    - $\Pr(G_{Q \to Y}|Q)$ 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*

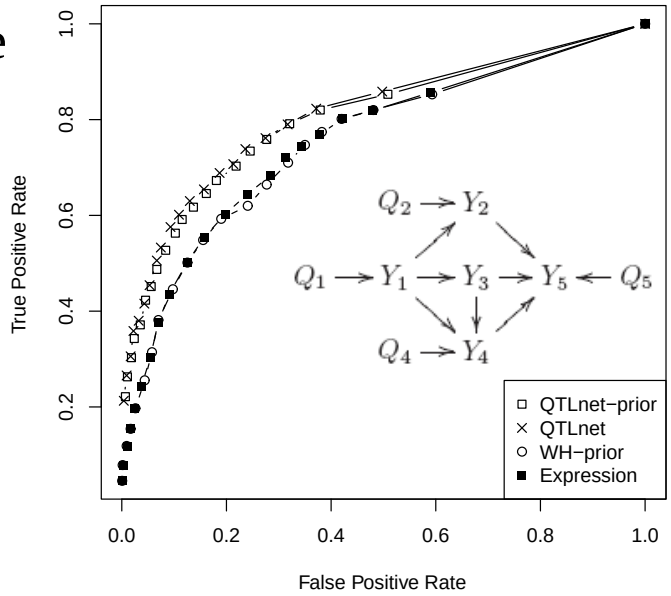
MCMC with pathway information

- 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

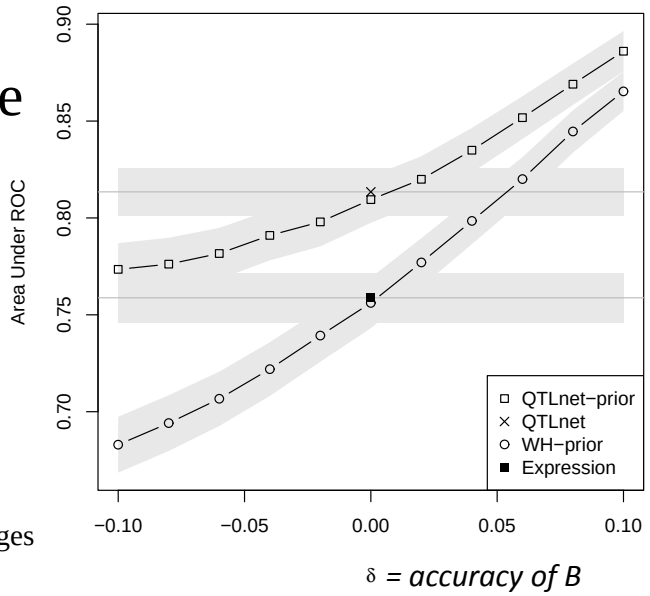
closed = phenotypes only



integrated ROC curve

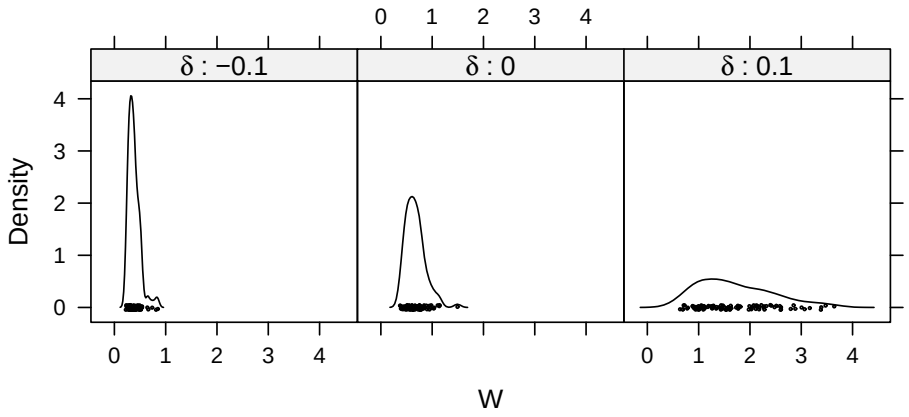
2x2: genetics pathways

probability classifier ranks true > false edges



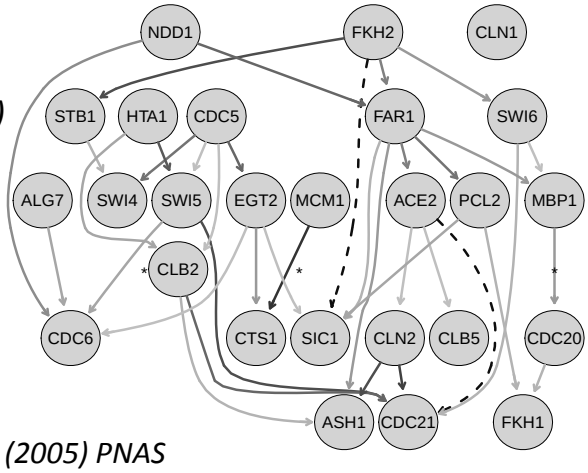
weight on biological knowledge

incorrect *non-informative* *correct*



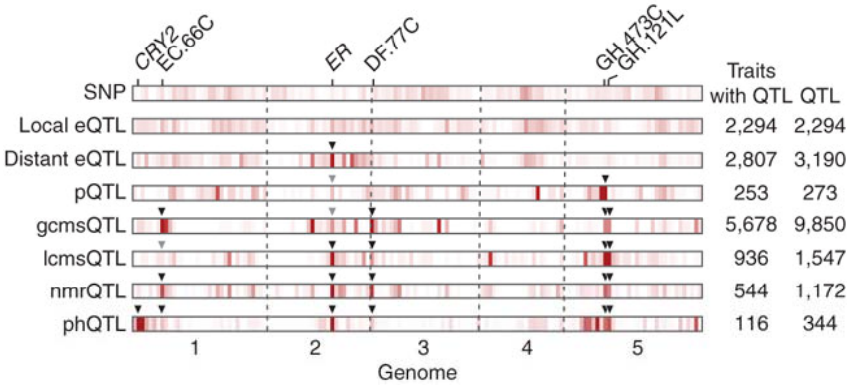
yeast data—partial success

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



Data: Brem, Kruglyak (2005) PNAS

phenotypic buffering of molecular QTL

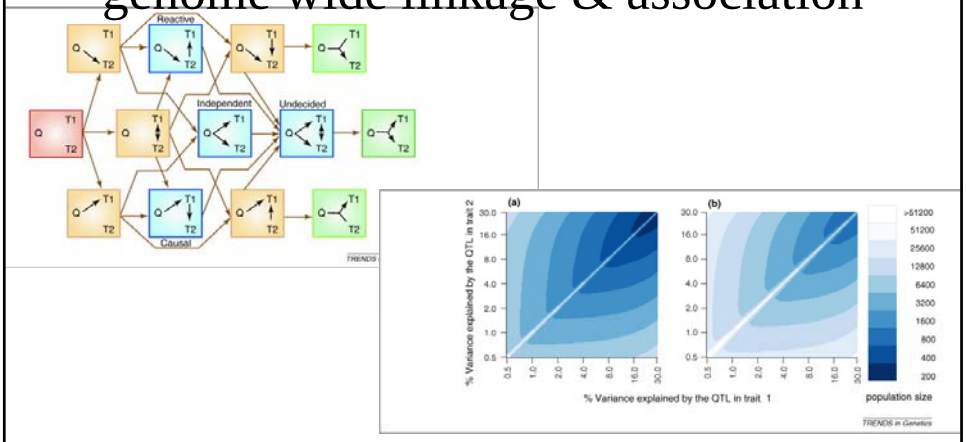


Fu et al. Jansen (2009 Nature Genetics)

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→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

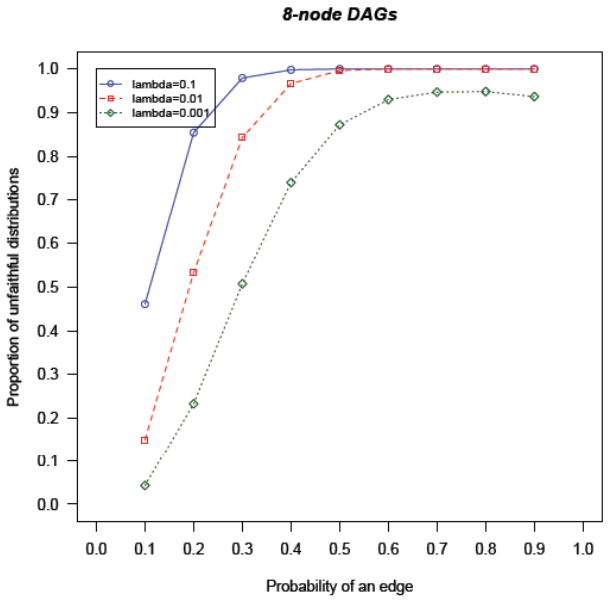
$\lambda = \min|\text{cor}(Y_i, Y_j)|$

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

d=max degree

p=# nodes

n=sample size



Uhler, Raskutti, Buhlmann, Yu (2012 in p

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

Computational Infrastructure for Systems Genetics Analysis

Brian Yandell, UW-Madison

high-throughput analysis of systems data
enable biologists & analysts to share tools

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

- UW-Madison
  - Alan Attie
  - Christina Kendziorski
  - Karl Broman
  - Mark Keller
  - Andrew Broman
  - Aimee Broman
  - YounJeong Choi
  - Elias Chaibub Neto
  - Jee Young Moon
  - John Dawson
  - Ping Wang
  - NIH Grants DK58037, DK66369, GM74244, GM69430 , EY18869
- Jackson Labs (HTDAS)
  - Gary Churchill
  - Ricardo Verdugo
  - Keith Sheppard
- UC-Denver (PhenoGen)
  - Boris Tabakoff
  - Cheryl Hornbaker
  - Laura Saba
  - Paula Hoffman
- Labkey Software
  - Mark Igra
- U Groningen (XGA)
  - Ritsert Jansen
  - Morris Swertz
  - Pjotr Pins
  - Danny Arends
- Broad Institute
  - Jill Mesirov
  - Michael Reich

experimental context

- B6 x BTBR obese mouse cross
  - model for diabetes and obesity
  - 500+ mice from intercross (F2)
  - collaboration with Rosetta/Merck
- genotypes
  - 5K SNP Affymetrix mouse chip
  - care in curating genotypes! (map version, errors, …)
- phenotypes
  - clinical phenotypes (>100 / mouse)
  - gene expression traits (>40,000 / mouse / tissue)
  - other molecular phenotypes

how does one filter traits?

- want to reduce to “manageable” set
  - 10/100/1000: depends on needs/tools
  - How many can the biologist handle?
- how can we create such sets?
  - data-driven procedures
    - correlation-based modules
      - Zhang & Horvath 2005 *SAGMB*, Keller et al. 2008 *Genome Res*
      - Li et al. 2006 *Hum Mol Gen*
    - mapping-based focus on genome region
  - function-driven selection with database tools
    - GO, KEGG, etc
    - Incomplete knowledge leads to bias
  - random sample

why build Web eQTL tools?

- common storage/maintainence of data
  - one well-curated copy
  - central repository
  - reduce errors, ensure analysis on same data
- automate commonly used methods
  - biologist gets immediate feedback
  - statistician can focus on new methods
  - codify standard choices

how does one build tools?

- no one solution for all situations
- use existing tools wherever possible
  - new tools take time and care to build!
  - downloaded databases must be updated regularly
- human component is key
  - need informatics expertise
  - need continual dialog with biologists
- build bridges (interfaces) between tools
  - Web interface uses PHP
  - commands are created dynamically for R
- continually rethink & redesign organization

perspectives for building a community where disease data and models are shared

Benefits of wider access to datasets and models:

1- catalyze new insights on disease & methods

2- enable deeper comparison of methods & results

Lessons Learned:

1- need quick feedback between biologists & analysts

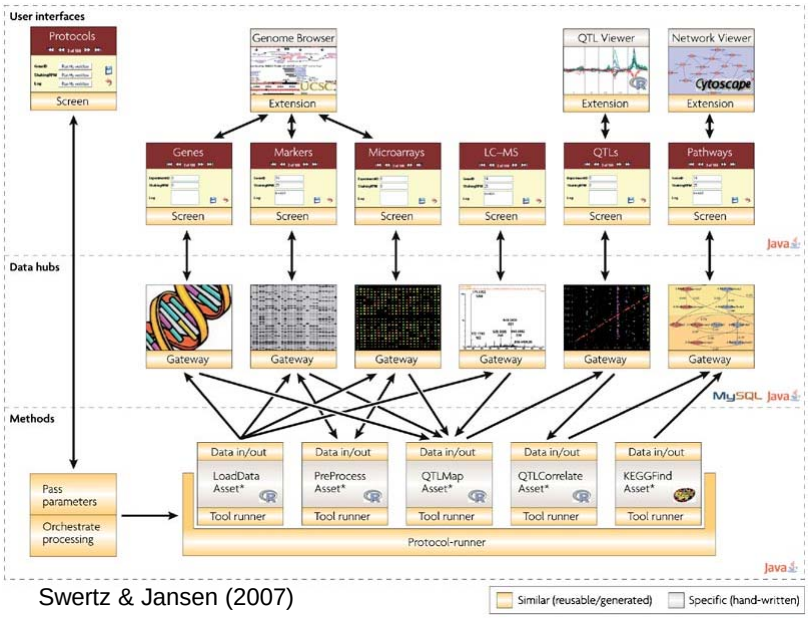
2- involve biologists early in development

3- repeated use of pipelines leads to
documented learning from experience
increased rigor in methods

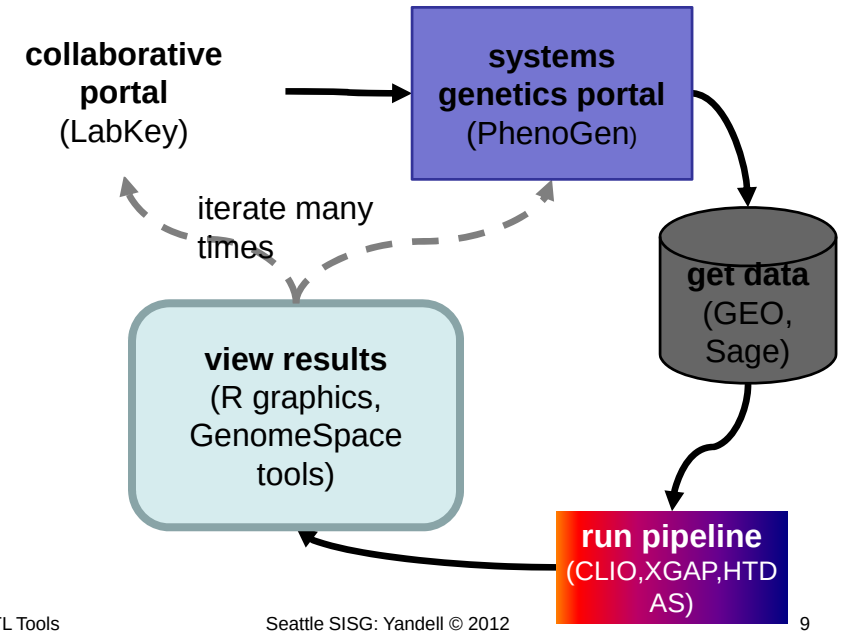
Challenges Ahead:

1- stitching together components as coherent system

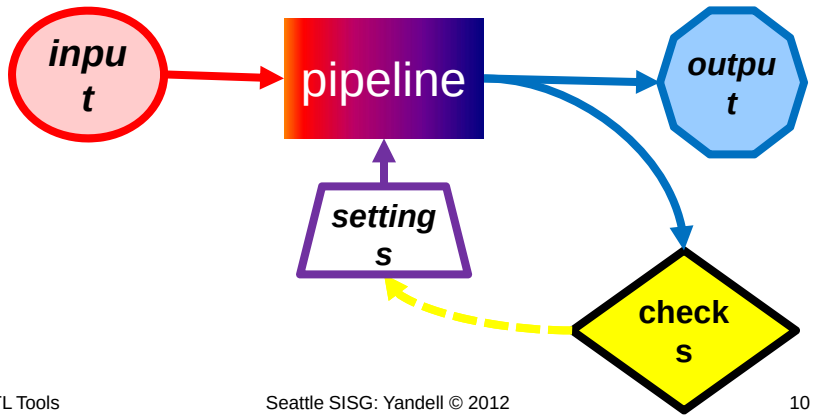
2- ramping up to ever larger molecular datasets



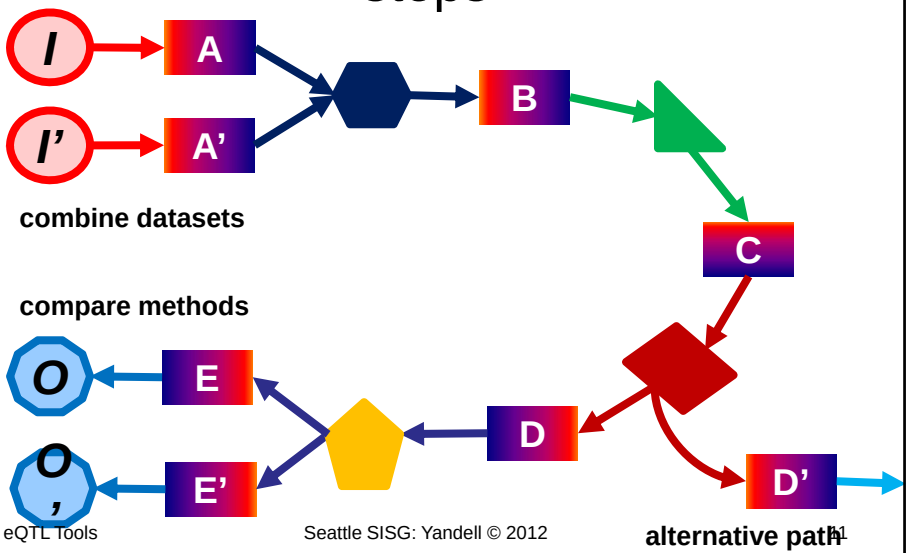
Swertz & Jansen (2007)



analysis pipeline acts on objects (extends concept of GenePattern)

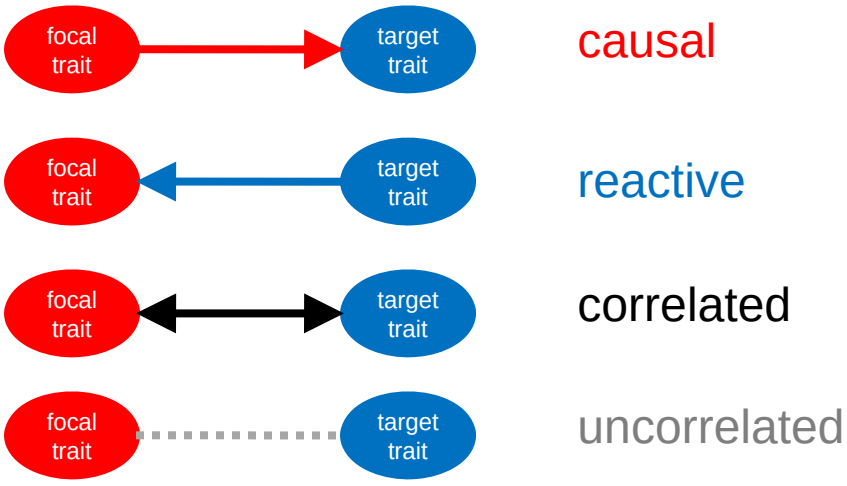


pipeline is composed of many steps

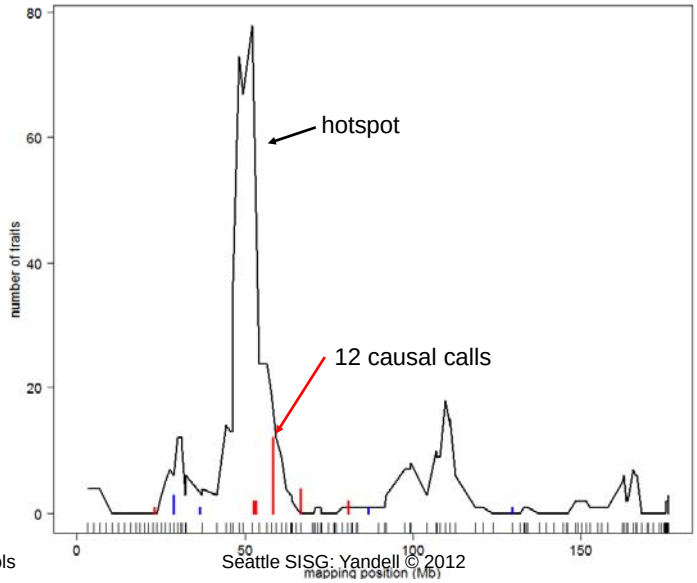


causal model selection choices

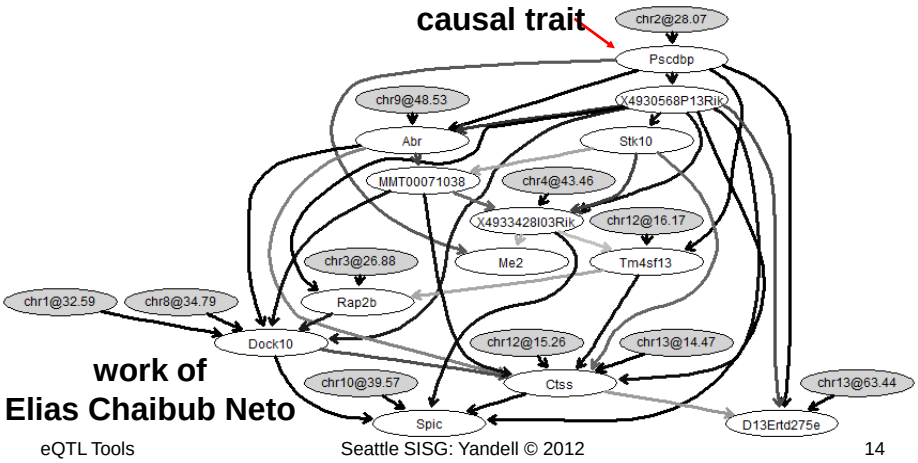
in context of larger, unknown network



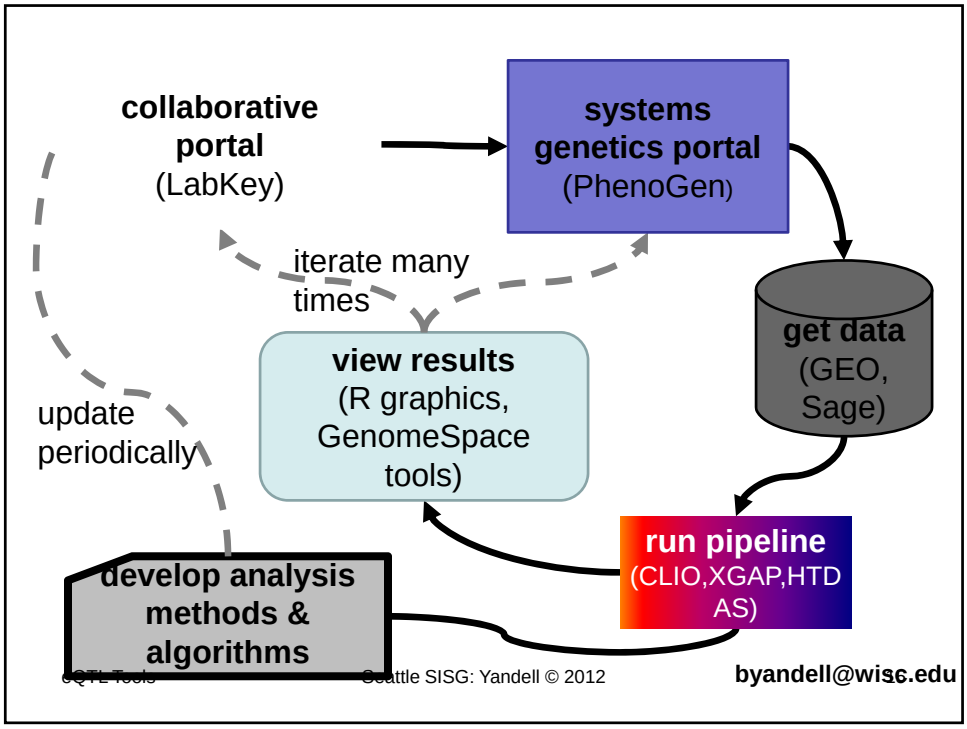
BxH ApoE-/- chr 2: causal architecture

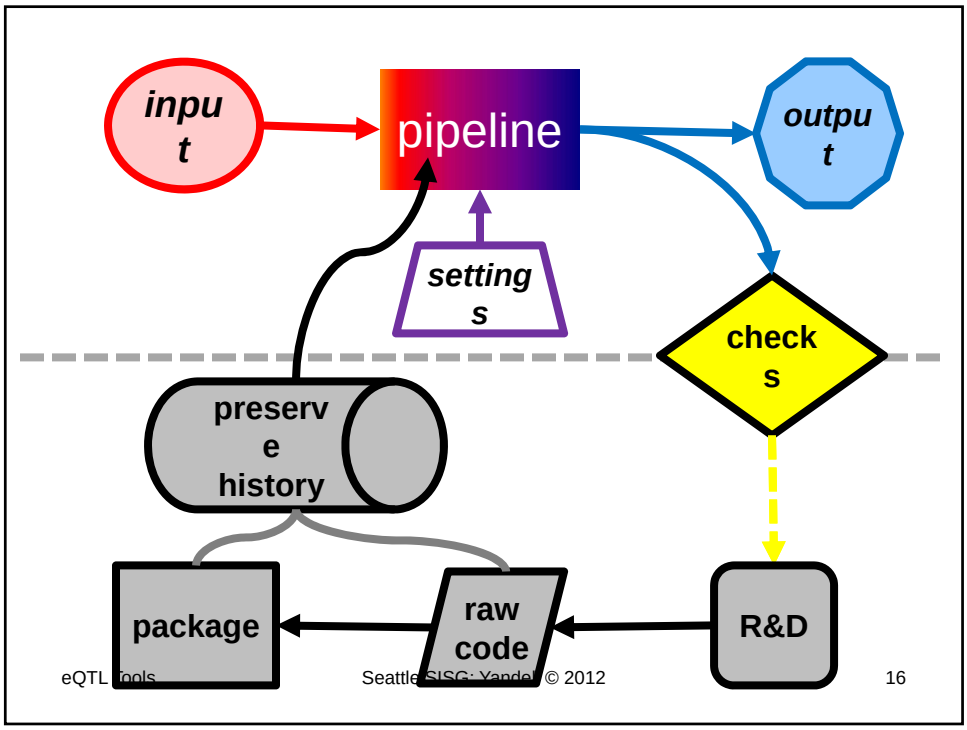


BxH ApoE-/- causal network for transcription factor Pscdbp



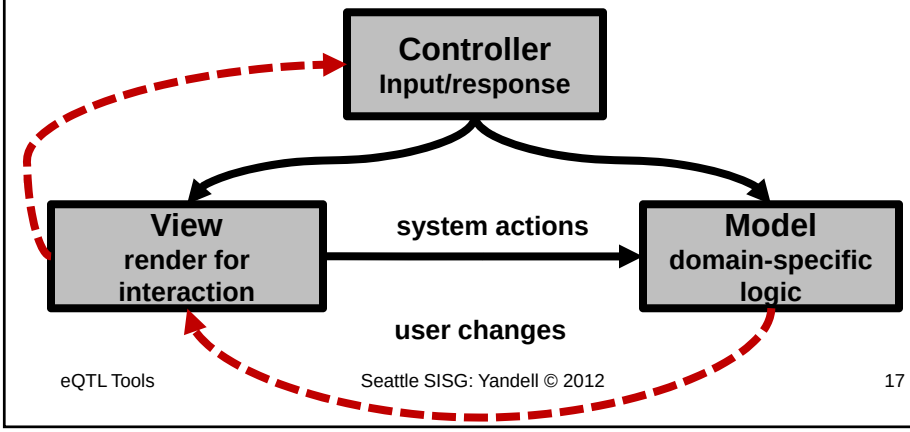
work of Elias Chaibub Neto

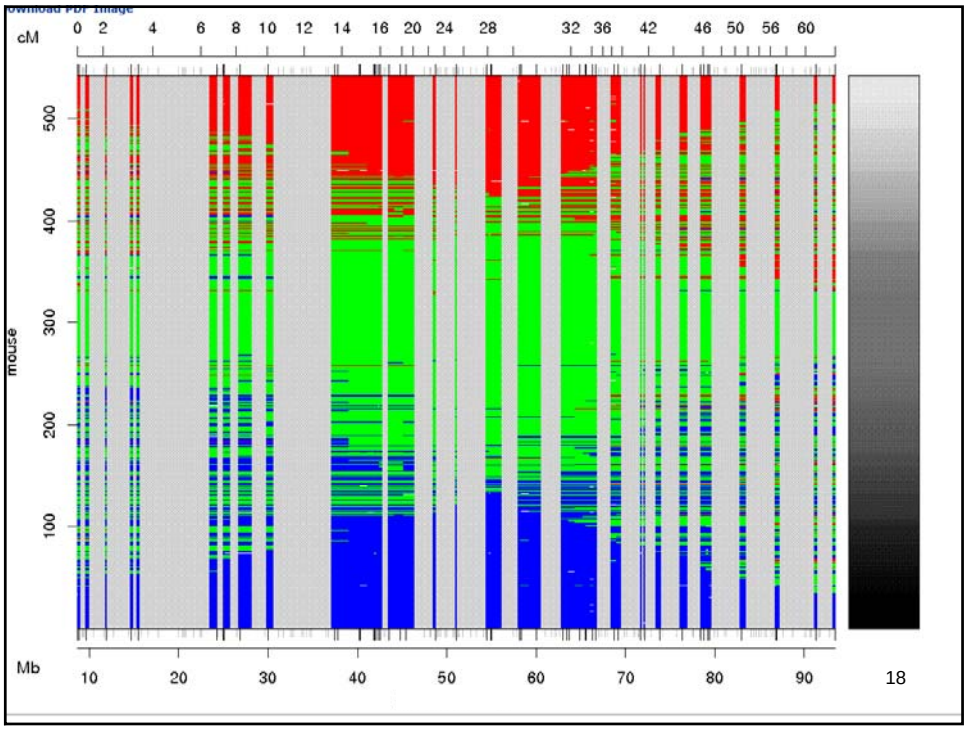

collaborative portal (LabKey)
systems genetics portal (PhenoGen)
iterate many times
get data (GEO, Sage)
view results (R graphics, GenomeSpace tools)
update periodically
run pipeline (CLIO,XGAP,HTDAS)
develop analysis methods & algorithms
eQTL Tools
Seattle SISG: Yandell © 2012
byandell@wisc.edu

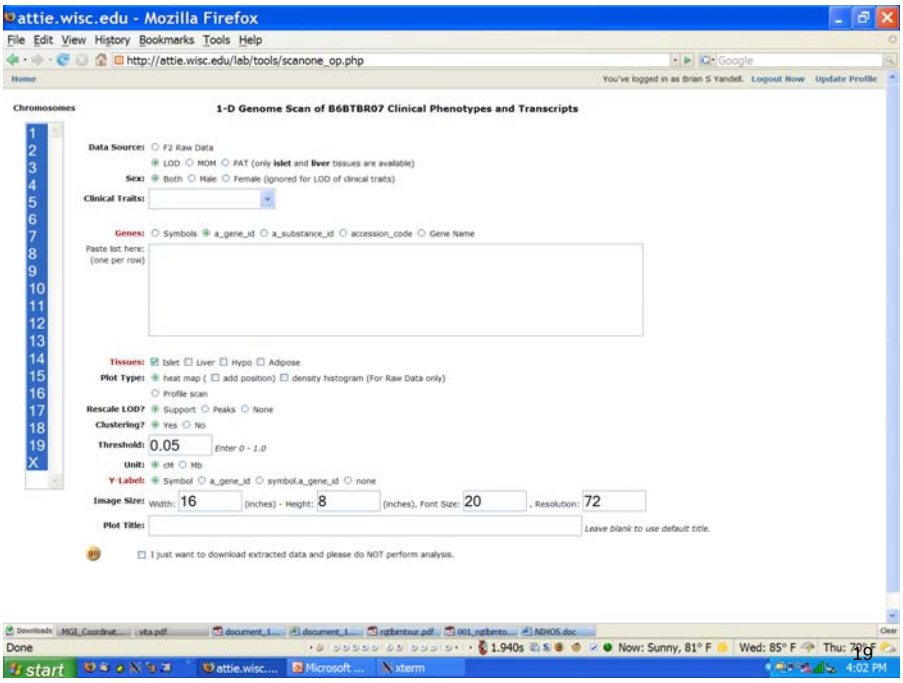

input
pipeline
output
setting s
check s
preserve history
package
raw code
R&D
eQTL Tools
Seattle SISG: Yandell © 2012
16

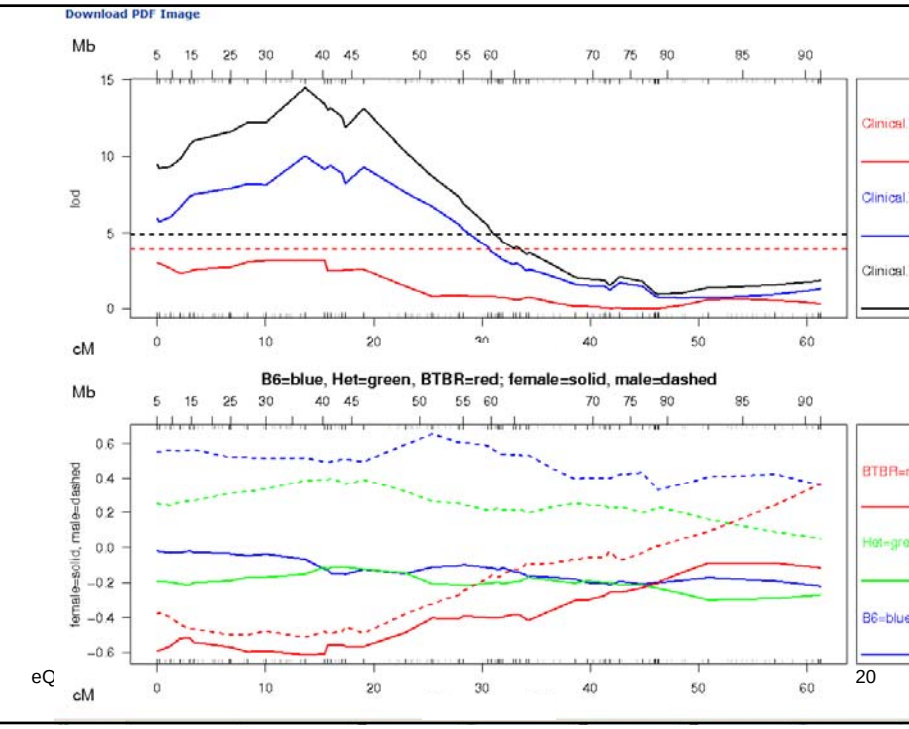
Model/View/Controller (MVC)
software architecture

- isolate domain logic from input and presentation
- permit independent development, testing, maintenance





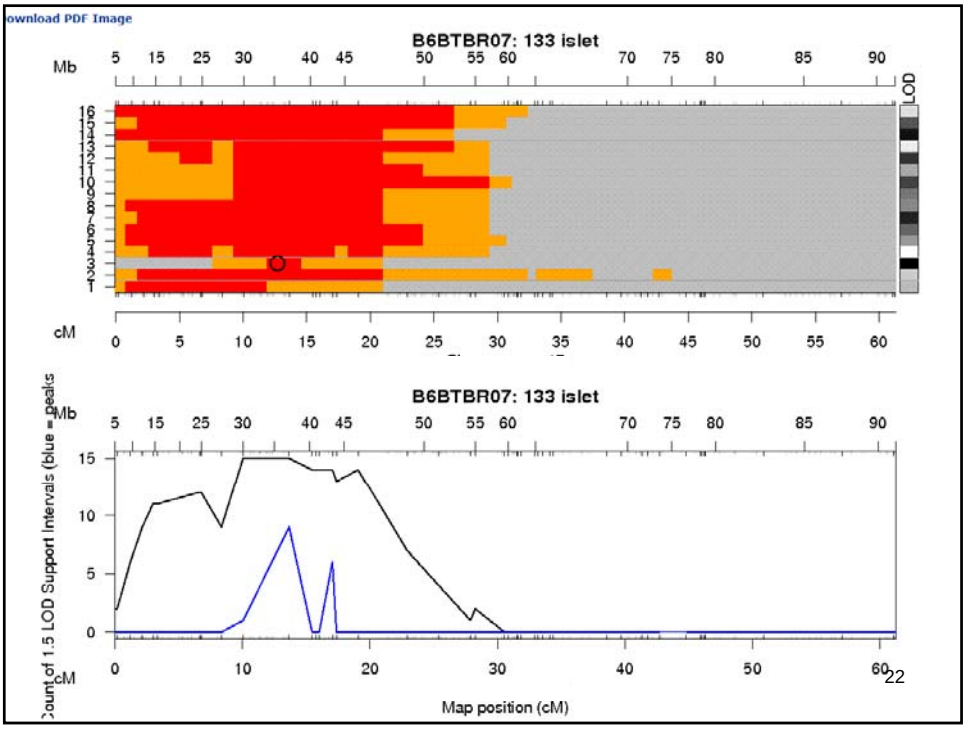

attie.wisc.edu - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://attie.wisc.edu/lab/tools/scanone_op.php
Home
You've logged in as Brian S Yandell. Logout Now Update Profile
Chromosomes
1-D Genome Scan of B6BTBR07 Clinical Phenotypes and Transcripts
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 X
Data Source: F2 Raw Data
LOD MOM PAT (only islet and liver tissues are available)
Sex: Both Male Female (ignored for LOD of clinical traits)
Clinical Traits:
Genes: Symbols a_gene_id a_substance_id accession_code Gene Name
Paste list here: (one per row)
Tissues: Islet Liver Hypo Adipose
Plot Type: heat map (add position) density histogram (For Raw Data only)
Profile scan
Rescale LOD? Support Peaks None
Clustering? Yes No
Threshold: 0.05 Enter 0 - 1.0
Unit: cM Mb
Y Label: Symbol a_gene_id symbol.a_gene_id none
Image Size: Width: 16 (inches) - Height: 8 (inches), Font Size: 20 , Resolution: 72
Plot Title: Leave blank to use default title.
I just want to download extracted data and please do NOT perform analysis.
Done
1.940s
Now: Sunny, 81° F
Wed: 85° F
Thu:
start
attie.wisc....
Microsoft ...
xterm
4:02 PM

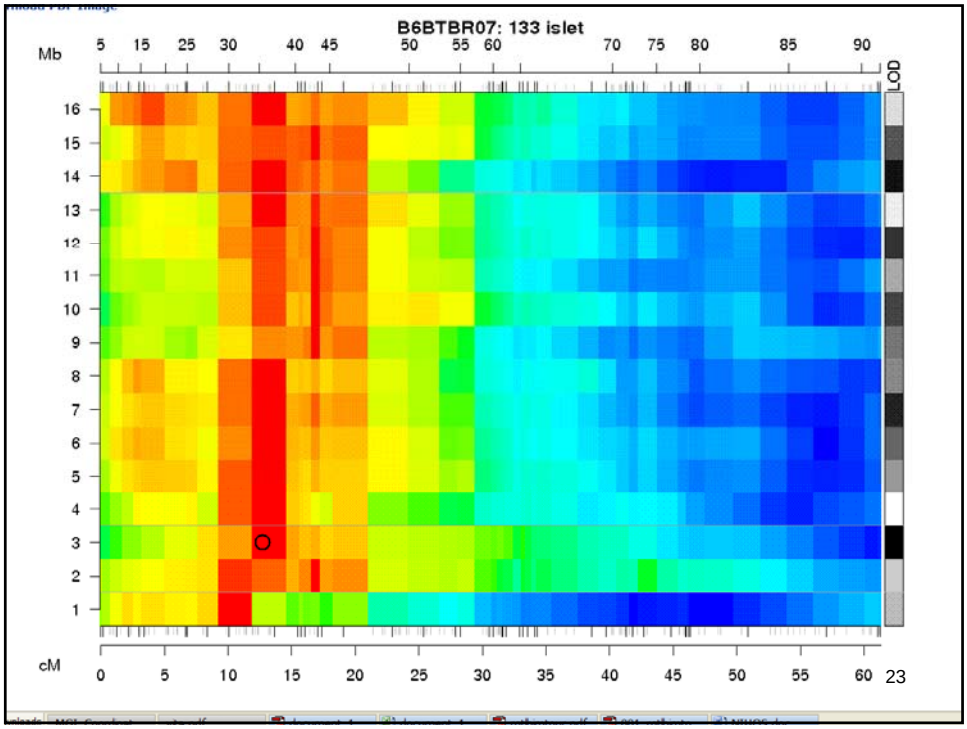

Download PDF Image
Mb
lod
cM
Clinical.Tot
Clinical.Tot
Clinical.Tot
B6=blue, Het=green, BTBR=red; female=solid, male=dashed
Mb
female=solid, male=dashed
cM
BTBR=red
Het=green
B6=blue

eQ

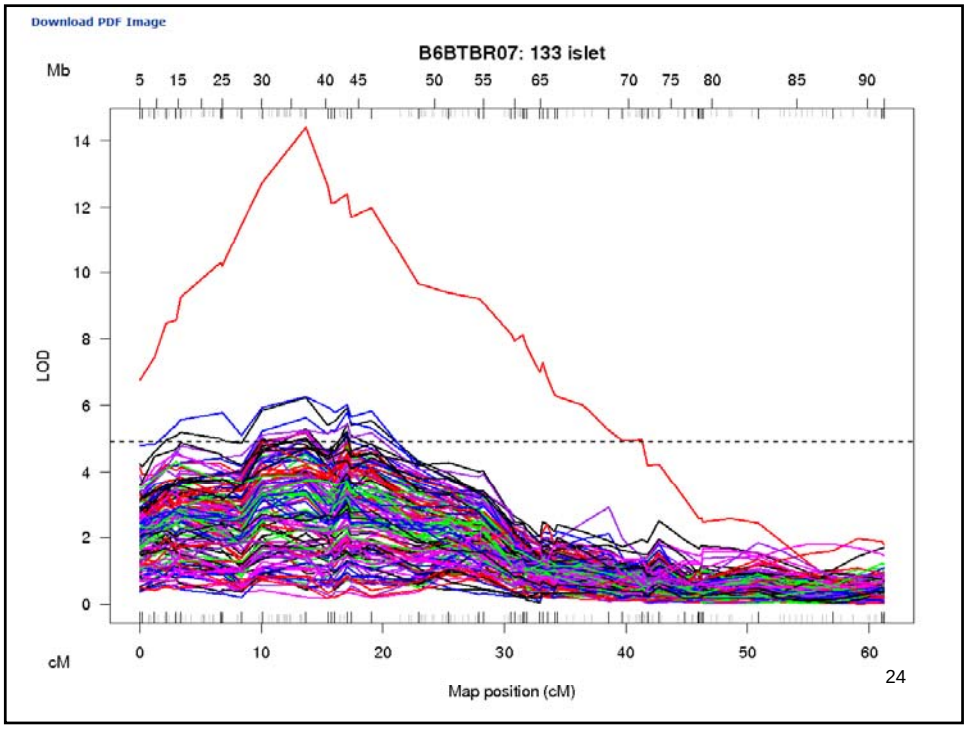
automated R script

```
library('B6BTBR07')

out <- multtrait(cross.name='B6BTBR07',
  filename = 'scanone_1214952578.csv',
  category = 'islet', chr = c(17),
  threshold.level = 0.05, sex = 'both',)

sink('scanone_1214952578.txt')
print(summary(out))
sink()

bitmap('scanone_1214952578%03d.bmp',
  height = 12, width = 16, res = 72, pointsize = 20)
plot(out, use.cM = TRUE)
dev.off()
```

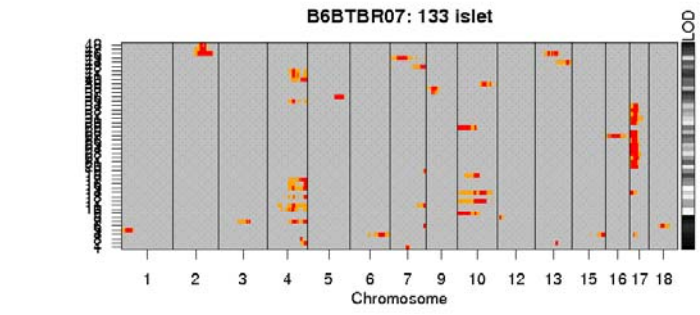



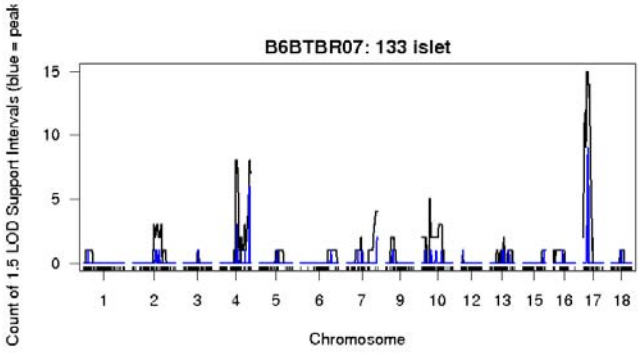

B6BTBR07: 133 islet
Mb
5 15 25 30 40 45 50 55 60 70 75 80 85 90
LOD
16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
cM
0 5 10 15 20 25 30 35 40 45 50 55 60

23

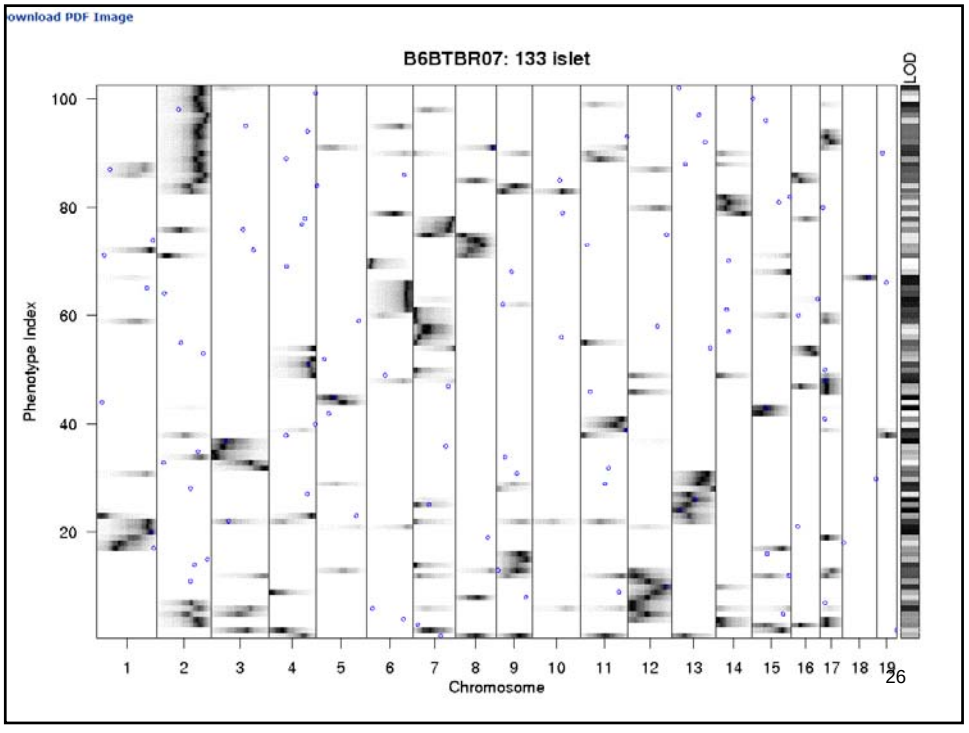

Download PDF Image
B6BTBR07: 133 islet
Mb
5 15 25 30 40 45 50 55 65 70 75 80 85 90
LOD
14 12 10 8 6 4 2 0
cM
0 10 20 30 40 50 60
Map position (cM)

24


B6BTBR07: 133 islet
LOD
1 2 3 4 5 6 7 9 10 12 13 15 16 17 18
Chromosome


B6BTBR07: 133 islet
Count of 1.5 LOD Support Intervals (blue = peak
15
10
5
0
1 2 3 4 5 6 7 9 10 12 13 15 16 17 18
Chromosome

25


ownload PDF Image
B6BTBR07: 133 islet
LOD
Phenotype Index
100
80
60
40
20
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
Chromosome