QTLs & Microarrays Overview

Brian S. Yandell
University of Wisconsin-Madison
www.stat.wisc.edu/~yandell/statgen

- what is a QTL?
- what is the goal of your QTL study?
- why worry about multiple QTL?
- why study multiple traits together?

what is a QTL?

- QTL = quantitative trait locus (or loci)
  - trait = phenotype = characteristic of interest
  - quantitative = measured somehow
    - qualitative traits can often be directly mapped
    - quantitative traits not readily mapped
  - locus = location in genome affecting trait
    - gene or collection of tightly linked genes
    - some physical feature of genome

what is the goal of QTL study?

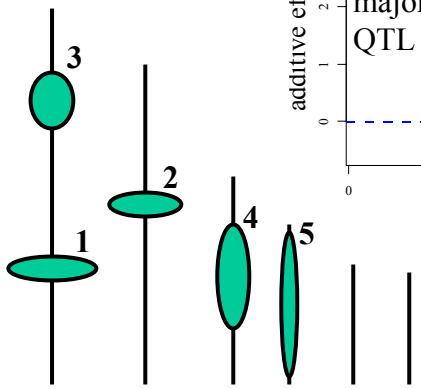
- uncover underlying biochemistry
  - identify how networks function, break down
  - find useful candidates for (medical) intervention
  - epistasis may play key role
  - statistical goal: maximize number of correctly identified QTL
- basic science/evolution
  - how is the genome organized?
  - identify 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

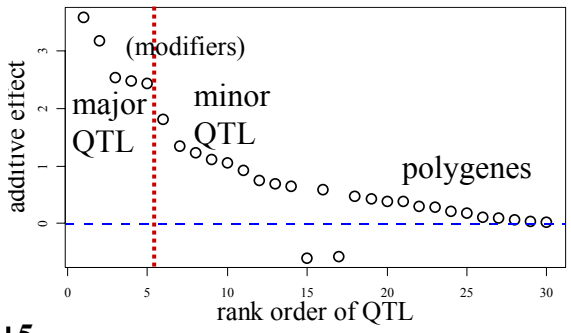
why worry about multiple QTL?

- so many possible genetic architectures!
  - number and positions of loci
  - gene action: additive, dominance, epistasis
  - how to efficiently search the model space?
- how to select "best" or "better" model(s)?
  - what criteria to use? where to draw the line?
  - shades of gray: exploratory vs. confirmatory study
  - how to balance false positives, false negatives?
- what are the key "features" of model?
  - means, variances & covariances, confidence regions
  - marginal or conditional distributions

Pareto diagram of QTL effects

major QTL on linkage map





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

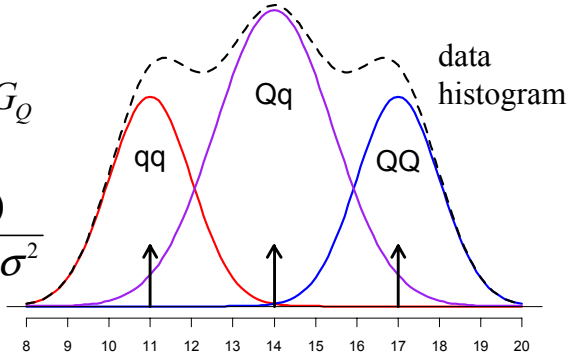
typical phenotype assumptions

- normal "bell-shaped" environmental variation
- genotypic value G_Q is composite of m QTL
- genetic uncorrelated with environment

$$E(Y \mid Q) = \mu + G_Q$$

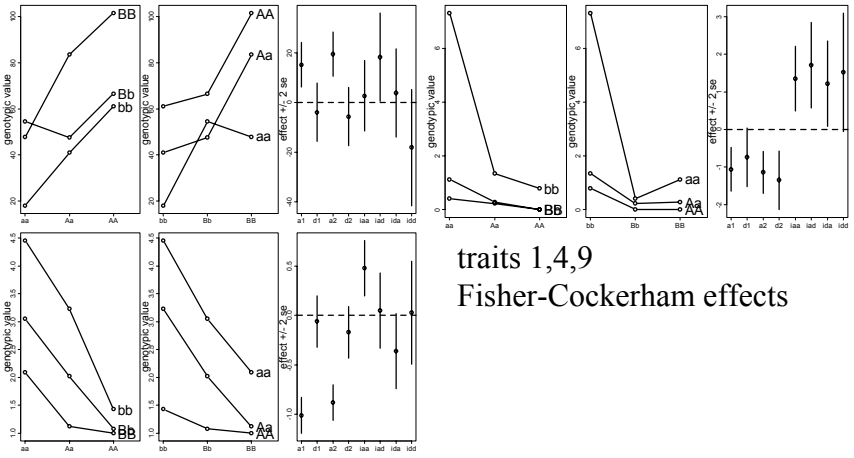
$$\text{var}(Y \mid Q) = \sigma^2$$

$$h^2 = \frac{\text{var}(G_Q)}{\text{var}(G_Q) + \sigma^2}$$



epistasis examples

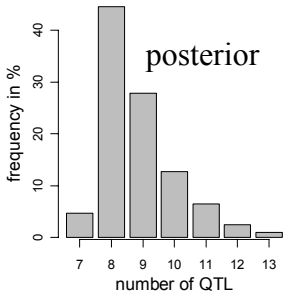
(Doebley Stec Gustus 1995; Zeng pers. comm.)



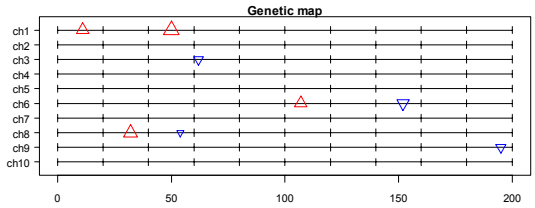
traits 1,4,9
Fisher-Cockerham effects

a complicated simulation

- simulated F2 intercross, 8 QTL
  - (Stephens, Fisch 1998)
  - n=200, heritability = 50%
  - detected 3 QTL
- increase to detect all 8
  - n=500, heritability to 97%



QTL	chr	loci	effect
1	1	11	–3
2	1	50	–5
3	3	62	+2
4	6	107	–3
5	6	152	+3
6	8	32	–4
7	8	54	+1
8	9	195	+2



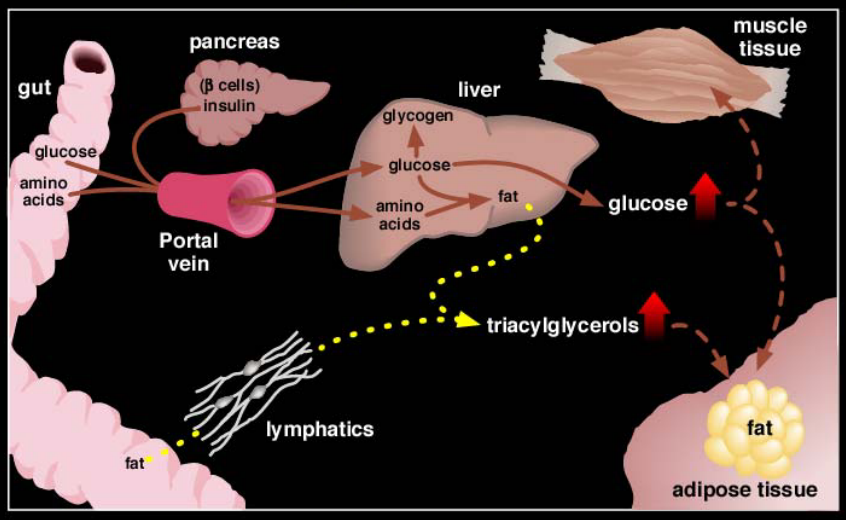
loci pattern across genome

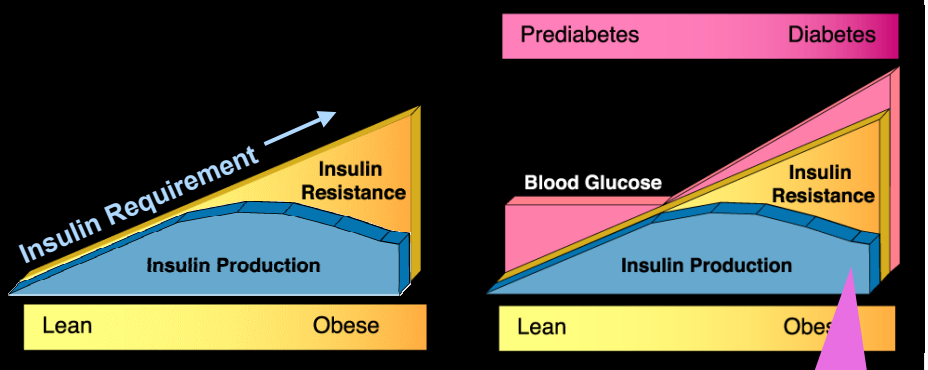
- notice which chromosomes have persistent loci
- best pattern found 42% of the time

Chromosome

m	**1**	2	**3**	4	5	**6**	7	**8**	**9**	10	**Count of 8000**
8	**2**	**0**	**1**	**0**	**0**	**2**	**0**	**2**	**1**	**0**	3371
9	*3*	0	1	0	0	2	0	2	1	0	751
7	2	0	1	0	0	2	0	*1*	1	0	377
9	2	0	1	0	0	2	0	2	1	0	218
9	2	0	1	0	0	*3*	0	2	1	0	218
9	2	0	1	0	0	2	0	2	*2*	0	198

Type 2 Diabetes Mellitus





from Unger & Orci *FASEB J.* (2001) 15,312

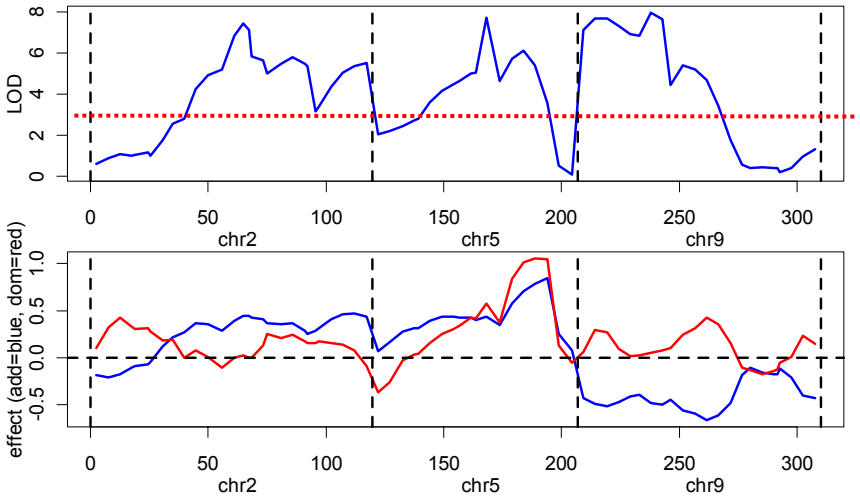
studying diabetes in an F2

- segregating cross of inbred lines
  - B6.ob x BTBR.ob → F1 → F2
  - selected mice with ob/ob alleles at leptin gene (chr 6)
  - measured and mapped body weight, insulin, glucose at various ages (Stoehr et al. 2000 Diabetes)
  - sacrificed at 14 weeks, tissues preserved
- gene expression data
  - Affymetrix microarrays on parental strains, F1
    - key tissues: adipose, liver, muscle, β-cells
    - novel discoveries of differential expression (Nadler et al. 2000 PNAS; Lan et al. 2002 in review; Ntambi et al. 2002 PNAS)
  - RT-PCR on 108 F2 mice liver tissues
    - 15 genes, selected as important in diabetes pathways
    - SCD1, PEPCK, ACO, FAS, GPAT, PPARgamma, PPARalpha, G6Pase, PDI,…

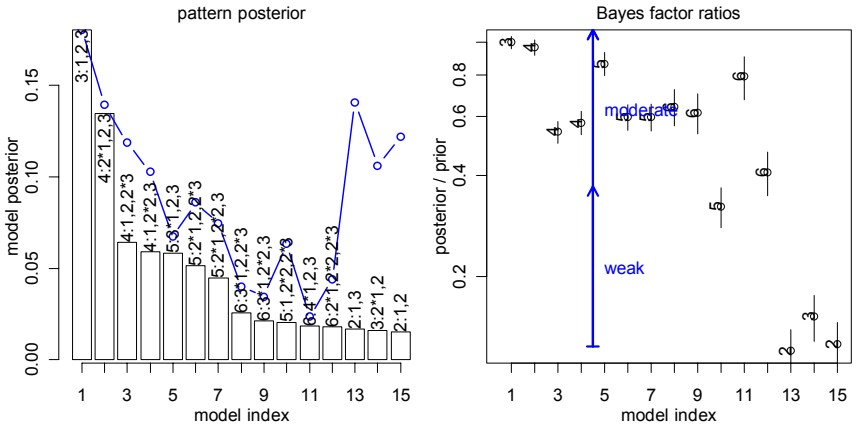
why map gene expression as a quantitative trait?

- *cis*- or *trans*-action?
  - does gene control its own expression?
  - evidence for both modes (Brem et al. 2002 Science)
- mechanics of gene expression mapping
  - measure gene expression in intercross (F2) population
  - map expression as quantitative trait (QTL technology)
  - adjust for multiple testing via false discovery rate
- research groups working on expression QTLs
  - review by Cheung and Spielman (2002 *Nat Gen Suppl*)
  - Kruglyak (Brem et al. 2002 *Science*)
  - Doerge et al. (Purdue); Jansen et al. (Waginingen)
  - Williams et al. (U KY); Lusis et al. (UCLA)
  - Dumas et al. (2000 *J Hypertension*)

Multiple Interval Mapping
SCD1: multiple QTL plus epistasis!

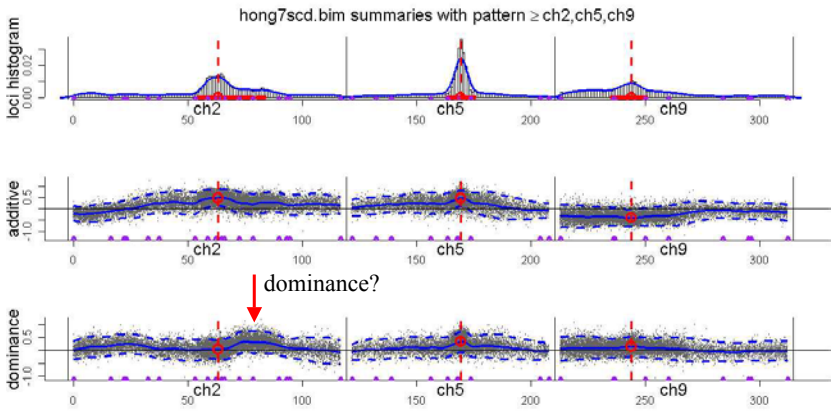


Bayesian model assessment:
chromosome QTL pattern for SCD1

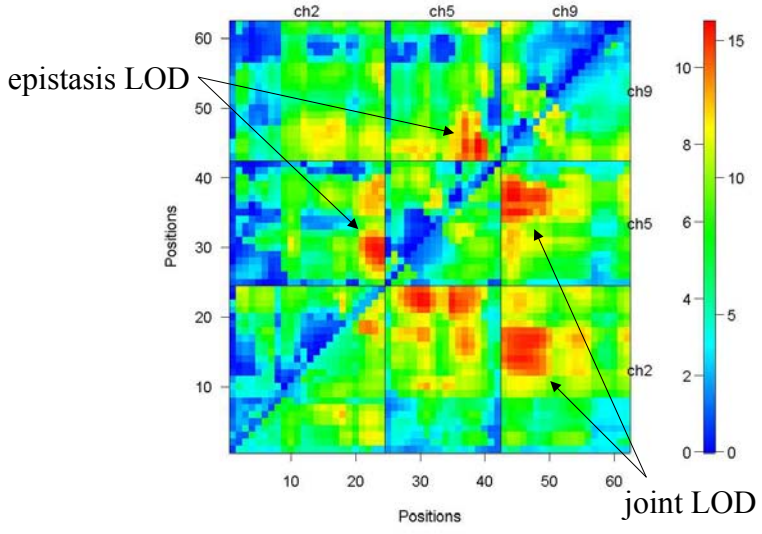


trans-acting QTL for SCD1

(no epistasis yet: see Yi, Xu, Allison 2003)



2-D scan: assumes only 2 QTL!



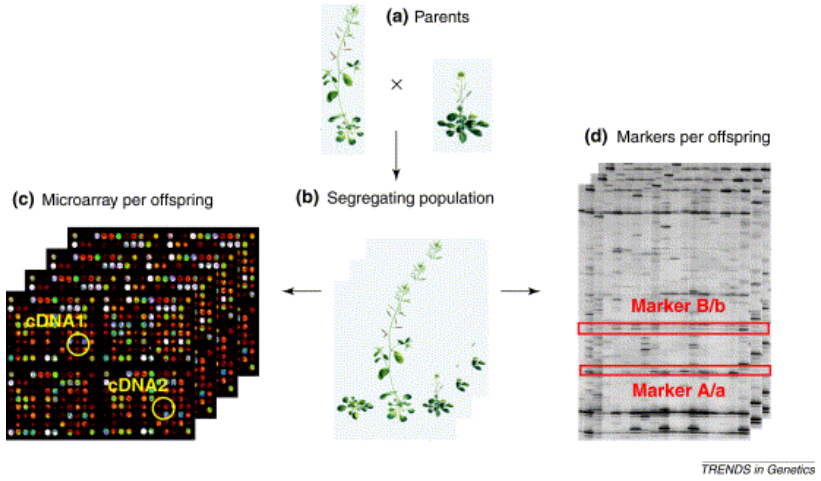
why study multiple traits together?

- environmental correlation
  - non-genetic, controllable by design
  - historical correlation (learned behavior)
  - physiological correlation (same body)
- genetic correlation
  - pleiotropy
    - one gene, many functions
    - common biochemical pathway, splicing variants
  - close linkage
    - two tightly linked genes
    - genotypes *Q* are collinear

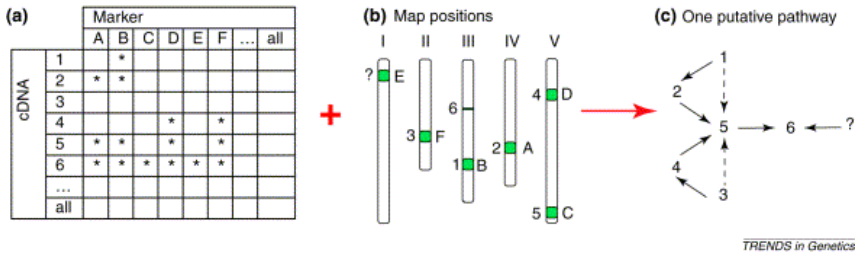
high throughput dilemma

- want to focus on gene expression network
  - ideally capture pathway in a few dimensions
  - allow for complicated genetic architecture
- may have multiple controlling loci
  - could affect many genes in coordinated fashion
  - could show evidence of epistasis
  - quick assessment via interval mapping may be misleading
- try mapping principle components as super-traits
  - capture key multivariate features of multiple traits
  - elicit biochemical pathways (Henderson et al. Hoeschele 2001; Ong Page 2002)

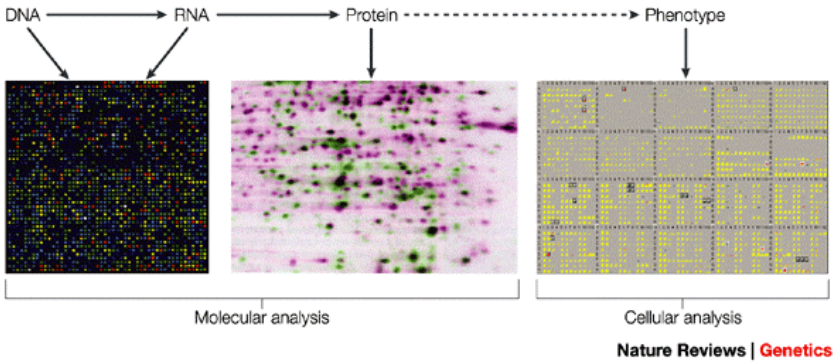
idea of mapping microarrays (Jansen Nap 2001)



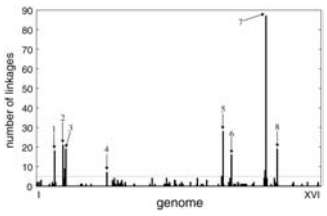
goal: unravel biochemical pathways (Jansen Nap 2001)



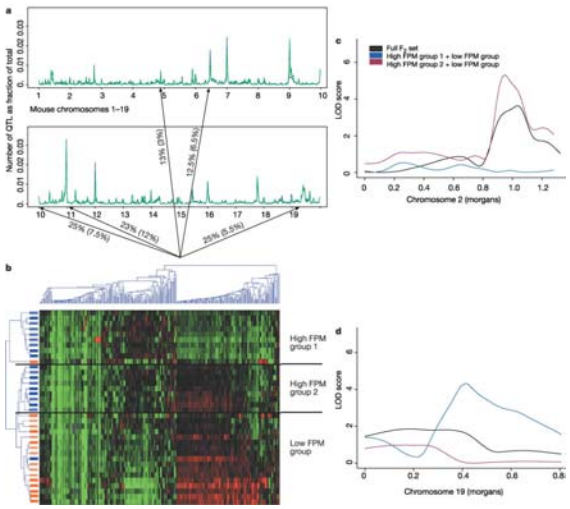
central dogma via microarrays (Bochner 2003)



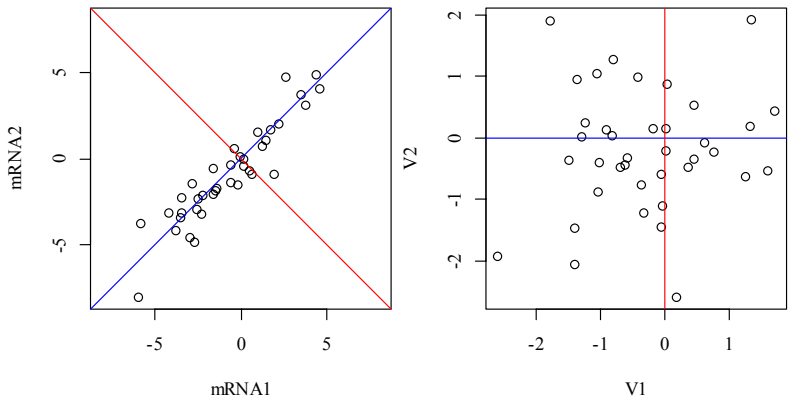
expression pleiotropy in yeast genome (Brem et al. 2002)



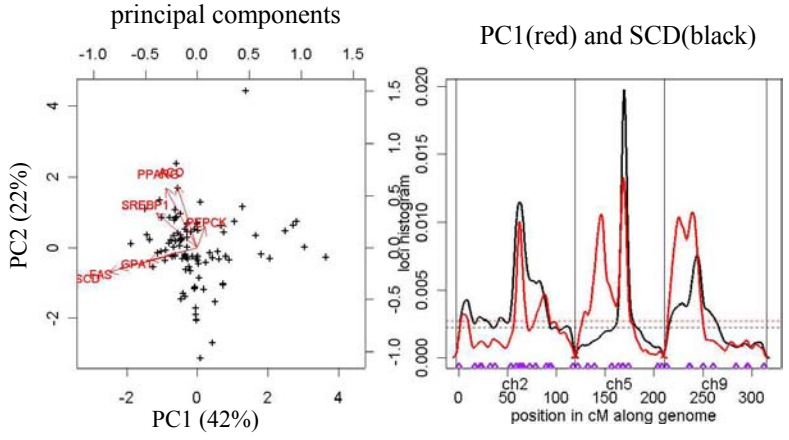
coordinated expression in mouse genome (Schadt et al. 2003)



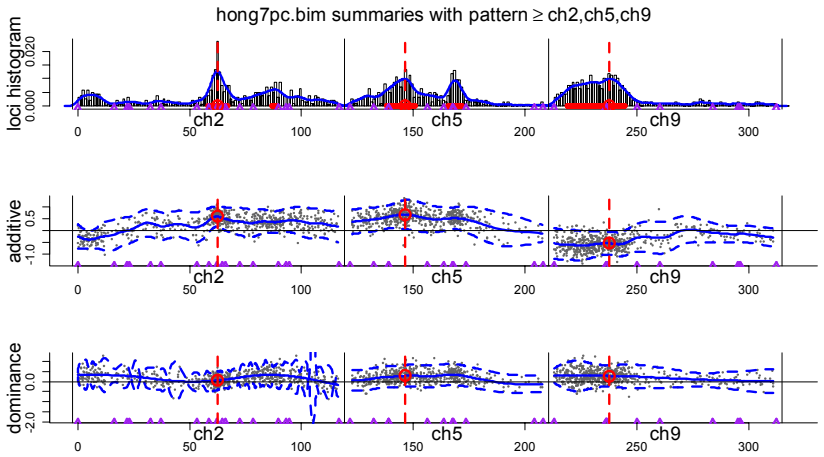
PC simply rotates & rescales to find major axes of variation



multivariate screen for gene expressing mapping



mapping first diabetes PC as a trait



pFDR for PC1 analysis