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

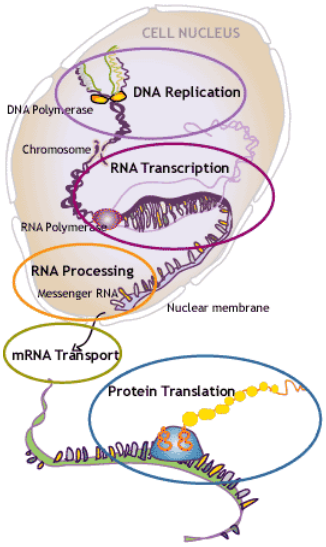
Monday		
8:30-10	Introductions; Overview of System Genetics	1-50
10:30-12	QTL Model Selection	51-100
1:30-3	Gene Mapping for Multiple Correlated Traits	101-150
3:30-5	Hands On Lab: R/qtl	151-200
Tuesday		
8:30-10	Permutation Tests for Correlated Traits	201-250
10:30-12	Scanning the Genome for Causal 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

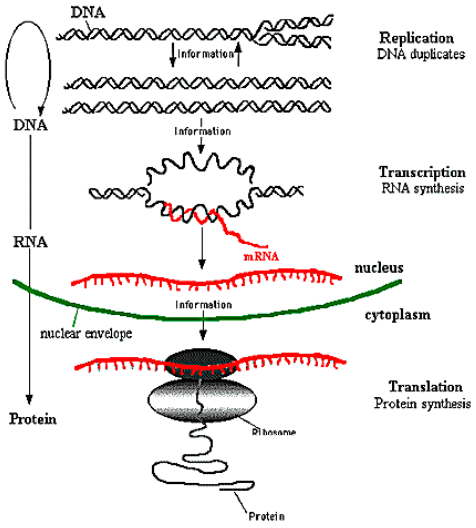
- Big idea: how do genes affect organisms?
- Measuring system(s) state(s) of an organism
- QTL mapping as tool toward goal
- Making sense of multiple traits
- Connecting traits to biochemical pathways
- Putting it all together: workflows

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



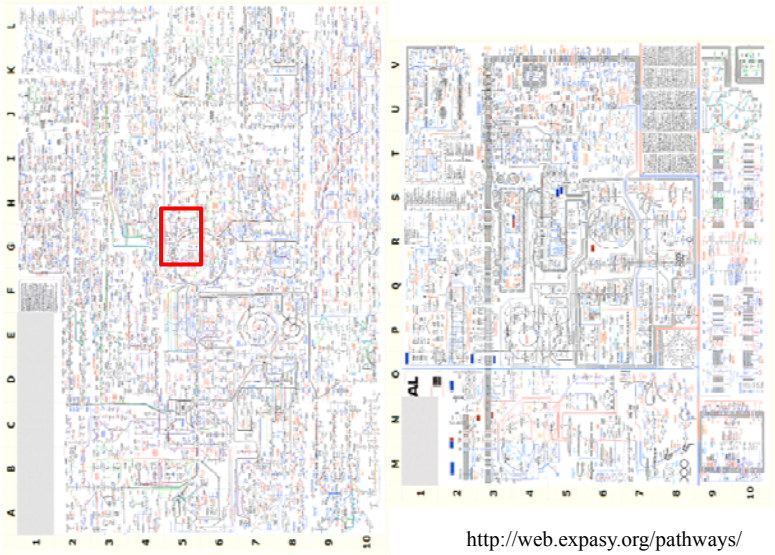
www.nobelprize.org/educational/medicine/dna



The Central Dogma of Molecular Biology

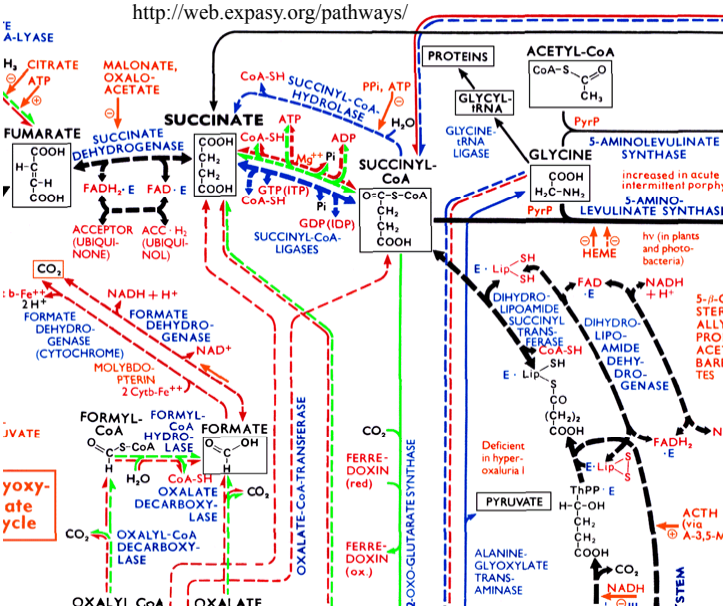
www.accessexcellence.org/RC/VL/GG/central.php

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.)?

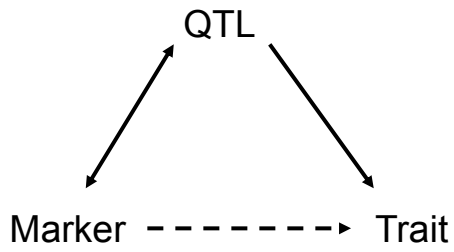
cross two inbred lines

→ linkage disequilibrium

→ associations

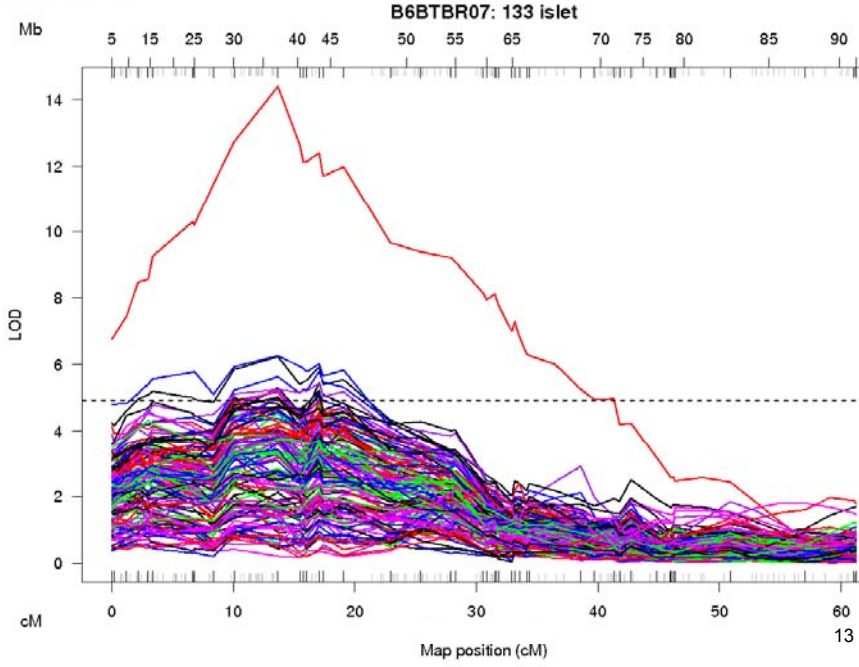
→ linked segregating QTL

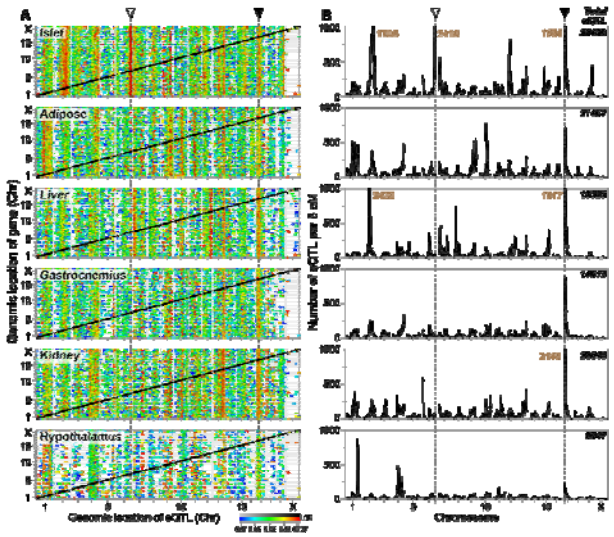
(after Gary Churchill)



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).*

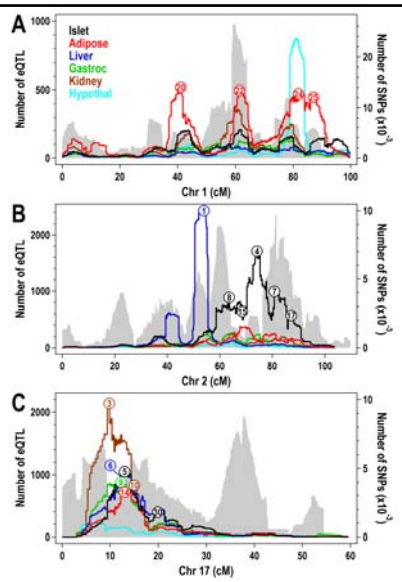
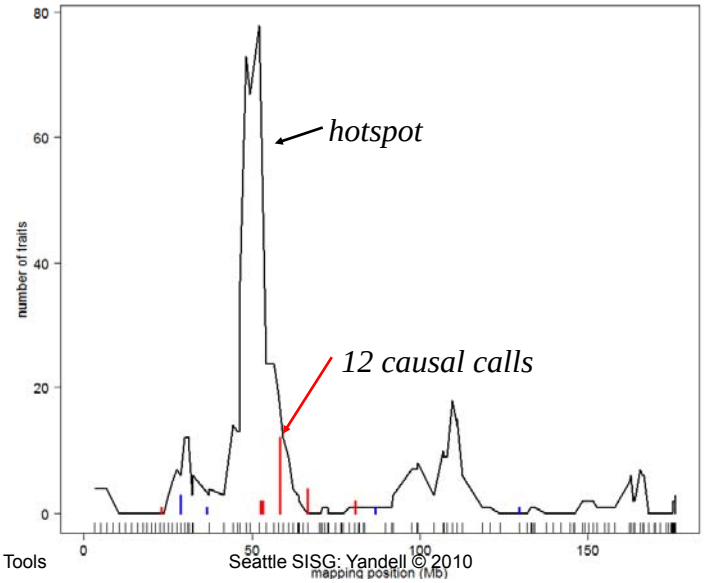


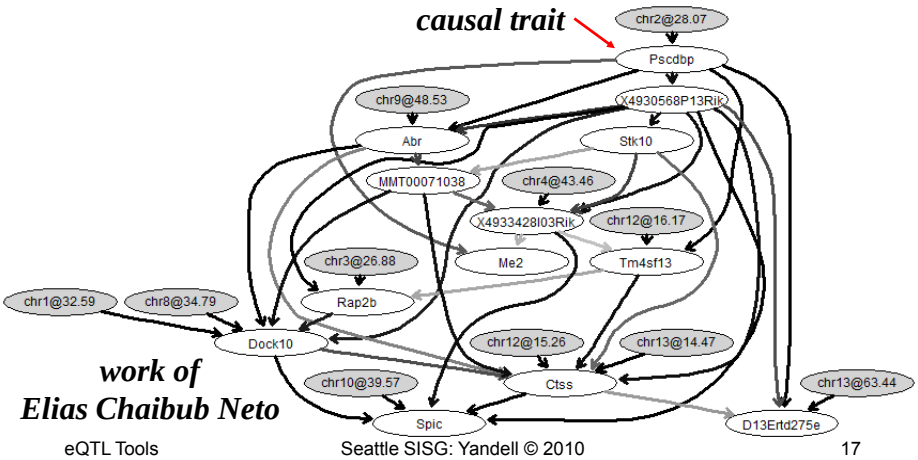
Figure 4 *Tissue-specific hotspots with eQTL and SNP architecture for Chrs 1, 2 and 17.*

*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



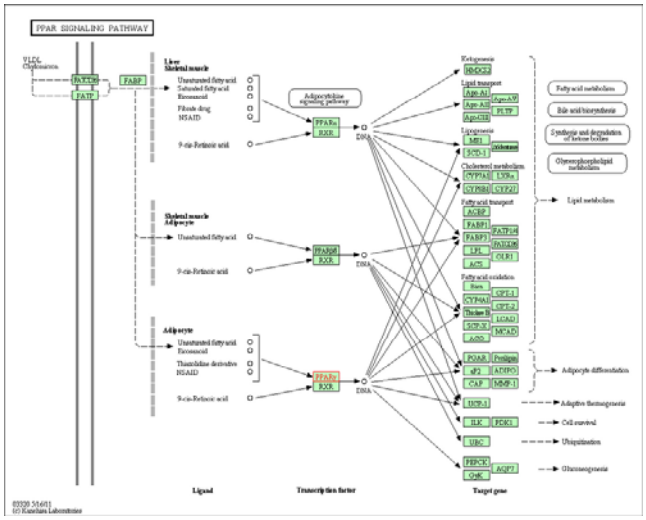
BxH ApoE-/- causal network for transcription factor Pscdbp



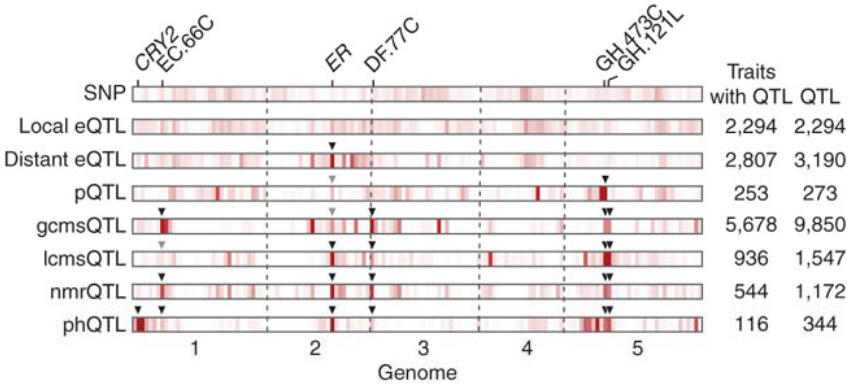
Connecting to biochemical pathways

- 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



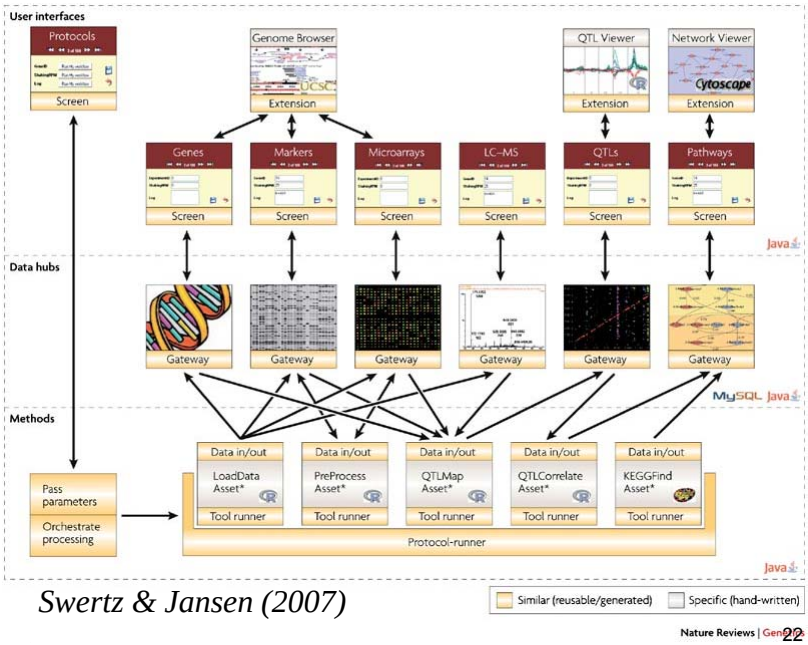
phenotypic buffering of molecular QTL



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

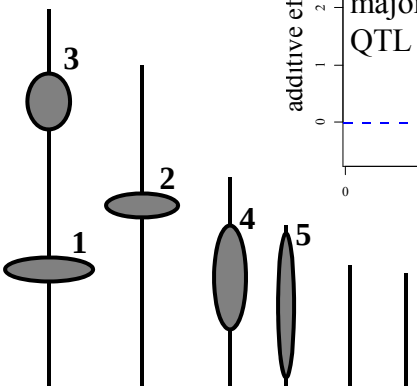
- wrong model: biased view
  - fool yourself: bad guess at locations, effects
  - detect ghost QTL between linked loci
  - miss epistasis completely
- low power
- bad science
  - use best tools for the job
  - maximize scarce research resources
  - leverage already big investment in experiment

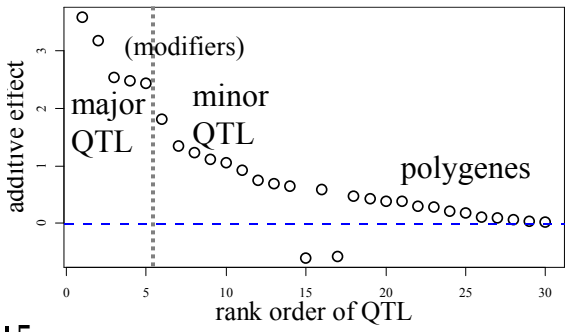
advantages of multiple QTL approach

- 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

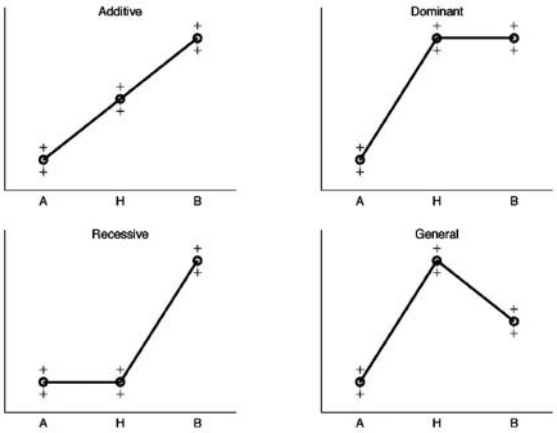
major QTL on linkage map





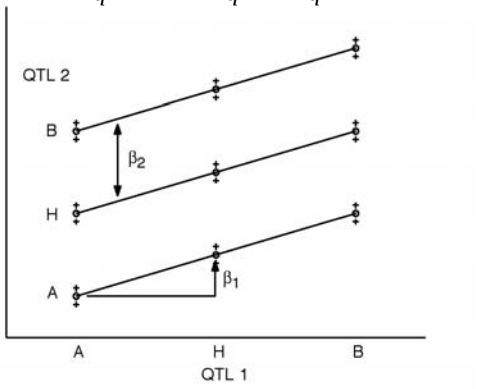
Gene Action and Epistasis

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



additive effects of two QTL (Gary Churchill)

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



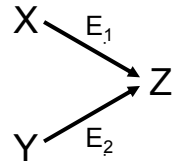
Epistasis (Gary Churchill)

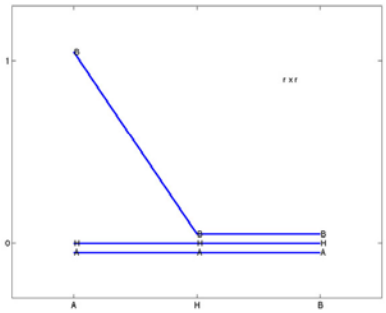
The allelic state at one locus can mask or uncover the effects of allelic variation at another.

- W. Bateson, 1907.

epistasis in parallel pathways (GAC)

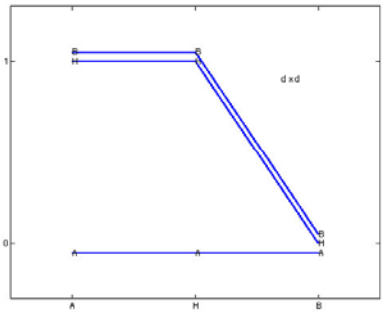
- Z keeps trait value low
- neither E_1 nor E_2 is rate limiting
- loss of function alleles are segregating from parent A at E_1 and from parent B at E_2





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

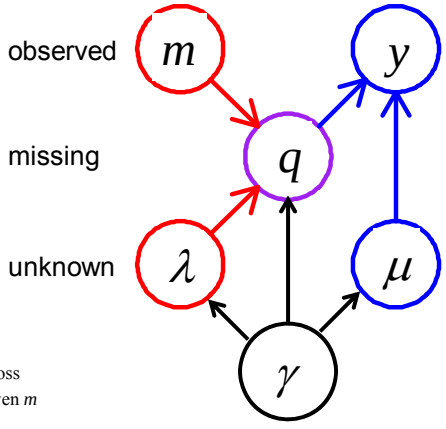
- classical study
  - ***maximize*** over unknown effects
  - ***test*** for detection of QTL at loci
  - model selection in stepwise fashion
- Bayesian study
  - ***average*** over unknown effects
  - ***estimate*** chance of detecting QTL
  - sample all possible models
- both approaches
  - average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

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

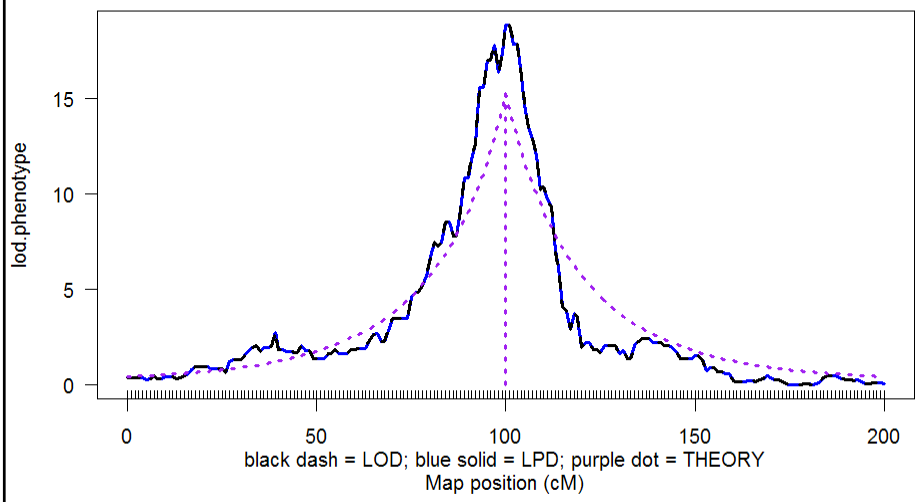
$$\mathrm{LPD}(\lambda) = \log_{10}\{\mathrm{pr}(\lambda \mid m)\int \mathrm{pr}(y \mid m, \mu, \lambda)\mathrm{pr}(\mu)d\mu\} + C$$

likelihood mixes over missing QTL genotypes:

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

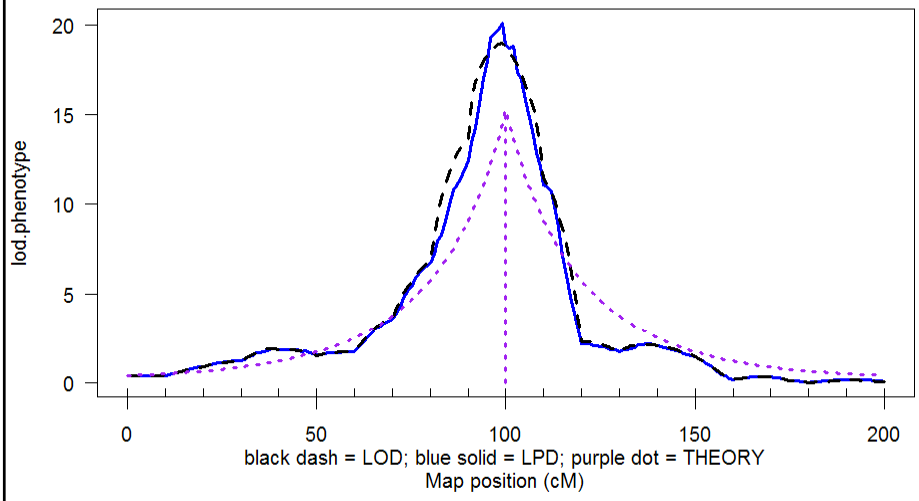
LOD & LPD: 1 QTL

n.ind = 100, 1 cM marker spacing



LOD & LPD: 1 QTL

n.ind = 100, 10 cM marker spacing



marginal LOD or LPD

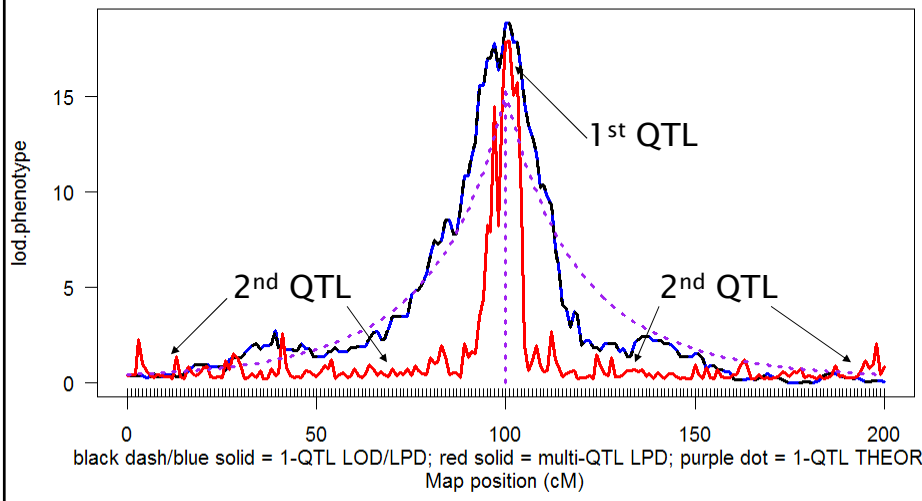
- compare two genetic architectures (γ_2,γ_1) at each locus
  - with (γ_2) or without (γ_1) another QTL at locus λ
    - preserve model hierarchy (e.g. drop any epistasis with QTL at λ)
  - with (γ_2) or without (γ_1) epistasis with QTL at locus λ
  - γ_2 contains γ_1 as a sub-architecture
- allow for 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

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

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

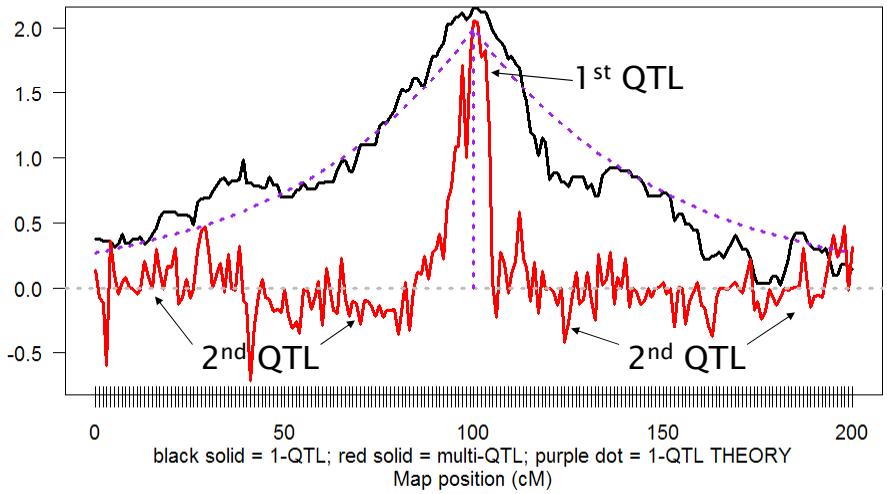
LPD: 1 QTL vs. multi-QTL

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