Association Mapping

Brian S. Yandell, UW-Madison

January 2017

evolution of QTL models

original ideas focused on rare & costly markers
models & methods refined as technology advanced

- single marker regression
- QTL (quantitative trait loci)
  - single locus models: interval mapping for QTL
  - QTL model search: QTLs & epistasis
- **polygenes (association mapping)**
  - adjust for population structure
  - capture "missing heritability"
- genome-wide selection

polygene big idea

- only detect some genetic effects
  - significant QTL
- effects of modest or small effect ignored
  - non-significant QTL
  - effects too small to observe or test
- these other effects have two sources
  - many small effects on phenotype
  - population admixture reflected in genome structure

missing heritability

"actual" genetic model $y = \mu_q + e$
with J genetic effects (recode q_j to have variance 1)

$$\mu_q = \mu + \sum_{j=1}^{J} \beta_j q_j$$

actual heritability:

$$h^2 = \frac{\sum_{j=1}^{J} \beta_j^2}{\sigma^2 + \sum_{j=1}^{J} \beta_j^2}$$

(consider backcross and ignore epistasis from here forward)

best QTL misses most of heritability

but "best" QTL model has 2 terms

$$\mu_q = \mu + \beta_1 q_1 + \beta_2 q_2$$

with heritability:

$$h^2_{\mathtt{QTL}} = \frac{\beta_1^2 + \beta_2^2}{\sigma^2 + \sum_{j=1}^{J} \beta_j^2}$$

missing genetic variability:

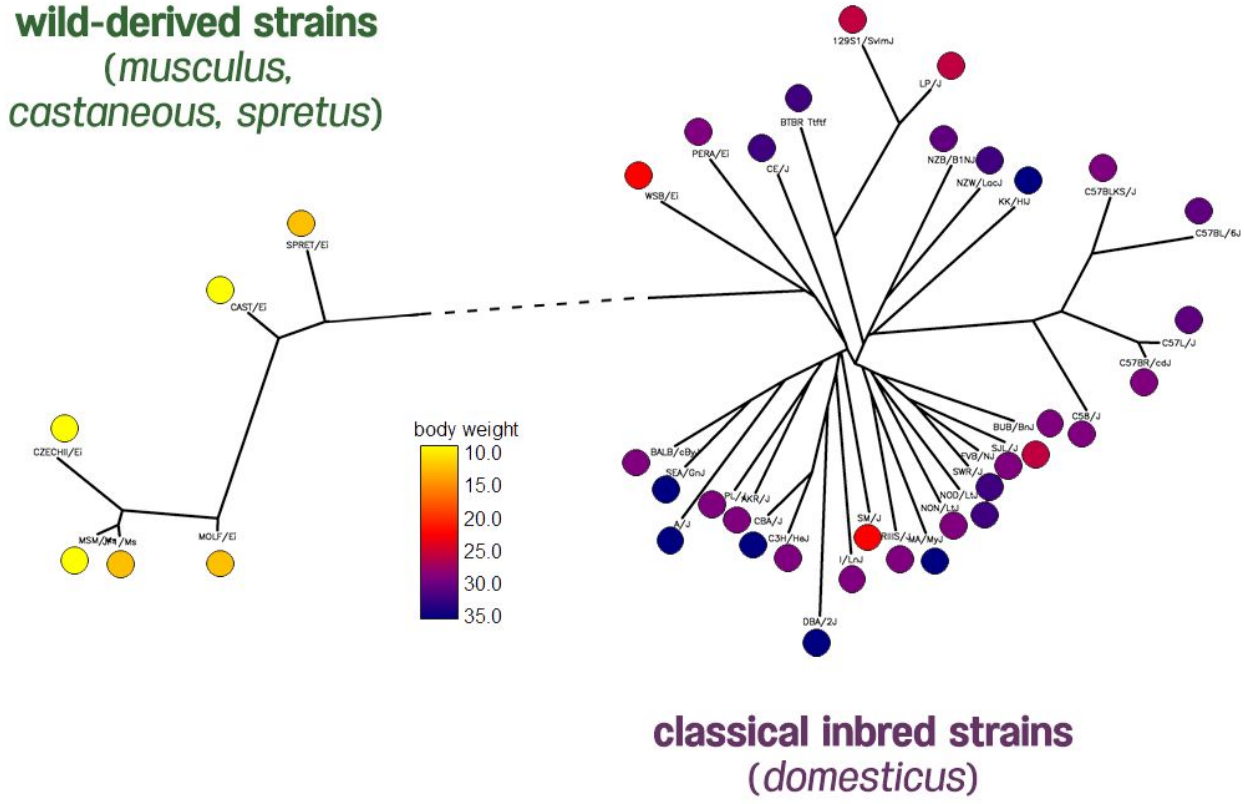
$$h^2 - h^2_{\mathtt{QTL}} > 0$$

Eskin (2105) http://dx.doi.org/10.1145/2817827

population structure inflates effects

- spurious association of phenotype
- population structure affects
  - some phenotypes
  - some genetic loci
- but genotype may not affect phenotype
  - if we adjust for population structure

mouse strains & body weight



Eskin (2106 draft) Review doi:10.1101/092106

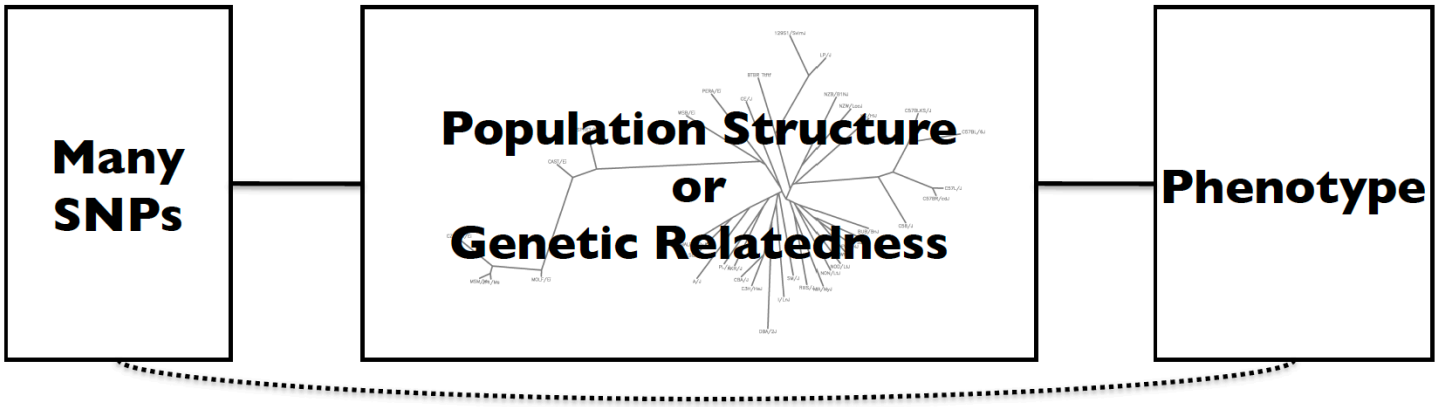
indirect correlation

SNPs and phenotypes become indirectly correlated

~~H_0: [Phenotype] ⊥ [SNP]~~

H_0: [Phenotype]~[SNP]

H_1: [Phenotype]~[SNP]



Eskin (2106 draft) Review doi:10.1101/092106

mixed model: QTL + poly

$$y = \mu_q + g + e$$

μ_q = QTL effects (fixed)
g = polygenic effects (random)

$$g \sim N(0, \sigma_g^2 K)$$

e = unexplained variation (random)

$$e \sim N(0, \sigma^2 I)$$

K = kinship matrix
I = identity matrix (1s on diagonal, 0s off diagonal)

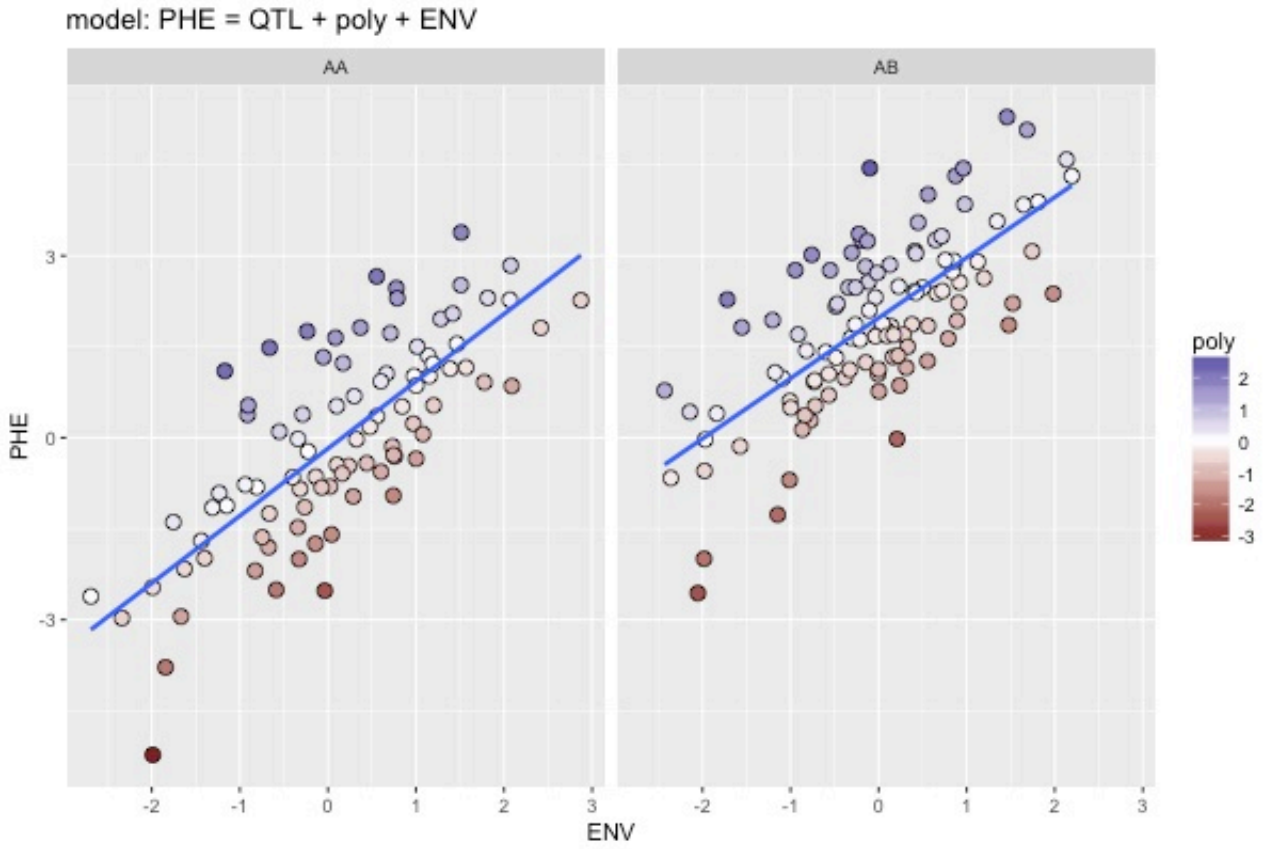
polygenes and kinship K

- estimate kinship K via pedigree: all we had in past
  - average / predicted relationship
  - works globally, might be inaccurate locally
  - think siblings vs parent/offspring
- estimate kinship K via SNP or GBS
  - estimate K from marker data M
  - in past, selected "neutral" markers
  - now use markers away from QTL q

$$K = c * M^{\mathrm{T}} M$$

c set so diagonal of K is 1

PHE = QTL + poly + ENV example



fitting the mixed model

distribution of phenotype

$$y = \mu_q + g + e$$

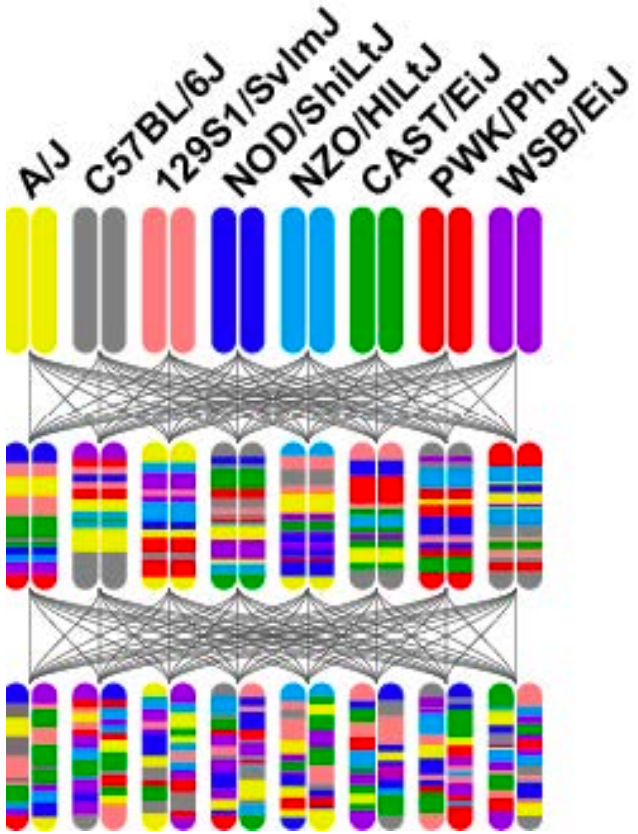
$$y \sim N(\mu_q, V), V = \sigma_g^2 K + \sigma^2 I$$

iterate to solve (similar to EM idea)

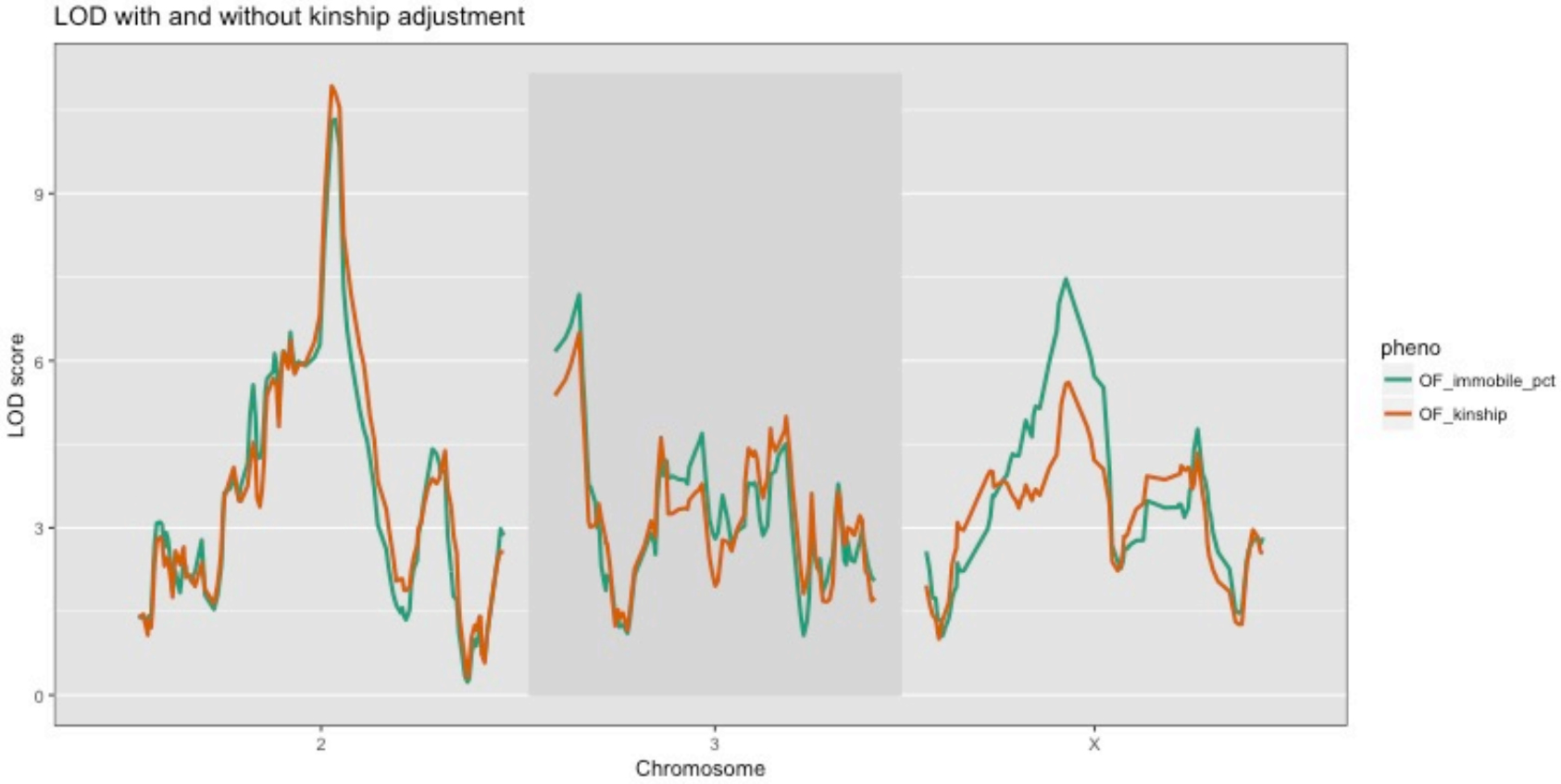
- get MLE of μ_q given V: $\hat{\mu}_q = (V^{\mathrm{T}} V)^{-1} V^{\mathrm{T}} y$
- estimate σ_g and σ^2 given μ_q

Diversity Outbred experiment

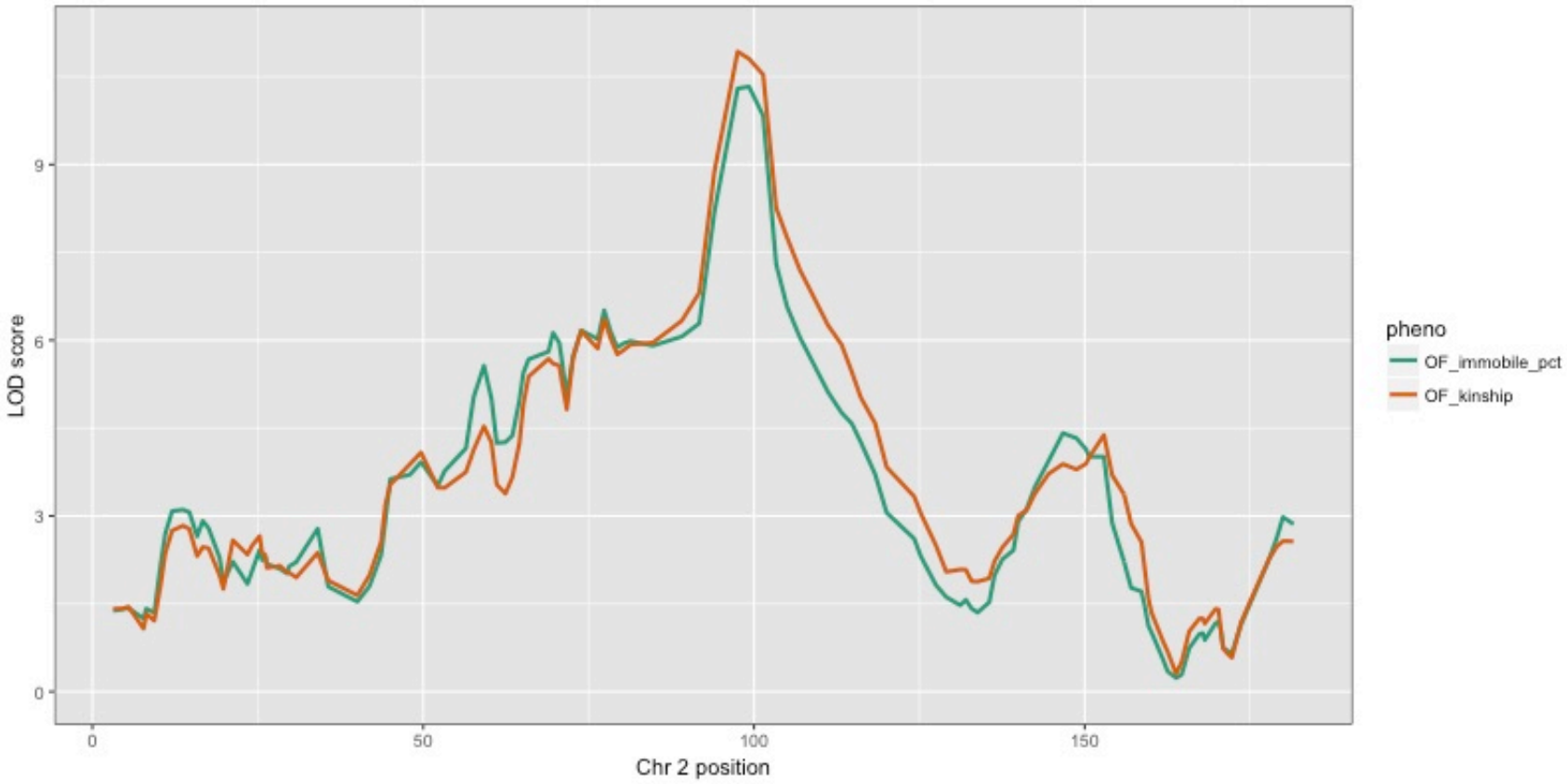
- 283 mice
- Diversity Outbred cross
- generations 4 & 5
- 320 (of 7851) SNP markers
- phenotype = `OF_immobile_pct`
- Data: https://github.com/rqtl/qtl2data/
- Recla, Robledo, Gatti, Bult, Churchill, Chesler (2014)



genome scans with kinship



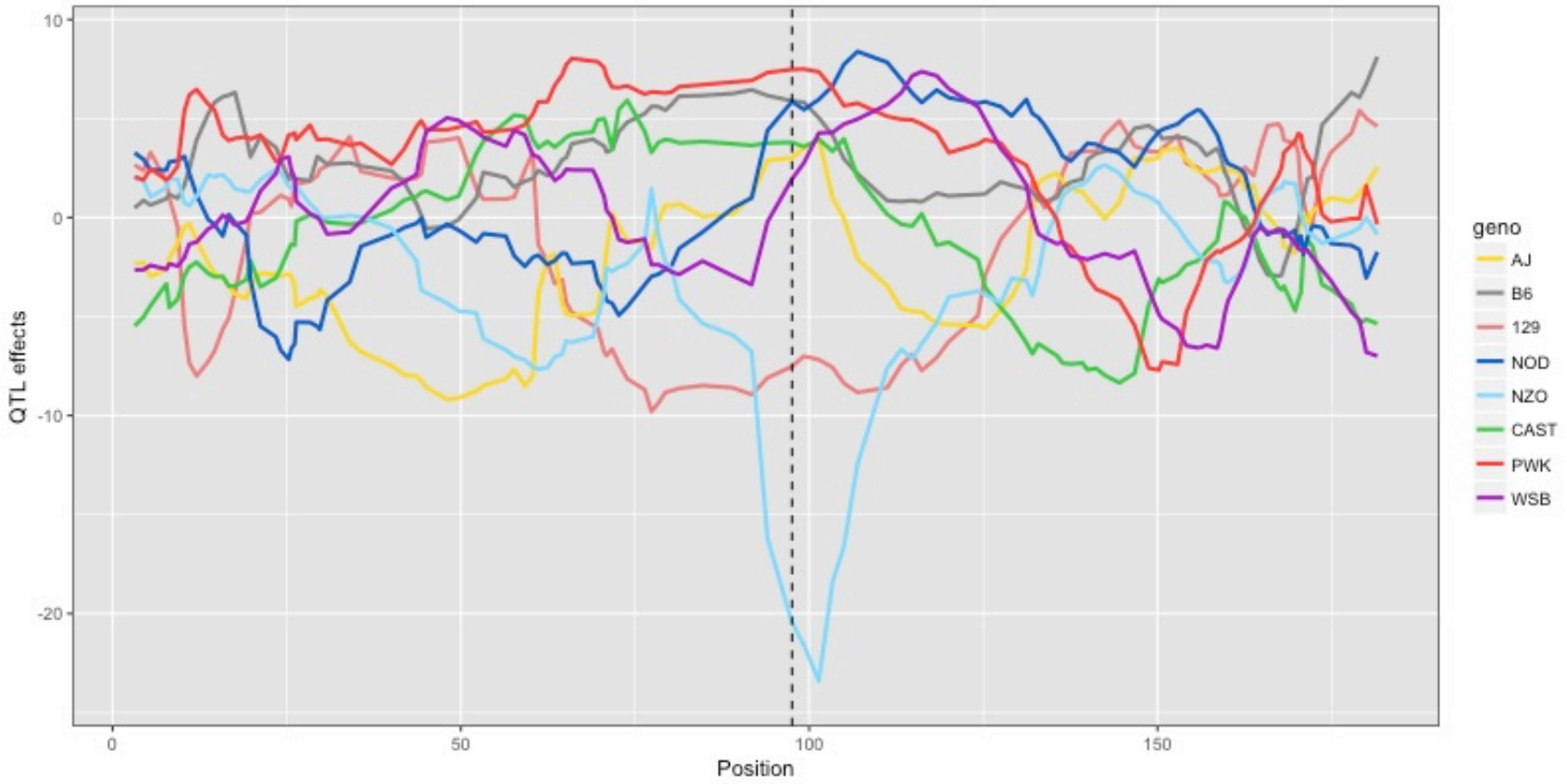
detail for key chromosome



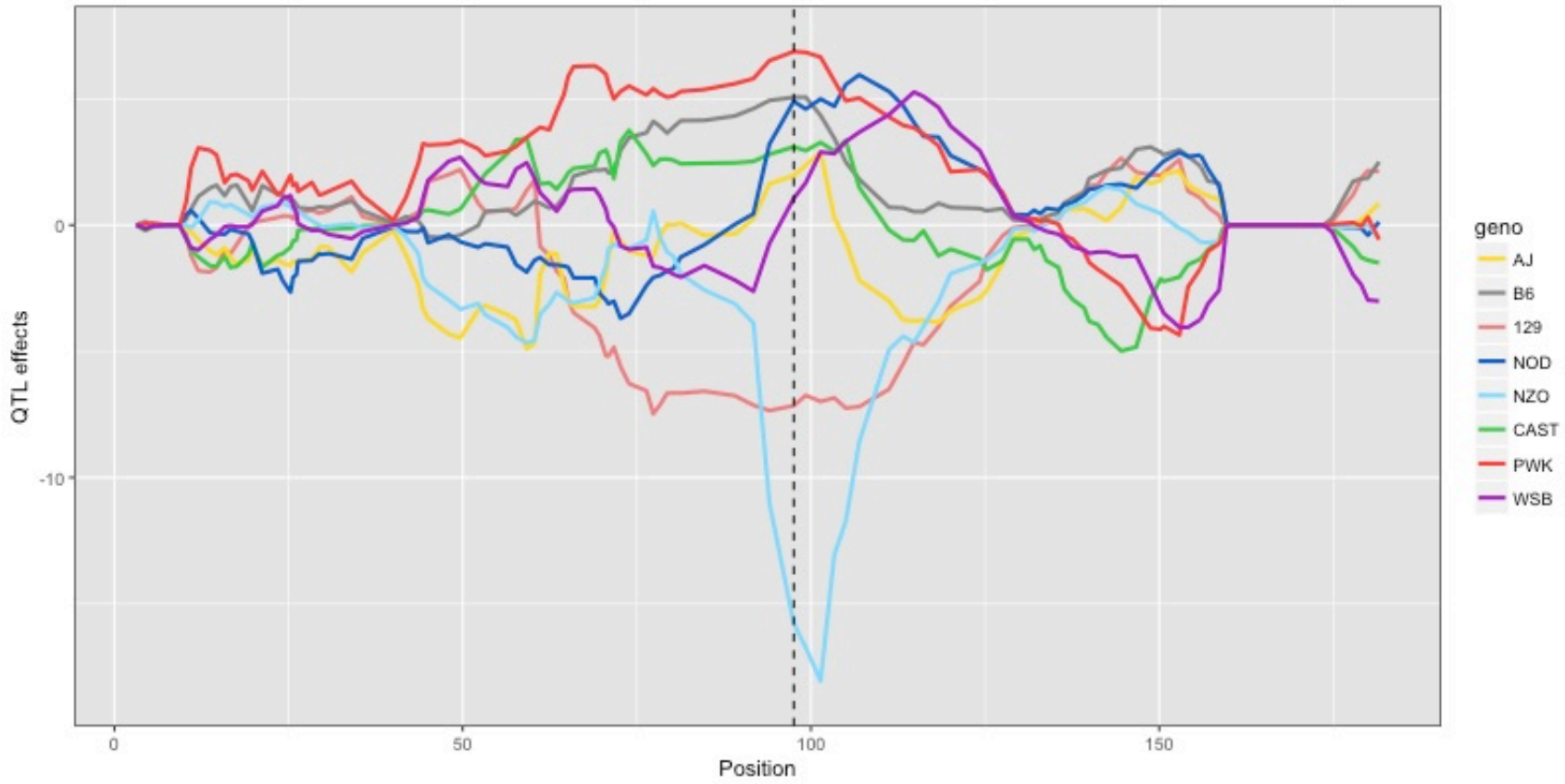
allele vs SNP scans

- allele-based genome scan: LOD maps
  - continuous curve across loci
  - interval mapping for missing data
  - model effect of founder alleles
- DO founder alleles: A,B,C,D,E,F,G,H
- response ~ sum of effects of alleles
- predict allele effects
  - naive: allele means based on geno probs
  - BLUP: predicted allele effects using kinship

naive allele scan



BLUP allele scan



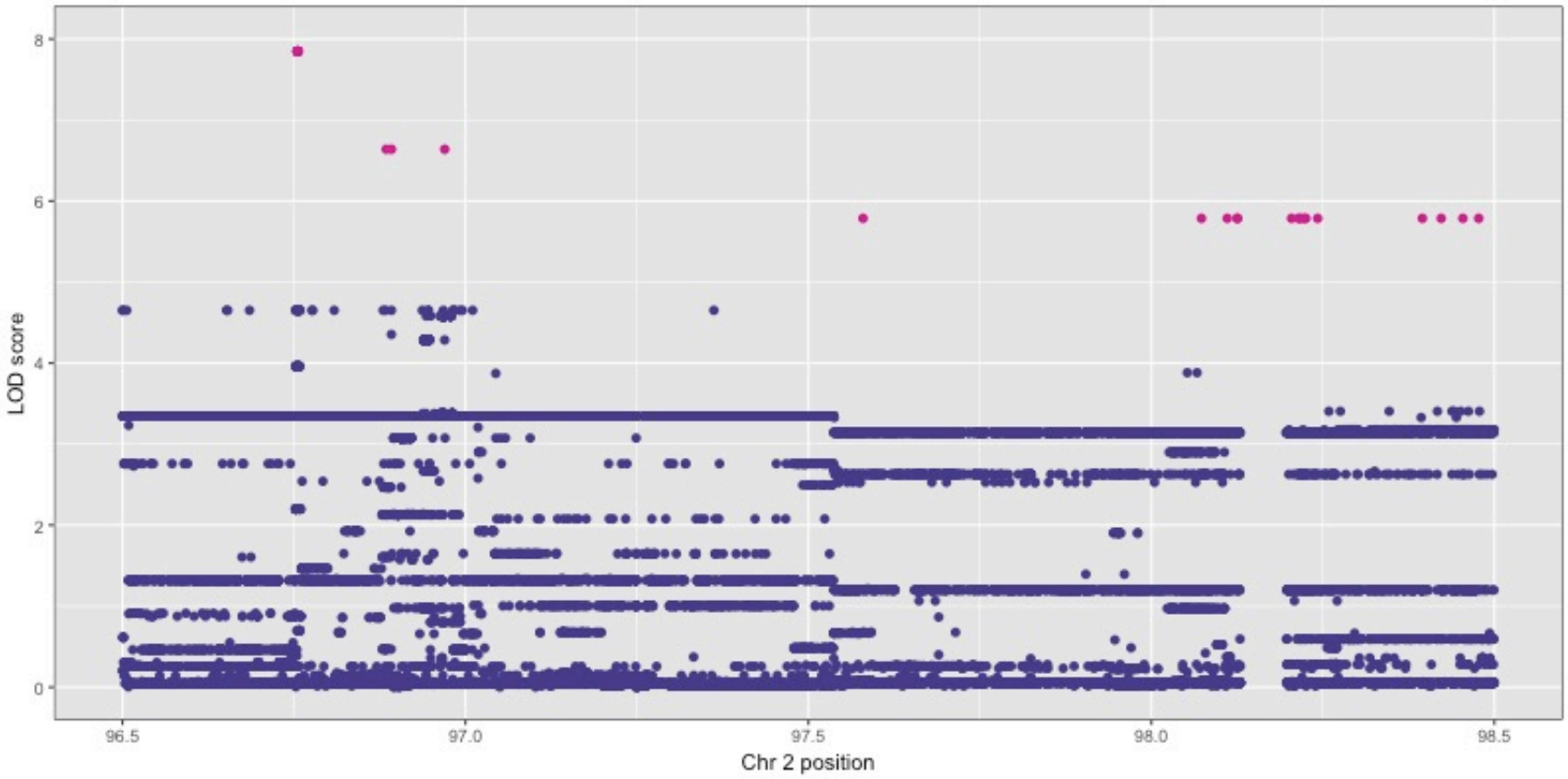
SNP association mapping

- SNP-based genome scan: GWAS Manhattan plots
  - discrete tests of SNPs or other features
  - typically 2 SNP alleles
  - model effect of number of non-ref SNP copies
- SNP recorded as pair of DNA base pairs (A,C,G,T)
  - SNPs typically have two values (G/T)
  - individual has genotype GG, GT or TT
- account for other associations with kinship
  - effects of other SNPs
  - population structure

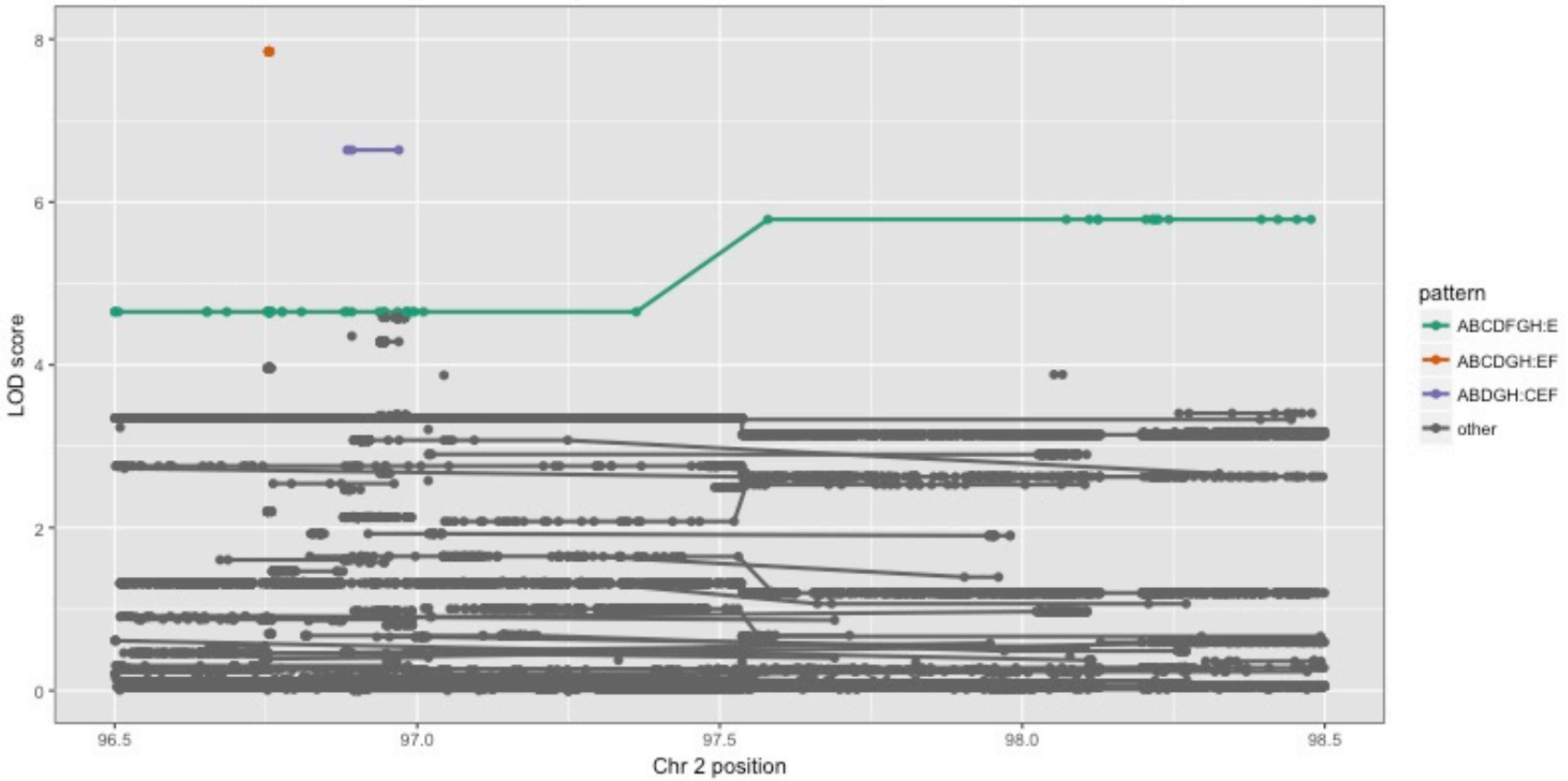
Detailed look in region of LOD peak

- consider SNPs within region of LOD peak
- use SNPs to refine search
  - relate SNPs to genomic features
  - compare SNP pattern across founders
  - DO reference is B = B6
  - s = 0,1,2 copies of non-reference nucleotide
- mixed model $y = a + bs + g + e$
  - μ_q replaced by $a + bs$
- test slope $b = 0$ using LOD score

SNP scans



SNP patterns



SNP best patterns