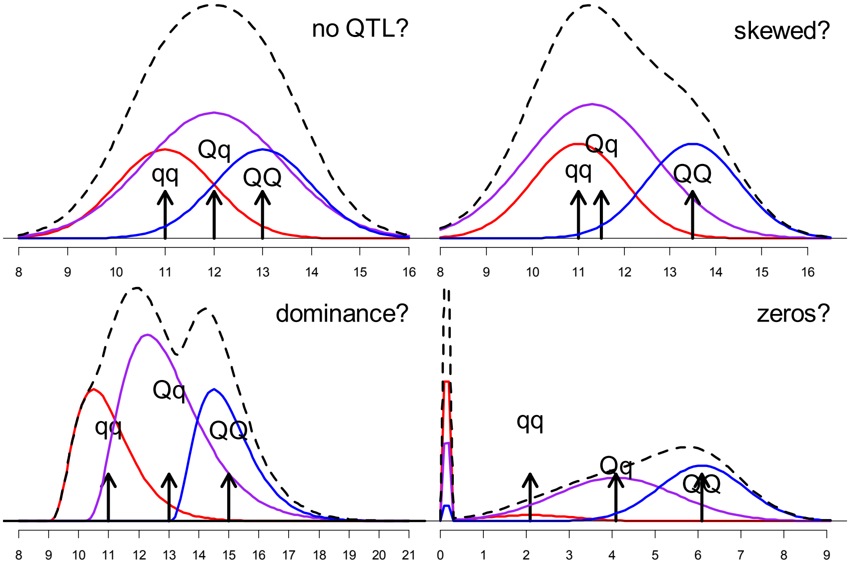
10 Extension of Phenotype Model

- limitations of parametric models
- diagnostic tools for QTL analysis
- QTL mapping with other parametric "families"
- quick fixes via data transformations
- semi-parametric approaches
- non-parametric approaches
- bottom line:
  - normal phenotype model works well to pick up loci, but may be poor at estimating effects if data not normal

limitations of parametric models

- measurements not normal
  - categorical traits: counts (*e.g.* number of tumors)
    - use methods specific for counts
    - binomial, Poisson, negative binomial
  - traits measured over time and/or space
    - survival time (*e.g.* days to flowering)
    - developmental process; signal transduction between cells
    - TP Speed (pers. comm.); Ma, Casella, Wu (2002)
- false positives due to miss-specified model
  - how to check model assumptions?
- want more robust estimates of effects
  - parametric: only center (mean), spread (SD)
  - shape of distribution may be important

what if data are far away from ideal?



diagnostic tools for QTL (Hackett 1997)

- illustrated with BC, adapt regression diagnostics
- normality & equal variance (fig. 1)
  - plot fitted values vs. residuals--football shaped?
  - normal scores plot of residuals--straight line?
- number of QTL: likelihood profile (fig. 2)
  - flat shoulders near LOD peak: evidence for 1 vs. 2 QTL
- genetic effects
  - effect estimate near QTL should be (1–2*r*)*a*
  - plot effect vs. location

marker density & sample size: 2 QTL

modest sample size
dense vs. sparse markers

large sample size
dense vs. sparse markers

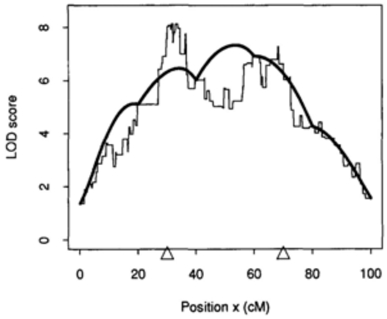


FIGURE 1.—The two-QTL true model with a QTL at 30 cM and a second QTL of somewhat smaller effect at 70 cM (true locations indicated by △). A normal single-QTL model is assumed and the LOD score for 100 simulated individuals is given for dense markers (thin curve) and markers at 20-cM intervals (bold curve).

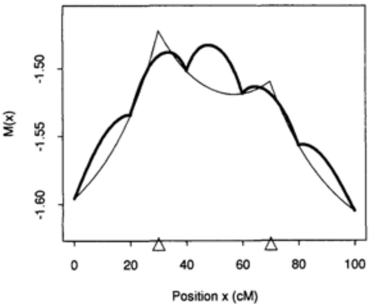


FIGURE 4.—$M(x)$ for a normal single-QTL assumed model under a two-QTL true model when both of the genes lie on the chromosome under study. This scenario was originally depicted in Figure 1. With dense markers (thin curve), $M(x)$ peaks at exactly 30 cM, the location of the QTL of stronger effect. With nondense markers at 20-cM intervals, $M(x)$ peaks at 47 cM in an incorrect interval (bold curve). Note the similarity in shape between the LODs in Figure 1 and the limiting forms depicted here.

Wright Kong (1997 *Genetics*)

robust locus estimate for non-normal phenotype

large sample size &
dense marker map:
no need for normality

but what happens for
modest sample sizes?

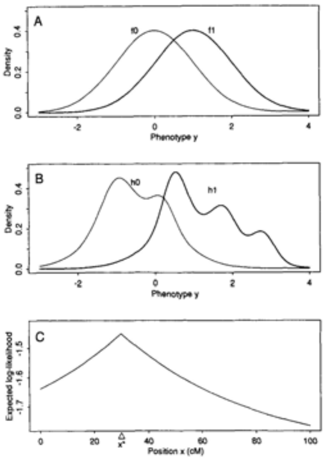


FIGURE 2.—Misspecification of the phenotype model. (A) The assumed distributions f_0 and f_1. (B) The true distributions h_0, h_1. (C) The expected log-likelihood across the chromosome when the markers are dense. Despite the misspecification, the function is maximized at exactly the true location $x^* = 30$ cM (indicated by △).

Wright Kong (1997 *Genetics*)

What shape is your histogram?

- histogram conditional on known QT genotype
  - – $\text{pr}(Y|\text{qq},\theta)$ model shape with genotype qq
  - – $\text{pr}(Y|\text{Qq},\theta)$ model shape with genotype Qq
  - – $\text{pr}(Y|\text{QQ},\theta)$ model shape with genotype QQ
- is the QTL at a given locus λ?
  - – no QTL $\text{pr}(Y|\text{qq},\theta) = \text{pr}(Y|\text{Qq},\theta) = \text{pr}(Y|\text{QQ},\theta)$
  - – QTL present mixture if genotype unknown
- mixture across possible genotypes
  - – sum over Q = qq, Qq, QQ
  - – $\text{pr}(Y|X,\lambda,\theta) = \text{sum}_Q\ \text{pr}(Q|X,\lambda)\ \text{pr}(Y|Q,\theta)$

interval mapping likelihood

- likelihood: basis for scanning the genome
  - – product over $i = 1,\ldots,n$ individuals

$$L(\theta,\lambda|Y) = \text{product}_i\ \text{pr}(Y_i|X_i,\lambda)$$
$$= \text{product}_i\ \text{sum}_Q\ \text{pr}(Q|X_i,\lambda)\ \text{pr}(Y_i|Q,\theta)$$

- problem: unknown phenotype model
  - – parametric $\text{pr}(Y|Q,\theta) = f(Y \mid \mu, G_Q, \sigma^2)$
  - – semi-parametric $\text{pr}(Y|Q,\theta) = f(Y)\exp(Y\beta_Q)$
  - – non-parametric $\text{pr}(Y|Q,\theta) = F_Q(Y)$

useful models & transformations

- binary trait (yes/no, hi/lo, …)
  - map directly as another marker
  - categorical: break into binary traits?
  - mixed binary/continuous: condition on $Y > 0$?
- known model for biological mechanism

– counts	Poisson
– fractions	binomial
– clustered	negative binomial

- transform to stabilize variance

– counts	$\surd Y = \text{sqrt}(Y)$
– concentration	$\log(Y)$ or $\log(Y+c)$
– fractions	$\arcsin(\surd Y)$

- transform to symmetry (approx. normal)

– fraction	$\log(Y/(1-Y))$ or $\log((Y+c)/(1+c-Y))$

- empirical transform based on histogram
  - watch out: hard to do well even without mixture
  - probably better to map untransformed, then examine residuals

semi-parametric QTL

- phenotype model $\text{pr}(Y|Q,\theta) = f(Y)\exp(Y\beta_Q)$
  - unknown parameters $\theta = (f, \beta)$
    - $f\,(Y)$ is a (unknown) density if there is no QTL
    - $\beta = (\beta_{\text{qq}}, \beta_{\text{Qq}}, \beta_{\text{QQ}})$
    - $\exp(Y\beta_Q)$ \`tilts' f based on genotype Q and phenotype Y
- test for QTL at locus λ
  - $\beta_Q = 0$ for all Q, or $\text{pr}(Y|Q,\theta) = f(Y)$
- includes many standard phenotype models

normal	$\text{pr}(Y|Q,\theta) = N(G_{\text{Q}},\sigma^2)$
Poisson	$\text{pr}(Y|Q,\theta) = \text{Poisson}(G_{\text{Q}})$

exponential, binomial, …, but not negative binomial

QTL for binomial data

- approximate methods: marker regression
  - Zeng (1993,1994); Visscher et al. (1996); McIntyre et al. (2001)
- interval mapping, CIM
  - Xu Atchley (1996); Yi Xu (2000)
  - $Y \sim$ binomial$(1,\pi)$, π depends on genotype Q
  - pr$(Y|Q) = (\pi_Q)^Y (1 - \pi_Q)^{(1-Y)}$
  - substitute this phenotype model in EM iteration
- or just map it as another marker!
  - but may have complex

EM algorithm for binomial QTL

- E-step: posterior probability of genotype Q

$$\text{pr}(Q \mid Y_i, X_i, \lambda, \pi_Q) = \frac{\text{pr}(Q \mid X_i, \lambda)(\pi_Q)^{Y_i}(1-\pi_Q)^{(1-Y_i)}}{\text{sum}_Q \text{ of numerator}}$$

- M-step: MLE of binomial probability π_Q

$$\pi_Q = \frac{\text{sum}_i\ Y_i \text{pr}(Q \mid Y_i, X_i, \lambda, \pi_Q)}{\text{sum}_i\ \text{pr}(Q \mid Y_i, X_i, \lambda, \pi_Q)}$$

threshold or latent variable idea

- "real", unobserved phenotype Z is continuous
- observed phenotype Y is ordinal value
  - no/yes; poor/fair/good/excellent
  - $\text{pr}(Y = j) = \text{pr}(\tau_{j-1} < \text{Z} \leq \tau_j)$
  - $\text{pr}(Y \leq j) = \text{pr}(\text{Z} \leq \tau_j)$
- use logistic regression idea (Hackett Weller 1995)
  - substitute new phenotype model in to EM algorithm
  - or use Bayesian posterior approach
  - extended to multiple QTL (papers in press)

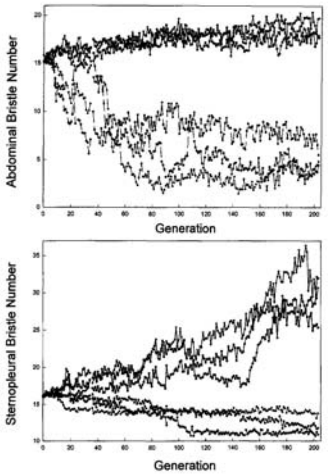
$$\text{pr}(Y \leq j \mid Q) = \text{pr}(Z \leq \tau_j \mid Q) = [1 + \exp(\mu + G_Q - \tau_j)]^{-1}$$

quantitative & qualitative traits

- Broman (2003): spike in phenotype
  - large fraction of phenotype has one value
  - map binary trait (is/is not that value)
  - map continuous trait given not that value
- multiple traits
  - Williams et al. (1999)
    - multiple binary & normal traits
    - variance component analysis
  - Corander Sillanpaa (2002)
    - multiple discrete & continuous traits
    - latent (unobserved) variables

other parametric approaches

- Poisson counts
  - Mackay Fry (1996)
    - trait = bristle number
  - Shepel et al (1998)
    - trait = tumor count
- negative binomial
  - Lan *et al.* (2001)
    - number of tumors
- exponential
  - Jansen (1992)



Mackay Fry (1996 *Genetics*)

semi-parametric empirical likelihood

- phenotype model $\mathrm{pr}(Y|Q,\theta) = f(Y)\ \exp(Y\beta_Q)$
  - "point mass" at each measured phenotype Y_i
  - subject to distribution constraints for each Q:
    $1 = \mathrm{sum}_i\, f(Y_i)\ \exp(Y_i\beta_Q)$
- non-parametric empirical likelihood (Owen 1988)

  $$\begin{aligned} L(\theta,\lambda|Y,X) &= \mathrm{product}_i\ [\mathrm{sum}_Q\ \mathrm{pr}(Q|X_i,\lambda)\, f(Y_i)\ \exp(Y_i\beta_Q)] \\ &= \mathrm{product}_i\, f(Y_i)\ [\mathrm{sum}_Q\ \mathrm{pr}(Q|X_i,\lambda)\ \exp(Y_i\beta_Q)] \\ &= \mathrm{product}_i\, f(Y_i)\ w_i \end{aligned}$$

  - weights $w_i = w(Y_i|X_i,\beta,\lambda)$ rely only on flanking markers
    - 4 possible values for BC, 9 for F2, etc.
- profile likelihood: $L(\lambda|Y,X) = \max_\theta L(\theta,\lambda|Y,X)$

semi-parametric formal tests

- clever trick: use partial empirical LOD
  - Zou, Fine, Yandell (2002 *Biometrika*)
  - Lange, Whittaker (2001 *Genetics*) GEE
- has same formal behavior as parametric LOD
  - single locus test: approximately χ^2 with 1 d.f.
  - genome-wide scan: can use same critical values
  - permutation test: possible with some work
- can estimate cumulative distributions
  - nice properties (converge to Gaussian processes)

log empirical likelihood details

$\log(L(\theta,\lambda|Y,X)) = \text{sum}_i \log(f(Y_i)) + \log(w_i)$

now profile with respect to β,λ

$\log(L(\beta,\lambda|Y,X)) = \text{sum}_i \log(f_i) + \log(w_i)$

$+ \text{sum}_Q\, \alpha_Q(1 - \text{sum}_i f_i \exp(Y_i\beta_Q))$

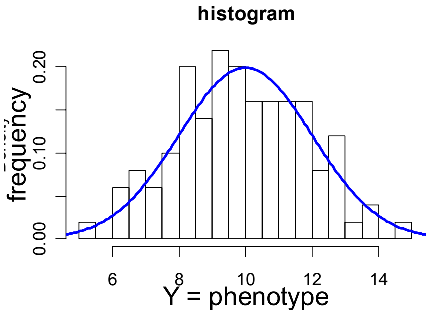
partial likelihood: set Lagrange multipliers α_Q to 0

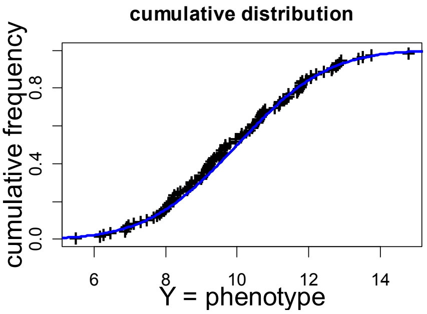
point mass density estimates

$$f_i = \left[\text{sum}_Q \exp(Y_i\beta_Q)\, p(Q \mid X, \lambda)\right]^{-1}$$

$$\text{with } p(Q \mid X, \lambda) = \text{sum}_i \text{pr}(Q \mid X_i, \lambda)$$

histograms and CDFs



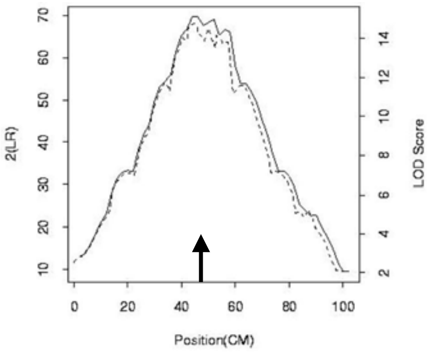


histograms capture shape but are not very accurate

CDFs are more accurate but not always intuitive

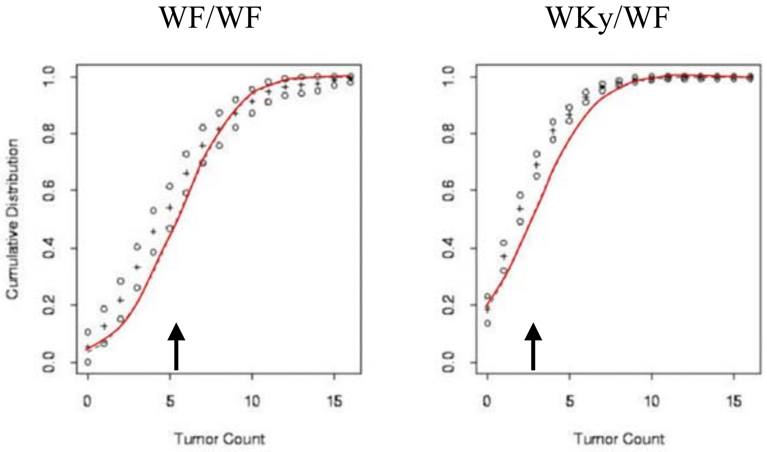
rat study of breast cancer Lan *et al.* (2001 *Genetics*)

- rat backcross
  - two inbred strains
    - Wistar-Furth susceptible
    - Wistar-Kyoto resistant
  - backcross to WF
  - 383 females
  - chromosome 5, 58 markers
- search for resistance genes
- *Y* = # mammary carcinomas
- where is the QTL?



dash = normal
solid = semi-parametric

what shape histograms by genotype?


WF/WF

WKy/WF

line = normal, + = semi-parametric, o = confidence interval

non-parametric methods

- phenotype model pr($Y|Q,\theta$) = $F_Q(Y)$
  - $\theta = F = (F_{qq},F_{Qq},F_{QQ})$ arbitrary distribution functions
- interval mapping Wilcoxon rank-sum test
  - replaced Y by rank(Y)
    - (Kruglyak Lander 1995; Poole Drinkwater 1996; Broman 2003)
  - claimed no estimator of QTL effects
- non-parametric shift estimator
  - semi-parametric shift (Hodges-Lehmann)
    - Zou (2001) thesis, Zou, Yandell, Fine (2002 in review)
  - non-parametric cumulative distribution
    - Fine, Zou, Yandell (2001 in review)
- stochastic ordering (Hoff et al. 2002)

rank-sum QTL methods

- phenotype model pr($Y|Q,\theta$) = $F_Q(Y)$
- replace Y by rank(Y) and perform IM
  - – extension of Wilcoxon rank-sum test
  - – fully non-parametric (Kruglyak Lander 1995; Poole Drinkwater 1996)
- Hodges-Lehmann estimator of shift β
  - – most efficient if pr($Y|Q,\theta$) = $F(Y+Q\beta)$
  - – find β that matches medians
    - problem: genotypes Q unknown
    - resolution: Haley-Knott (1992) regression scan
  - – works well in practice, but theory is elusive
    - Zou, Yandell Fine (*Genetics*, in review)

non-parametric QTL CDFs

- estimate non-parametric phenotype model
  - – cumulative distributions $F_Q(y) = \text{pr}(Y \leq y\ |Q)$
  - – can use to check parametric model validity
- basic idea:

  $\text{pr}(Y \leq y\ |X,\lambda) = \text{sum}_Q\ \text{pr}(Q|X,\lambda)F_Q(y)$
  - – depends on X only through flanking markers
  - – few possible flanking marker genotypes
    - 4 for BC, 9 for F2, etc.

finding non-parametric QTL CDFs

- cumulative distribution $F_Q(y) = \mathrm{pr}(Y \leq y \mid Q)$
- $F = \{F_Q$, all possible QT genotypes $Q\}$
  - BC with 1 QTL: $F = \{F_{\mathrm{QQ}}, F_{\mathrm{Qq}}\}$
- find F to minimize over all phenotypes y

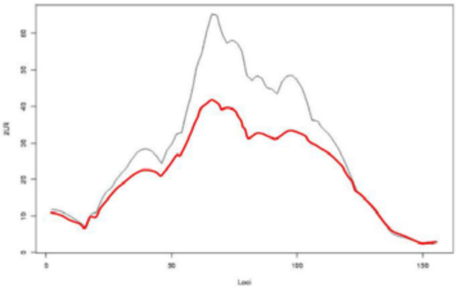
  $\mathrm{sum}_i\ [I(Y_i \leq y) - \mathrm{sum}_Q\ \mathrm{pr}(Q|X,\lambda)F_Q(y)]^2$
- looks complicated, but simple to implement

non-parametric CDF properties

- readily extended to censored data
  - time to flowering for non-vernalized plants
- nice large sample properties
  - estimates of $F(y) = \{F_Q(y)\}$ jointly normal
  - point-wise, experiment-wise confidence bands
- more robust to heavy tails and outliers
- can use to assess parametric assumptions

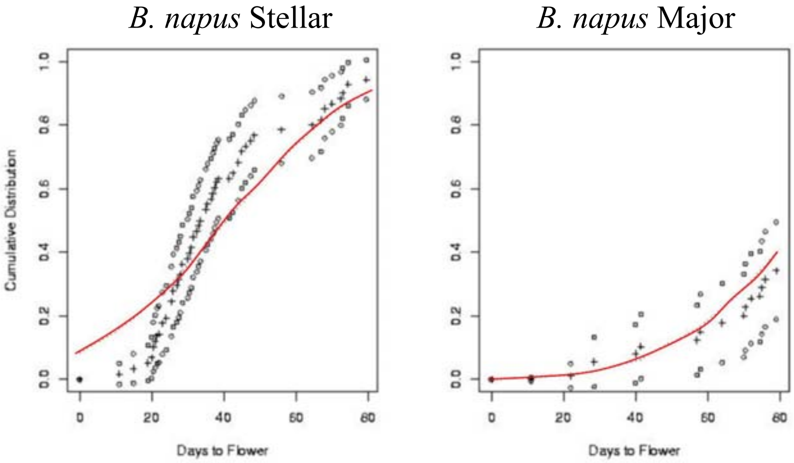
what QTL influence flowering time? no vernalization: censored survival

- *Brassica napus*
  - Major female
    - needs vernalization
  - Stellar male
    - insensitive
  - 99 double haploids
- *Y* = log(days to flower)
  - over 50% Major at QTL never flowered
  - log not fully effective



grey = normal, red = non-parametric

what shape is flowering distribution?



line = normal, + = non-parametric, o = confidence interval