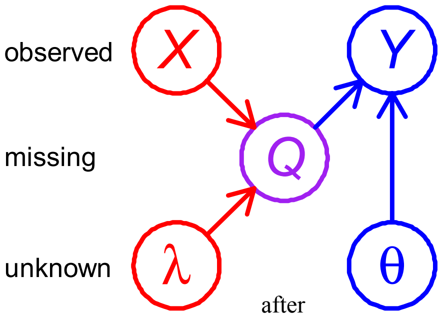
2 Key Statistical Issues for QTL

- general notation and data structure
- recombination model
  - two linked markers
  - flanking markers to a QTL
  - map distance and map functions
- modelling the phenotype
  - phenotype model
  - model likelihood
  - Bayesian posterior
- missing data concepts and algorithms
- model selection

interval mapping basics

- observed measurements
  - Y = phenotypic trait
  - X = markers & linkage map
    - i = individual index 1,…,n
- missing data
  - missing marker data
  - Q = QT genotypes
    - alleles QQ, Qq, or qq at locus
- unknown genetic architecture
  - λ = QT locus (or loci)
  - θ = genetic action
  - m = number of QTL
- pr($Q|X,\lambda,m$) recombination model
  - grounded by linkage map, experimental cross
  - recombination yields multinomial for Q given X
- pr($Y|Q,\theta,m$) phenotype model
  - distribution shape (assumed normal here)
  - unknown parameters θ (could be non-parametric)



after
Sen Churchill (2001)

2.1 general notation and data structure

- *Y* = phenotype values
  - as concept and realized (observed) values
- *X* = marker genotypes
  - type of experimental cross
  - linkage map construction
    - marker orders, positions, linkage phases
  - observed marker genotypes (possibly with error)
- pr(*Y,X*) = joint probability
  - what we “know” about *Y* and *X* for this experiment
  - usually assume linkage map is “known”

conditional data likelihood

- condition on markers and linkage map

$$\mathrm{pr}(Y \mid X) = \frac{\mathrm{pr}(Y, X)}{\mathrm{pr}(X)}$$

- pr(*X*) comprises information on linkage map
  - not influenced by phenotype
  - thus can “ignore” for QTL purposes

unknown QTL genotypes

- usually have sparse linkage map of markers
  - want to condition on actual QTL genotype *Q*

    pr(*Y*|*Q*)
  - but actual QTL affecting phenotype not known
- need to consider all possibilities
  - average pr(*Y*|*Q*) over all possible genotypes *Q*
  - weight by recombination pr(*Q*|*X*)

$$\text{pr}(Y \mid X) = \sum_{Q} \text{pr}(Y \mid Q)\text{pr}(Q \mid X)$$

enter the (Greek) parameters

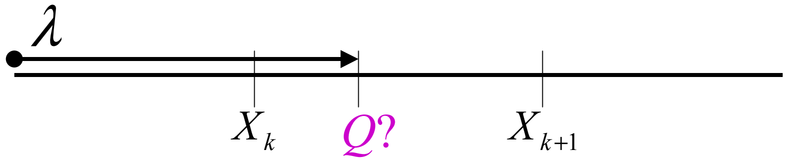
- θ = genetic effects, or gene action
  - additive, dominance, epistasis
  - may include reference values
    - grand mean (μ), environmental variance (σ^2)
- λ = location(s) of QTL
  - measured along “linear” genome
  - related to recombination and map distance

$$L(\theta,\lambda \mid Y,X) = \text{pr}(Y \mid X,\theta,\lambda) = \sum_{Q} \text{pr}(Y \mid Q,\theta)\text{pr}(Q \mid X,\lambda)$$

2.2 recombination model

- locus λ is distance along linkage map
  - identifies flanking marker region
- flanking markers provide good approximation
  - map assumed known from earlier study
  - inaccuracy slight using only flanking markers
    - extend to next flanking markers if missing data
  - could consider more complicated relationship
    - but little change in results

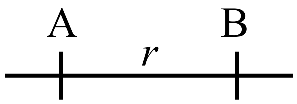
$$\text{pr}(Q|X,\lambda) = \text{pr}(\text{geno} \mid \text{map, locus}) \approx$$
$$\text{pr}(\text{geno} \mid \text{flanking markers, locus})$$



2.2.1 two linked markers

- backcross design
  - *n* individuals, 2 markers
  - follow one gamete
- recombinants
  - Ab, aB
  - $n_R = n_{12} + n_{21}$
- non-recombinants
  - ab, AB
  - $n_{NR} = n_{11} + n_{22}$
- recombination rate

$$\hat{r} = \frac{n_R}{n} = \frac{n_{12}+n_{21}}{n_{11}+n_{12}+n_{21}+n_{22}}$$



	b	B
a	$(1-r)/2$ n_{11}	$r/2$ n_{12}
A	$r/2$ n_{21}	$(1-r)/2$ n_{22}

NR	*R*
ab, AB	Ab, aB
$n_{NR} = n_{11} + n_{22}$	$n_R = n_{12} + n_{21}$

no linkage?

- test for no linkage: r = 1/2
- assumption: all individuals have same rate
  - implies binomial variation

$$\hat{r} = \frac{n_R}{n} = \frac{n_{12} + n_{21}}{n_{11} + n_{12} + n_{21} + n_{22}},\ \text{var}(\hat{r}) \approx \frac{\hat{r}(1-\hat{r})}{n}$$

- normal or chi-square test statistic

$$Z = \frac{\hat{r} - 1/2}{\sqrt{\text{var}(\hat{r})}} \sim N(0,1) \text{ or } \mathrm{Z}^2 = \frac{(n_R - n/2)^2}{(n_R n_{NR}/n)} \sim \chi_1^2$$

binomial probabilities

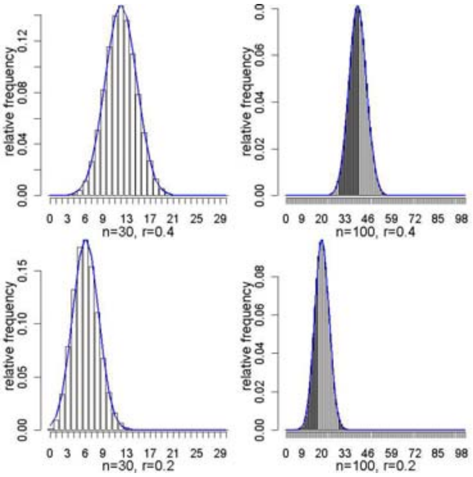
binomial prob

n = 30,100

r = 0.4,0.2

$$\text{pr}(n_R = k) = \binom{n}{k} r^k (1-r)^{n-k}$$

normal approx



likelihood ratio and LOD test

- likelihood for linked markers

$$L(r) = \mathrm{pr}(n_R \mid n, r) = Cr^{n_R}(1-r)^{n_{NR}}$$

- likelihood for unlinked markers

$$L(\tfrac{1}{2}) = C(\tfrac{1}{2})^n$$

- likelihood ratio and LOD

$$LR = 2^n(\hat{r})^{n_R}(1-\hat{r})^{n_{NR}}, G^2 = 2\log(LR) \sim \chi_1^2$$

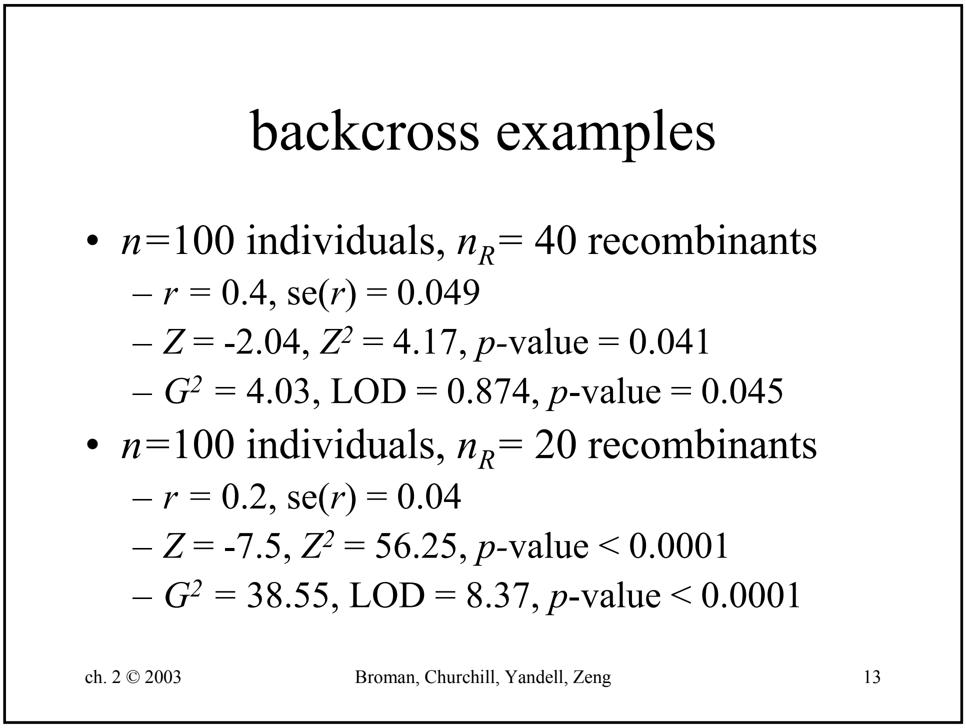
$$LOD = \log_{10}(LR) = \frac{G^2}{2\log(10)} = .217G^2$$

test statistic: distribution

- Z^2 and G^2 are generally close to each other
  - Z^2 based on properties of counts
  - G^2 and LOD based on likelihood principle
  - both have approximate chi-square distribution
- (non)central chi-square distribution

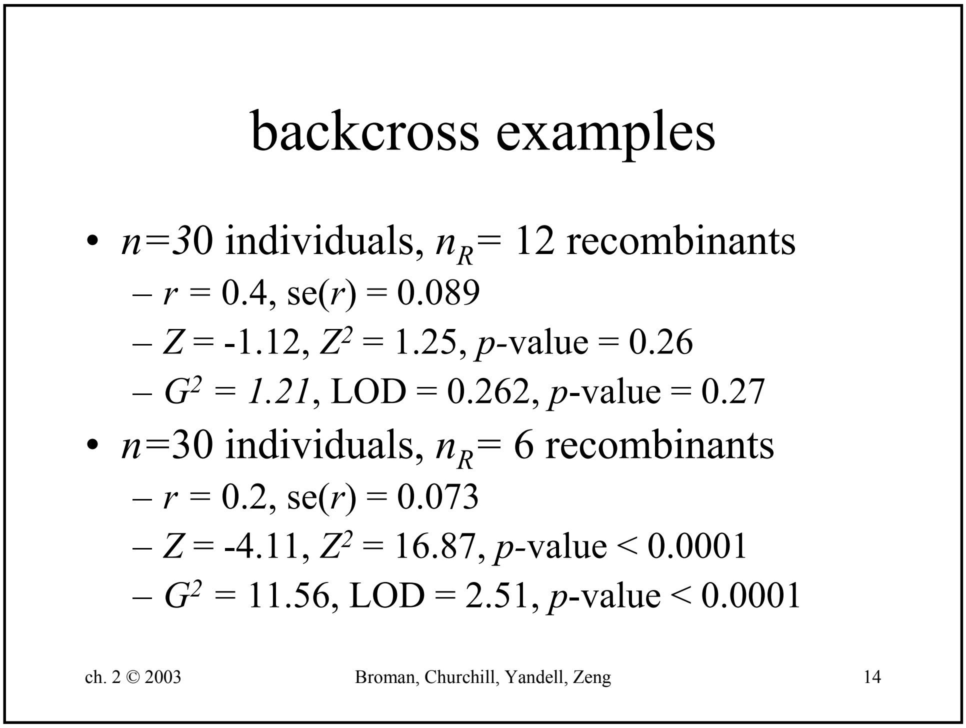
$$r = 0.5 : Z^2, G^2 \sim \chi_1^2$$

$$r < 0.5 : Z^2, G^2 \sim \chi_{1;ncp}^2, ncp = 4n(0.5-r)^2$$



backcross examples

- n=100 individuals, n_R= 40 recombinants
  - – r = 0.4, se(r) = 0.049
  - – Z = -2.04, Z^2 = 4.17, p-value = 0.041
  - – G^2 = 4.03, LOD = 0.874, p-value = 0.045
- n=100 individuals, n_R= 20 recombinants
  - – r = 0.2, se(r) = 0.04
  - – Z = -7.5, Z^2 = 56.25, p-value < 0.0001
  - – G^2 = 38.55, LOD = 8.37, p-value < 0.0001



backcross examples

- n=30 individuals, n_R= 12 recombinants
  - – r = 0.4, se(r) = 0.089
  - – Z = -1.12, Z^2 = 1.25, p-value = 0.26
  - – G^2 = 1.21, LOD = 0.262, p-value = 0.27
- n=30 individuals, n_R= 6 recombinants
  - – r = 0.2, se(r) = 0.073
  - – Z = -4.11, Z^2 = 16.87, p-value < 0.0001
  - – G^2 = 11.56, LOD = 2.51, p-value < 0.0001

simulations of LOD distribution

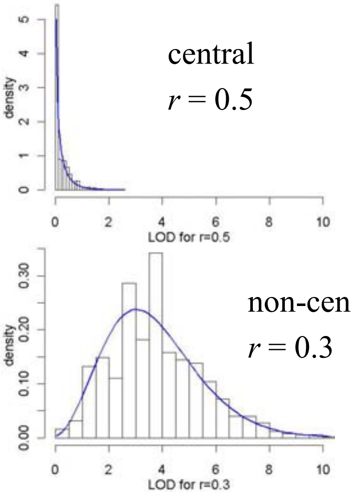
n=100,*r*=0.3,0.5

1000 samples

histogram

chi-square curve

rescaled by 2log(10)



central

r = 0.5

non-central

r = 0.3

LOD and LR over possible *r*

n = 30

n_R=12 or 6

evaluate at

all possible *r*

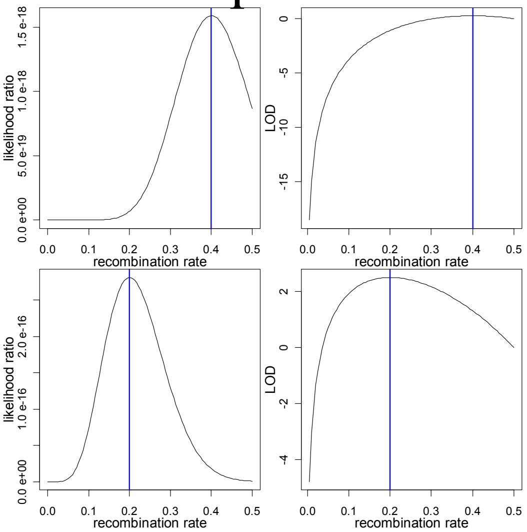
not just “best”

LR like a density

LOD is basis for

hypothesis test

estimate interval



LR, LOD and *p*-values

LR	LOD	*p*-value 1 d.f.	*p*-value 2 d.f.
10	1	0.0319	0.1
31.6	1.5	0.0086	0.0316
100	2	0.0024	0.01
1000	3	0.0002	0.001
10000	4	<0.0001	0.0001

LOD-based interval estimate for *r*

point estimate $\hat{r} = n_R / n$

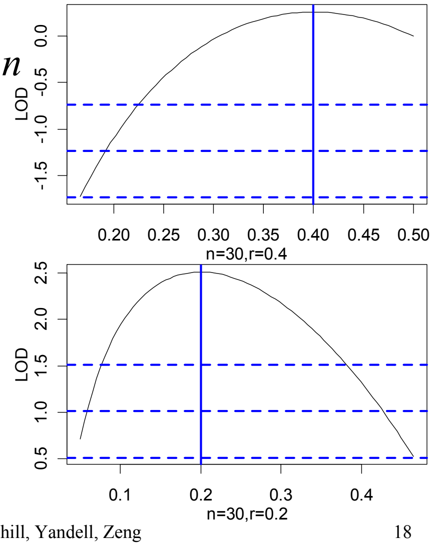
interval estimate

from LOD peak

down 1.5 LOD

(or 1 or 2 or …)

$n = 30$, $n_R = 12, 6$



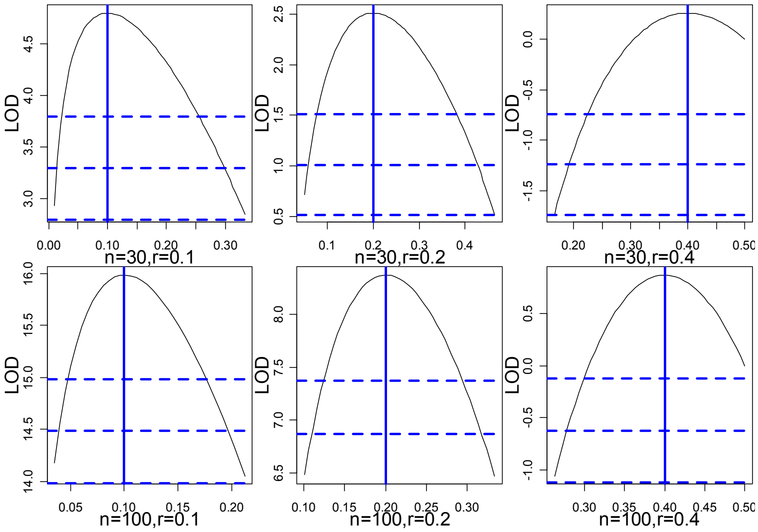
LOD-based interval calculations

confidence			96.8%	99.1%	99.76%
n	n_R	r	1 LOD	1.5 LOD	2 LOD
30	3	0.1	0.03-0.25	0.02-0.29	0.01-0.33
100	10	0.1	0.05-0.17	0.04-0.19	0.04-0.21
30	6	0.2	0.08-0.37	0.06-0.42	0.05-0.46
100	20	0.2	0.13-0.29	0.11-0.31	0.10-0.33
30	12	0.4	0.23-0.50	0.19-0.50	0.17-0.50
100	40	0.4	0.30-0.50	0.28-0.50	0.26-0.50

Note skew in intervals for small recombination rates.

Note upper boundary of 0.5.

LOD-based intervals



likelihood & Bayesian posterior

- recall the likelihood and likelihood ratio:

  $L(r) = \mathrm{pr}(n_R \mid n, r) = Cr^{n_R}(1-r)^{n_{NR}}$

  $LR(r) = 2^n r^{n_R}(1-r)^{n_{NR}}$
- posterior turns likelihood into a density
  - assume r may be any value <u>prior</u> to seeing data
  - posterior = likelihood x prior / constant

  $\mathrm{pr}(r \mid n, n_R) = L(r)/A$ or $= LR(r)/A$

  $A =$ area under likelihood or LR curve

  $\mathrm{sum}_r\ \mathrm{pr}(r \mid n, n_R) = 1$

LR and Bayes posterior

imagine LR as density

area under curve = 1

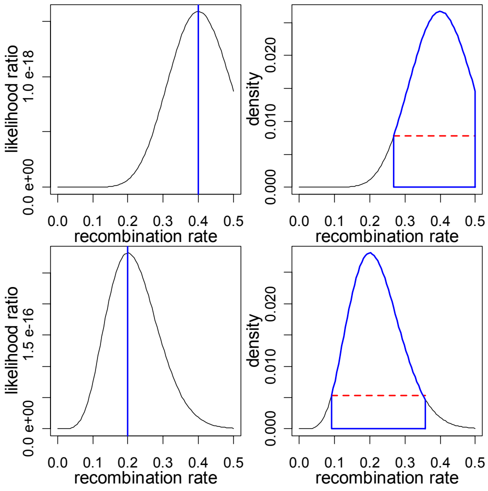
$\mathrm{pr}(r \mid n_R) = LR(r)/A$

what is probability that r is between 0.25 and 0.5?

where is interval with highest posterior mass? (HPD region)

example: n=30, n_R=12,6

95% HPD regions

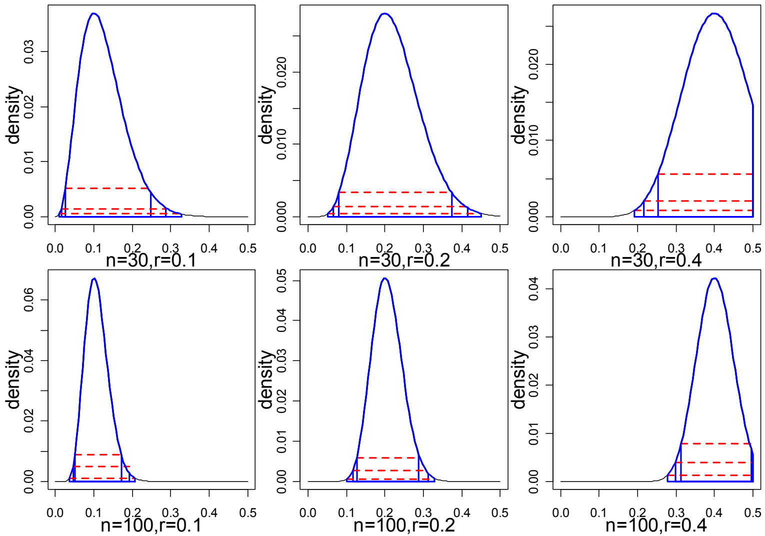


HPD-based interval calculations

HPD level			96.8%	99.1%	99.76%
n	n_R	*r*			
30	3	0.1	0.03-0.25	0.02-0.29	0.01-0.33
100	10	0.1	0.05-0.17	0.05-0.19	0.04-0.21
30	6	0.2	0.08-0.37	0.07-0.41	0.05-0.45
100	20	0.2	0.13-0.29	0.12-0.31	0.10-0.33
30	12	0.4	0.25-0.50	0.22-0.50	0.19-0.50
100	40	0.4	0.31-0.50	0.30-0.50	0.28-0.50

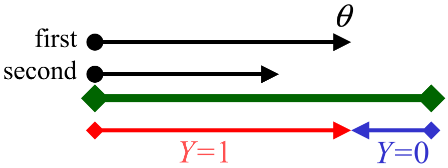
Note how these almost agree with LOD-based intervals.
Density height for HPD varies by *n* and *r*.

Bayesian posteriors for *r*



who was Bayes?

- 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)
- billiard balls on rectangular table
  - two balls tossed at random (uniform) on table
  - where is first ball if the second is to its left (right)?

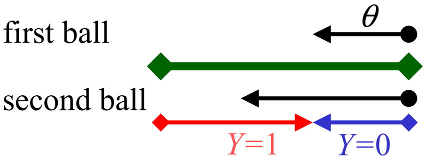


prior $\mathrm{pr}(\theta) = 1$

likelihood $\mathrm{pr}(Y|\theta) = \theta^Y(1-\theta)^{1-Y}$

posterior $\mathrm{pr}(\theta|Y) = ?$

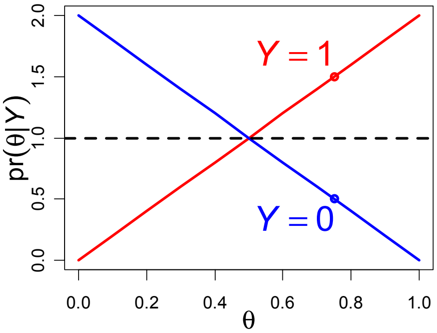
where is the first ball?



prior $\mathrm{pr}(\theta) = 1$

likelihood $\mathrm{pr}(Y|\theta) = \theta^Y(1-\theta)^{1-Y}$

posterior $\mathrm{pr}(\theta|Y) = ?$



$$\mathrm{pr}(\theta\,|\,Y) = \frac{\mathrm{pr}(Y\,|\,\theta)\mathrm{pr}(\theta)}{\mathrm{pr}(Y)}$$

$$\mathrm{pr}(Y) = \int_0^1 \theta^Y (1-\theta)^{1-Y} d\theta = \frac{1}{2}$$

$$\mathrm{pr}(\theta\,|\,Y) = \begin{cases} 2\theta & Y = 1 \\ 2(1-\theta) & Y = 0 \end{cases}$$

(now throw second ball n times)

what is Bayes theorem?

- before and after observing data
  - prior: $\mathrm{pr}(\theta)$ = pr(parameters)
  - posterior: $\mathrm{pr}(\theta|Y)$ = pr(parameters|data)
- posterior = likelihood * prior / constant
  - usual likelihood of parameters given data
  - normalizing constant pr(Y) depends only on data
    - constant often drops out of calculation

$$\mathrm{pr}(\theta\,|\,Y)=\frac{\mathrm{pr}(\theta,Y)}{\mathrm{pr}(Y)}=\frac{\mathrm{pr}(Y\,|\,\theta)\times\mathrm{pr}(\theta)}{\mathrm{pr}(Y)}$$

Bayes rule for recombination r

likelihood:

$$\mathrm{pr}(n_R\,|\,n,r)=L(r\,|\,n,n_R)=Cr^{n_R}(1-r)^{n_{NR}}$$

prior on recombination r:

$$\mathrm{pr}(a\le r\le b)=2(b-a)$$

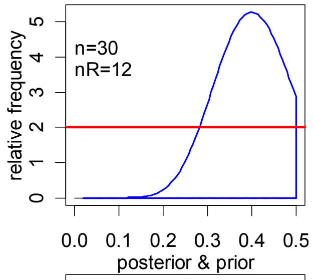
$$0\le a\le b\le 1/2$$

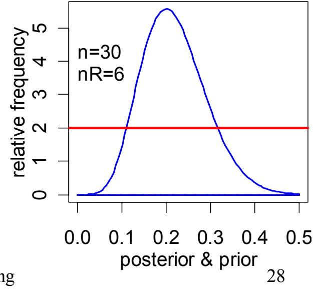
Bayes rule:

$$\mathrm{pr}(r\,|\,n,n_R)=\frac{\mathrm{pr}(n_R\,|\,n,r)\times\mathrm{pr}(r)}{\mathrm{pr}(n_R\,|\,n)}$$

normalizing constant:

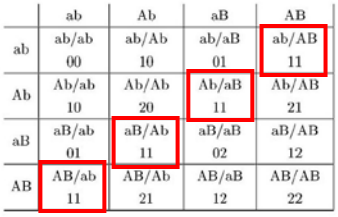
$$\mathrm{pr}(n_R\,|\,n)=\int_0^{1/2}\mathrm{pr}(n_R\,|\,n,r)2dr$$





two markers in F2 intercross

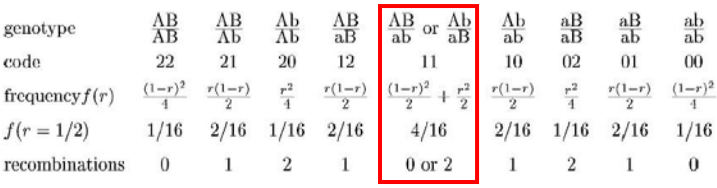
- two meioses
  - follow both gametes
  - 16 possibilities
- ambiguity with AaBb
  - 0 or 2 recombinations
- log likelihood ratio:



	ab	Ab	aB	AB
ab	ab/ab 00	ab/Ab 10	ab/aB 01	ab/AB 11
Ab	Ab/ab 10	Ab/Ab 20	Ab/aB 11	Ab/AB 21
aB	aB/ab 01	aB/Ab 11	aB/aB 02	aB/AB 12
AB	AB/ab 11	AB/Ab 21	AB/aB 12	AB/AB 22

lkjlj

$$\log LR = \text{sum}_x n_x \log\left(f_x(r) / f_x(0.5)\right)$$



genotype	$\frac{AB}{AB}$	$\frac{AB}{Ab}$	$\frac{Ab}{Ab}$	$\frac{AB}{aB}$	$\frac{AB}{ab}$ or $\frac{Ab}{aB}$	$\frac{Ab}{ab}$	$\frac{aB}{aB}$	$\frac{aB}{ab}$	$\frac{ab}{ab}$
code	22	21	20	12	11	10	02	01	00
frequency $f(r)$	$\frac{(1-r)^2}{4}$	$\frac{r(1-r)}{2}$	$\frac{r^2}{4}$	$\frac{r(1-r)}{2}$	$\frac{(1-r)^2}{2}+\frac{r^2}{2}$	$\frac{r(1-r)}{2}$	$\frac{r^2}{4}$	$\frac{r(1-r)}{2}$	$\frac{(1-r)^2}{4}$
$f(r=1/2)$	1/16	2/16	1/16	2/16	4/16	2/16	1/16	2/16	1/16
recombinations	0	1	2	1	0 or 2	1	2	1	0

two markers in F2 intercross

- γ= probability of double recombinant
  - for AaBb genotype, haplotype not known
  - need to "guess" the recombinant fraction of n_{11} offspring
- γ and r are inter-related
  - no "closed" solution, need to iterate
  - guess γ, use to estimate r, improve γ, etc.

$$\gamma = \text{pr}\left(\frac{\text{Ab}}{\text{aB}} \,\middle|\, \text{AaBb}\right) = \frac{r^2}{(1-r)^2 + r^2}$$

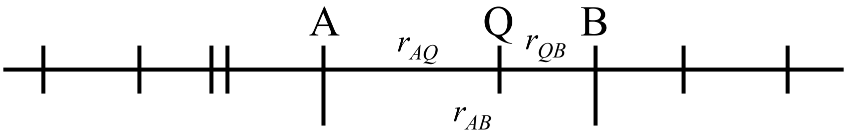
$$\hat{r} = \frac{1}{2n}\left[(n_{01} + n_{10} + n_{12} + n_{21}) + 2(n_{02} + n_{20} + \hat{\gamma}\, n_{11})\right]$$

EM algorithm for F2 recombination

- initial guess: $r = 0.5$, $\gamma = 0.5$
- Expectation (E) step
  - substitute expected values for nuisance γ
  - update γ given current value of r
- Maximization (M) step
  - maximize likelihood for parameter r
  - update r given current value of γ
- iterate E-step and M-step until “convergence”
  - stop when change in log-likelihood is small

    $$\log LR = \text{sum}_x n_x \log\left(f_x(r) / f_x(0.5)\right)$$
  - usually change in r is small at this point

2.2.2 flanking markers to QTL



- most genotype information is local
  - linkage drops off with distance
  - approximate by using only flanking markers
  - exception: linkage disequilibrium
    - different chromosome regions could be correlated
    - due to selection, etc.
    - not a problem for backcross or F2 intercross
- missing marker data: use next flanking marker

backcross QTL & flanking markers

1 meiosis

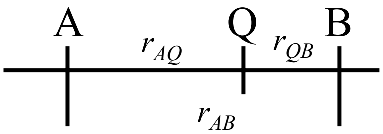
8 possible genotypes

3 recombination rates

small distances & rates?

no double crossovers

$\rho = r_{AQ} / r_{AB}$



	QTL genotype	
Marker genotype	QQ	Qq
AB/AB	$\frac{(1-r_{AQ})(1-r_{QB})}{1-r_{AB}} \approx 1$	$\frac{r_{AQ}r_{QB}}{1-r_{AB}} \approx 0$
Ab/AB	$\frac{(1-r_{AQ})r_{QB}}{r_{AB}} \approx 1-\rho$	$\frac{r_{AQ}(1-r_{QB})}{r_{AB}} \approx \rho$
aB/AB	$\frac{r_{AQ}(1-r_{QB})}{r_{AB}} \approx \rho$	$\frac{(1-r_{AQ})r_{QB}}{r_{AB}} \approx 1-\rho$
ab/AB	$\frac{r_{AQ}r_{QB}}{1-r_{AB}} \approx 0$	$\frac{(1-r_{AQ})(1-r_{QB})}{1-r_{AB}} \approx 1$

F2 QTL & flanking markers

2 meioses

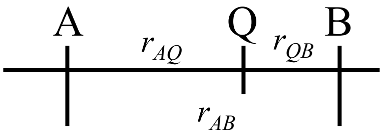
27 possible genotypes

3 recombination rates

EM steps on $\gamma$and r_{AB}

small distances & rates?

no double crossovers



$$\rho = r_{AQ} / r_{AB}, \gamma = \frac{r_{AB}^2}{(1-r_{AB}^2)+r_{AB}^2}$$

QTL	flanking marker genotypes								
	$\frac{\mathrm{AB}}{\mathrm{AB}}$	$\frac{\mathrm{AB}}{\mathrm{Ab}}$	$\frac{\mathrm{Ab}}{\mathrm{Ab}}$	$\frac{\mathrm{AB}}{\mathrm{aB}}$	$\frac{\mathrm{AB}}{\mathrm{ab}}$ or $\frac{\mathrm{Ab}}{\mathrm{aB}}$	$\frac{\mathrm{Ab}}{\mathrm{ab}}$	$\frac{\mathrm{aB}}{\mathrm{aB}}$	$\frac{\mathrm{aB}}{\mathrm{ab}}$	$\frac{\mathrm{ab}}{\mathrm{ab}}$
Q/Q	1	$1-\rho$	$(1-\rho)^2$	ρ	$\gamma\rho(1-\rho)$	0	ρ^2	0	0
Q/q	0	ρ	$2\rho(1-\rho)$	$1-\rho$	$(1-\gamma)+$ $\gamma[\rho^2+(1-\rho)^2]$	$1-\rho$	$2\rho(1-\rho)$	ρ	0
q/q	0	0	ρ^2	0	$\gamma\rho(1-\rho)$	ρ	$(1-\rho)^2$	$1-\rho$	1

2.2.3 map distance & map functions

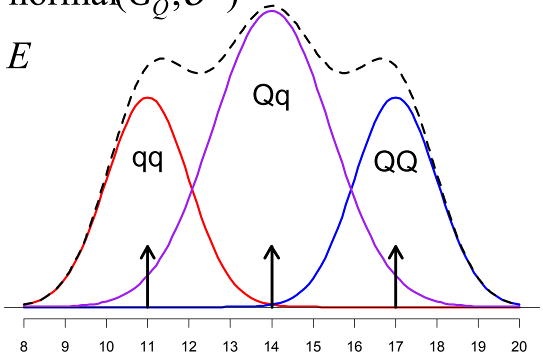
- How to relate genetic linkage to physical distance?
  - math assumptions = crude approximations
  - critical for map building, minor effect on QTL
- x = genetic map distance (1Morgan = 100cM)
  - expected number of crossovers per meiosis between two loci on a single chromatid strand (Sturtevant 1913)
- typical map functions
  - Morgan: interference $r_{AB} = r_{AQ} + r_{QB}$
  - Kosambi: intermediate $r_{AB} = (r_{AQ} + r_{QB})/(1+4r_{AQ}\, r_{QB})$
  - Haldane: no interference $r_{AB} = r_{AQ} + r_{QB} - 2r_{AQ}\, r_{QB}$

2.3 modelling the phenotype

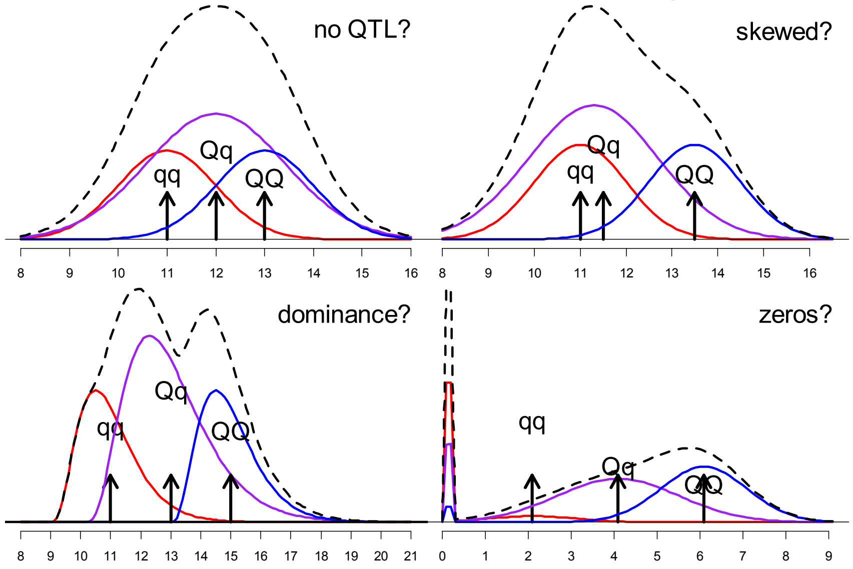
- trait = mean + genetic + environment
- pr(trait Y | genotype Q, effects θ)

$$\text{pr}(Y \mid Q,\theta) = \text{normal}(G_Q,\sigma^2)$$

$$Y = \mu + G_Q + E$$



caution: don’t trust raw histograms!



2.3.1 phenotype model

- how is phenotype related to genotype?
- typical assumptions
  - normal environmental variation
    - residuals *e* (not *Y*!) have bell-shaped histogram
  - genetic value G_Q is composite of a few QTL
    - other polygenic effects same across all individuals
  - genetic effect uncorrelated with environment

$$Y = \mu + G_Q + e, e \sim N(0, \sigma^2)$$

$$E(Y \mid Q, \theta) = \mu + G_Q, \mathrm{var}(Y \mid Q, \theta) = \sigma^2$$

$$\theta = (\mu, G_Q, \sigma^2) \text{ effects}$$

F2 intercross phenotype model

- here assume only one QTL
- genotypes QQ, Qq, qq
- genotypic values G_{QQ}, G_{Qq}, G_{qq}
- decompose as additive, dominance effects

genotype: $Q =$	QQ	Qq	qq
Mather-Jinx: $G_Q =$	$\mu+\alpha$	$\mu+\delta$	$\mu-\alpha$
Fisher-Cockerham: $G_Q =$	$\mu+\alpha-\frac{\delta}{2}$	$\mu+\frac{\delta}{2}$	$\mu-\alpha-\frac{\delta}{2}$

2.3.2 model likelihood

- why study the likelihood?
  - uncover hidden aspects of QTL
  - loci λ, effects θ, given data (Y,X)
- what is evidence to support a QTL?
- where are the QTL?
- how precise can estimate the loci & effects?
- what genetic architecture is supported?

building the model likelihood

- likelihood links phenotype & recombination
  - – through unknown QTL genotypes *Q*
  - – mixture over all possible genotypes
- contribution from individual *i*

$$\text{pr}(Y_i \mid X_i, \theta, \lambda) = \text{sum}_Q \ \text{pr}(Y_i \mid Q, \theta)\text{pr}(Q \mid X_i, \lambda)$$

- product over all individuals

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

$$L(\theta, \lambda \mid Y, X) = \text{prod}_i \ \text{pr}(Y_i \mid X_i, \theta, \lambda)$$

and if there are no QTL?

- $Y = \mu + e$, or $L(\mu \mid Y)$
- no relationship with markers & map *X*
- for normal data, maximum likelihood yields

$$L(\mu \mid Y) = N(Y \mid \mu, \sigma^2)$$

$$\hat{\mu} = \overline{Y}_{\bullet} = \text{sum}_i Y_i / n$$

$$\hat{\sigma}^2 = s^2 = \text{sum}_i (Y_i - \overline{Y})^2 / n$$

maximum likelihood & LOD

- likelihood peaks at some (θ,λ)
  - use “hat” (^) to signify value at maximum
- LOD profiles likelihood peak
  - find θ to maximize likelihood for each λ
  - profile (scan) loci λ over entire genome

$$L(\theta,\lambda \mid Y,X) = \mathrm{pr}(Y \mid X,\theta,\lambda) = \mathrm{prod}_i\ \mathrm{pr}(Y_i \mid X_i,\theta,\lambda)$$

$$LOD(\lambda \mid Y,X) = \log_{10}\left(\frac{\max_\theta L(\theta,\lambda \mid Y,X)}{\max_\mu L(\mu \mid Y)}\right)$$

2.3.3 Bayesian posterior

- treat unknowns as random
  - build “uncertainty” into model framework
  - genetic architecture: gene action θ, QTL locus λ
- interpret weighted likelihood as a density
  - weights based on prior “beliefs”

$$\mathrm{pr}(\theta,\lambda \mid Y,X) = \frac{\mathrm{pr}(Y \mid X,\theta,\lambda)\mathrm{pr}(\theta,\lambda \mid X)}{\mathrm{pr}(Y \mid X)}$$

$$\mathrm{pr}(\theta,\lambda \mid X) = \mathrm{pr}(\theta)\mathrm{pr}(\lambda \mid X)$$

choice of Bayesian priors

- elicited priors
  - higher weight for more probable parameter values
    - based on prior empirical knowledge
  - use previous study to inform current study
    - weather prediction, previous QTL studies on related organisms
- conjugate priors
  - convenient mathematical form
  - essential before computers, helpful now to simply computation
  - large variances on priors reduces their influence on posterior
- non-informative priors
  - may have “no” information on unknown parameters
  - prior with all parameter values equally likely
    - may not sum to 1 (improper), which can complicate use
- always check sensitivity of posterior to choice of prior

incorporate missing genotypes Q

- augment data with missing genotypes Q
  - use recombination model to state uncertainty
  - avoid likelihood mixture by augmentation
- marginal posterior on unknowns of interest
  - average over fully augmented posterior

$$\mathrm{pr}(\theta,\lambda,Q \mid Y,X) = \frac{\mathrm{pr}(Y \mid Q,\theta)\mathrm{pr}(Q \mid X,\lambda)\mathrm{pr}(\theta,\lambda \mid X)}{\mathrm{pr}(Y \mid X)}$$

$$\mathrm{pr}(\theta,\lambda \mid Y,X) = \mathrm{sum}_Q\ \mathrm{pr}(\theta,\lambda,Q \mid Y,X)$$

Bayesian parameter estimates

- estimates are posterior means or modes
  - mean = weighted average of all possible values
  - mode = maximum
- can get joint or marginal estimates

$$\hat{\theta}_{\text{mean}} = \text{sum}_{\theta,\lambda}\, \theta\, \text{pr}(\theta,\lambda \mid Y, X)$$

$$\hat{\theta}_{\text{mode}} = \text{argmax}_{\theta} \left(\text{sum}_{\lambda}\, \text{pr}(\theta,\lambda \mid Y, X)\right)$$

2.4 missing data concepts

- missing QTL genotype *Q*--see section 2.3
- missing marker data *X*
  - errors in genotyping
  - difficulty reading signal (on gel)
  - marker not fully informative
- distinguish full data *X* from observed X_{obs}
  - weighted average over all possible marker values

$$\text{pr}(Q \mid X_{\text{obs}}, \lambda) = \text{sum}_X \text{pr}(Q \mid X, \lambda)\text{pr}(X \mid X_{\text{obs}})$$