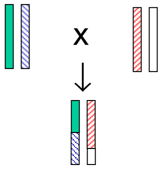
8. QTL for Multiple Crosses

- QTL for multiple crosses
  - four-way cross
  - BC1, BC2, F2 with same inbred parents
  - general crosses of inbred parents
- QTL for outbred pedigrees
  - mixed (effects) model for genotypic effect
  - linkage disequilibrium & inheritance vectors
  - mapping issues for pedigrees

4-way cross: outbred parents

- form “F1” from 2 outbred parents
- up to 4 possible alleles per locus
  - fully informative, heterozygous for one or both parents
- phase (coupling, repulsion) uncertain
  - resolve via parents and ancestors? (pedigree)
  - resolve via linkage (linkage map)

infer parental haplotypes from grandparents and/or offspring



2 outbred parents

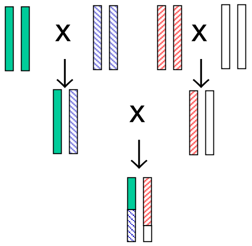
F1 offspring

likelihood-based outmapping

- Butcher et al. (2000)
  - OutMap software based on Ling (1999) thesis
  - R/qtl software incorporates these features
- variant of Lander-Green (1997)
  - ML for recombination rates along linkage group
  - extended from inbred lines to outbred (Ling 1999)
  - hidden Markov models
- caution on using only pair-wise linkage
  - JoinMap (Stam 1993) for arbitrary crosses
    - only need pairwise recombination rates
  - not optimal—not maximum likelihood
  - subtle marker order issues difficult to resolve

4-way cross: 4 inbred parents

- Xu (1996)
- cross in pairs to form 2 distinct F1s
  - cross F1s to get offspring
- phase known from grandparents
  - haplotypes of F1 parents derived from inbreds



4 inbred lines

2 F1 heterozygotes

F2 offspring

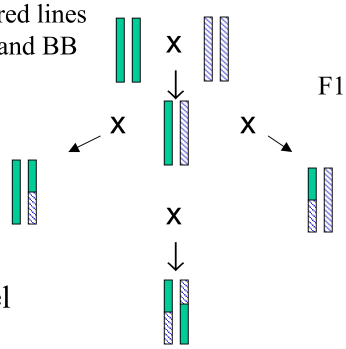
QTL for multiple crosses

- separate analysis by cross
  - simple but inefficient (less power)
- multiple crosses with different parents
  - more power
    - more individuals, more informative markers
  - effect of QTL in different backgrounds
    - genotype * cross, epistatic interactions
- combined analysis over crosses
  - allegedly identical parent stock?
    - crosses created or evaluated at different times
  - relate multiple projects in team

multiple related crosses

- *L* inbred lines (Liu Zeng 2000)
  - F2, BC1, BC2 based on 2 inbreds
  - Xu's (1996) 4-way cross
  - diallele cross: all possible crosses of *L* parents
    - full-diallele: each parent as both male & female
- advantages
  - unravel epistasis
  - increase efficiency of QTL study
    - more alleles = more informative loci
    - increase sample size across multiple crosses (BC1, BC2, F2)
- disadvantage: more complicated, fewer packages
  - related crosses are correlated…

combining BC1, BC2, F2



2 inbred lines
AA and BB

F1 heterozygote
AB

BC1 offspring
AA, AB

BC2 offspring
AB, BB

F2 offspring
AA, AB, BA, BB

recombination model
differs by cross

phenotype model
common overall

how to combine crosses?

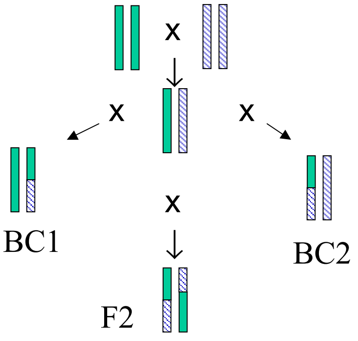
- founders unrelated between crosses
  - naïve sum of separate LODs by cross
    - different gene action in different crosses
  - combined analysis of independent crosses
    - common gene action: one phenotype model pr($Y \mid Q, \theta$)
- genetic relationships within & between crosses
  - constant genetic covariance within cross
    - all individuals have same genetic relationship
    - no effect on single cross analysis (compound symmetry)
  - genetic covariance differs between crosses
    - depends on expected number of alleles shared IBD
  - covariance across multiple crosses is NOT constant
  - “polygenes" usually assumed “independent” of QTL

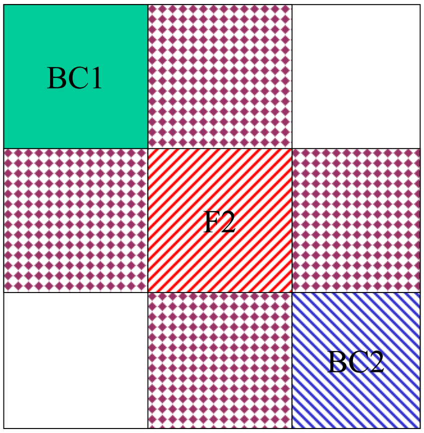
simple fix for multiple crosses

- introduce blocking factor for crosses
  - addresses constant covariance within each cross and different covariances between crosses
  - block is random effect for genetic relationship
- appropriate recombination model for cross
  - relation of recombination rate to distance
- common phenotype model across all crosses
  - could allow cross x genetic effect interactions

genetic covariance for BC, F2

$\text{cov}(Y_1, Y_2) =$





review of quantitative genetics

- genotypic effect is sum of many small effects
  - independent "polygenes" spread over genome
  - no effects localized to any region
- partition of variance of phenotype
  - sum over all polygenic effects
  - partition into additive, dominance, epistatic
  - analyze variance components, not effects

$$\begin{aligned}\text{var}(Y) &= \text{sum}_j \left(\sigma^2_{Aj} + \sigma^2_{Dj}\right) + \text{sum}_{jk}\ \sigma^2_{Ijk} \\ &= \sigma^2_A + \sigma^2_D + \sigma^2_I\end{aligned}$$

relating fixed to random effects

- consider one locus, 2 alleles
- p_{Q} = frequency of Q allele, $p_{\text{q}} = 1 - p_{\text{Q}}$
- a = additive effect per copy of Q allele
- d = dominance effect of Q over q allele

$$\begin{aligned}\sigma^2_A &= 2p_{\text{Q}}p_{\text{q}}\left[a + (1-2p_{\text{Q}})d\right]^2 \\ &= \frac{a^2}{2}, \frac{3(a-\frac{d}{2})^2}{8}, \frac{3(a+\frac{d}{2})^2}{8} \text{ for F2, BC1, BC2}\end{aligned}$$

$$\sigma^2_D = \left[2p_{\text{Q}}p_{\text{q}}d\right]^2 = \frac{d^2}{4}, \frac{9d^2}{64}, \frac{9d^2}{64} \text{ for F2, BC1, BC2}$$

identity by descent (IBD)

- individuals are genetically related
  - measured as correlation or covariance
  - depends directly on degree of genetic relatedness
- IBD allele sharing is key to relatedness
  - IBD = identity by descent (common ancestor)
  - IBS = identity by state (same allele, different sources)
  - IBD = IBS for many inbred crosses (distinct founders)
- variance component or mixed model analysis
  - allow for correlation in mixed model
  - estimate variance components, not effects
    - how variable is additive component?

IBD and QTL covariance

- consider a particular locus (not necessarily QTL) and two individuals Y_1, Y_2, related in some fashion
- k_j = pr(Y_1, Y_2 share j alleles IBD), j = 0,1,2
- $\pi = k_2 + k_1/2$ = coefficient of relationship
  = pr(random allele is IBD at locus)
- genetic covariance from m QTL
  - additive depends on coefficient of relationship
  - dominance depends on both alleles

$$\text{cov}(Y_1, Y_2) = \text{sum}_{j=1}^{m} \pi_j \sigma_{Aj}^2 + k_{2j} \sigma_{Dj}^2$$

IBD and polygenic covariance

- polygenic covariance depends on expection
  - average over all polygenic loci in genome
  - polygenic genotype typically unknown
  - (what if you have complete genomic sequence by individual? how could you improve this?)
- $E(\pi)$ = expected coefficient of relationship
- $E(k_2)$ = expected coefficient of double coancestry

$$\text{cov}(Y_1, Y_2) = E(\pi)\sigma_A^2 + E[k_2]\sigma_D^2$$

IBD and polygenic covariance

- $E(\pi)$ = expected coefficient of relationship
  - 0.5 for F1, 0.75 for BC, 0.625 for F2
  - 0.625 for F2 & BC, 0.5 for BC1 & BC2
- $E(k_2)$ = expected coefficient of double coancestry
  - 1 for F1, 0.5 in BC, 0.375 for F2
  - 0.375 for F2 & BC, 0.25 for BC1 & BC2

$$\begin{aligned}\text{cov}(Y_1, Y_2) &= E(\pi)\sigma_A^2 + E[k_2]\sigma_D^2 \\ &= \tfrac{3}{4}\sigma_A^2 + \tfrac{1}{2}\sigma_D^2 \text{ for BC} \\ &= \tfrac{5}{8}\sigma_A^2 + \tfrac{3}{8}\sigma_D^2 \text{ for F2}\end{aligned}$$

combining QTL and polygenes

- assume QTL and polygenes are independent
- combine in variance component model
- likelihood-based analysis
  - can extend to Bayesian analysis with priors
- null: no QTL effect (QTL variances = 0)

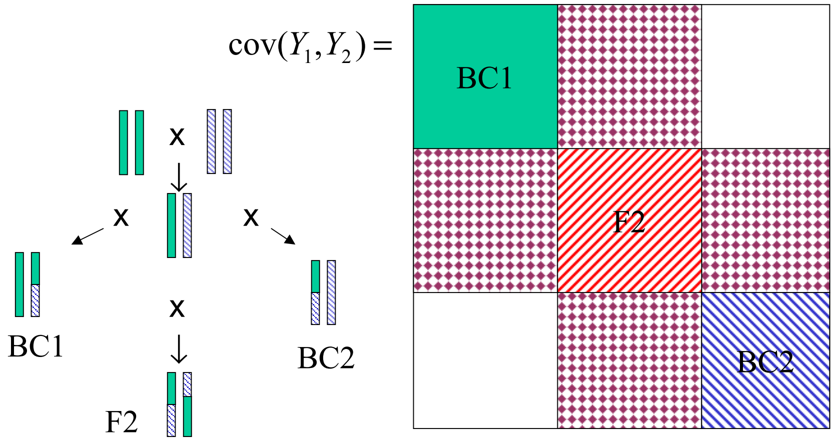
$$\mathrm{cov}(Y_1, Y_2) = \mathrm{sum}_{j=1}^{m} \left[\pi_j \sigma_{Aj}^2 + k_{2j} \sigma_{Dj}^2\right] + E(\pi)\sigma_A^2 + E[k_2]\sigma_D^2$$

$$V = \mathrm{cov}(Y), |V| = \det(V)$$

$$LOD(\theta \mid Y) = c\,|V| + \log_{10}\left((Y-\mu)^{\mathrm{T}} V^{-1} (Y-\mu)\right) - \log_{10}(\mathrm{null})$$

genetic covariance for BC, F2

$$\mathrm{cov}(Y_1, Y_2) =$$



EM approach for multiple crosses

- keep track of parental haplotypes with *L* inbreds
  - follow each allelic contribution separately
  - mostly known phase with inbred founders
    - recall unknown phase in F2: AB/ab vs. Ab/aB
- use in EM or other estimation procedure
  - E step: estimate posterior genotypes pr($Q \mid Y_i, X_i, \theta, \lambda$)
    - relation of recombination to distance
    - depends on type of cross for each individual
  - M steps: maximize likelihood to update effects θ
    - additive, dominance, variance in phenotype model pr($Y \mid Q, \theta$)
    - phenotypic covariance within and between crosses
- LOD (or *LR*) for your favorite hypothesis test

issues in combining crosses

- ignoring polygenic effects can bias results
  - additive effect biases
  - detect dominance when none exists
  - variance increased: less efficient, less power
  - location estimate OK
- increase power by combining crosses
  - important when several related crosses created
  - best power found with F2 alone
- threshold idea for testing and loci intervals
  - extends naturally to multiple crosses (Zou Fine Yandell 2001)
  - permutation based tests possible …

general pedigrees

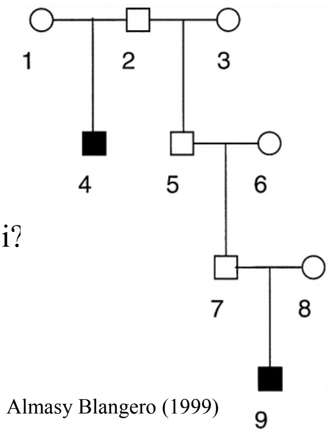
- combine QTL and polygenic effects
  - mixed model (variance components) approach
  - complicated covariance matrix (see above)
- many possible alleles
  - shift from fixed to random effects θ
  - keep track of parental haplotypes (inheritance vectors)
- ambiguities in haplotypes
  - alleles IBD or IBS? sort out using pedigrees & marker linkage
  - many missing values, loops in pedigrees
- calculations can be very complicated
  - software more complicated (SOLAR; Almasy Blangero 199)
  - less progress on QTL analysis than with inbreds
    - Haley-Knott regression common
    - single vs. multiple QTL implementation (Yi Xu 2000)

diversity of pedigree studies

- one or a few large pedigrees
  - common in animal science (cow, pig)
    - 1000 to 100,000 in a single pedigree
    - markers for founders often known
  - similar methods to those described already
- many small pedigrees
  - common in human studies
    - multi-generational; many founders may have died
    - missing marker and phenotype data through pedigree
  - insufficient power to examine only 1 pedigree
  - exceptions: large pedigree studies
    - Iceland, Hutterites, Finland

half-grand avuncular pairs

- founders: 1,2,3,6,8
  - assumed unrelated
- 4&9 may share 0,1 alleles IBD
  - E(π) = 1/16
  - pr(share 1 allele) = 1/8
- what is prob for pair of linked loci?
  - relate to recombination rate r
  - $p_{11} = (1 - r)^2 [r^2 + (1 - r)^2] / 8$



Almasy Blangero (1999)

sorting out missing data

- missing marker *j* for individual *i*?
  - chromosome peeling: use flanking markers
    - almost same idea as for inbreds
    - but relation of probability to *r* depends on pedigree
    - meiosis sampler (Thompson Heath)
  - pedigree peeling: use parents & offspring
    - predict from known marker *j* of parents & offspring
    - single-locus peeling sampler (Thompson Heath)
    - descent graph sampling of alleles (Thompson 1994)
- problem: many missing data!
  - solution: use MCMC to repeatedly fill in gaps

genotype (probability) peeling

- find nuclear families
  - depend on 2 individuals
- find peeling sequence
  - follow nuclear families
  - simplify chain rule

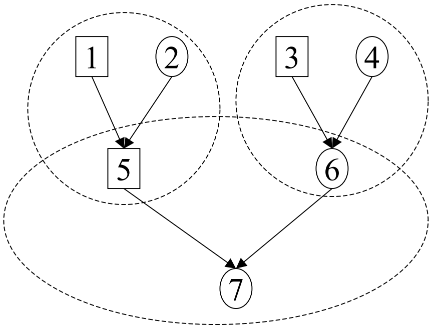
    pr(A,B,C) = pr(A)pr(B|A)pr(C|A,B)
  - use Bayes rule

    pr(A|B) = c x pr(A)pr(B|A)

    $\mathrm{pr}(Q_4|Q_3,Q_6) = c \times \mathrm{pr}(Q_4)\mathrm{pr}(Q_6|Q_3, Q_4)$
- use phenotype to improve
  - posterior for genotype

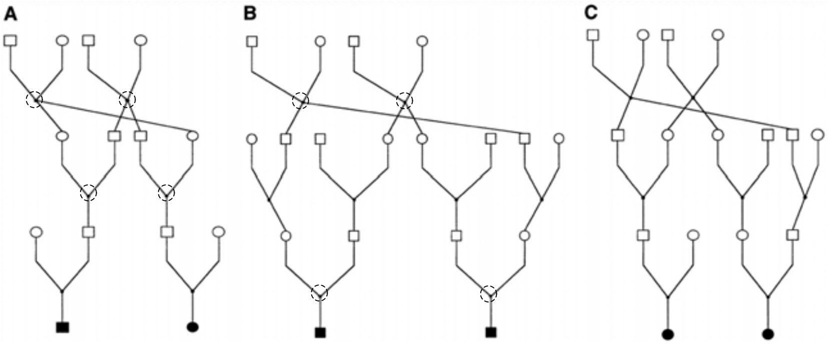
    $\mathrm{pr}(Q_4|Q_3,Q_6\,,Y_4) =$

    $c \times \mathrm{pr}(Q_4) f(Y_4|Q_4)\ \mathrm{pr}(Q_6|Q_3, Q_4)$

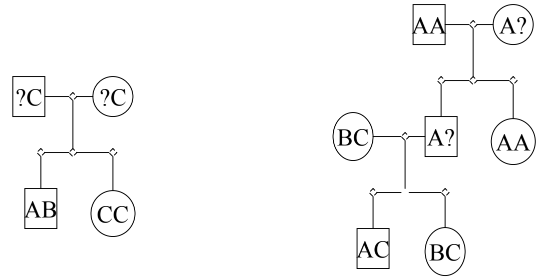


double-second cousins loops in pedigrees!

(Almasy Blangero 1999)

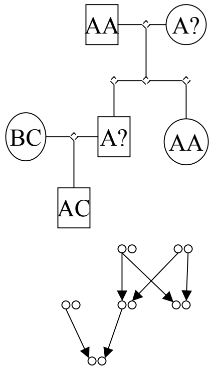


ambiguities in genotype phase
(Hoeschele 2001)



decent graph sampling
(Thompson 1994)

- follow alleles
  - decent through pedigree
  - which grandparent?
- decent graph synonyms
  - segregation patterns
  - meiosis indicators
  - inheritance vectors
- several allele descent graphs may be possible for genetic descent states



fine mapping sketch of idea

- identify small genomic region with QTL
  - ideally less than 1cM or 1M base pairs
- develop advanced intercross lines
  - follow segregation of phenotype & genotype
  - reduce to 100K base pairs via congenics
- identify genes (& pseudo-genes) in region
  - hunt literature, genbank, ncgr, …
- sequence for polymorphisms
  - exons, introns, promoter region,…
  - comparative genomics
- create transgenics to prove function

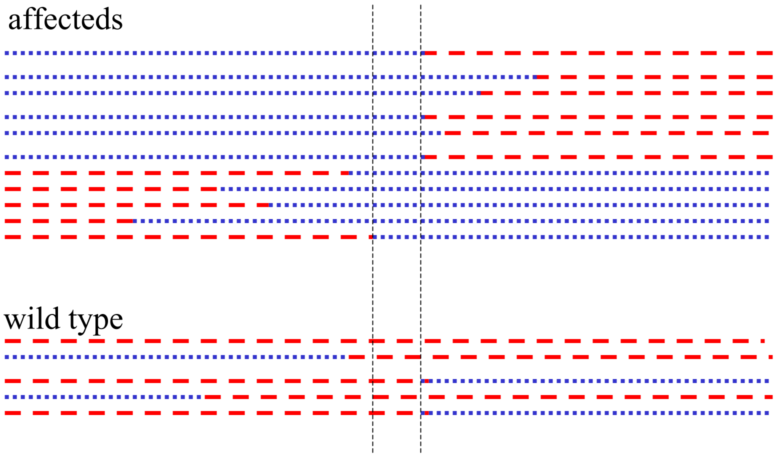
Fine Mapping & Linkage Disequilibrium

- fine mapping with current recombinations
  - QTL localized to 5-20cM: few recombinations nearby
  - additional markers to refine subinterval (Hoeschele 2001)
    - haplotype groups based on recombinant events
    - need highly heritable trait
- fine mapping with historic recombinations
  - linkage (gametic phase) disequilibrium
  - used extensively for qualitative traits
  - influenced by selection, mutation, migration,…
  - assume allele introduced once (e.g. by mutation)

linkage disequilibrium

- phenotypes & markers for current generation(s)
- no pedigree information back to founders
- phenotype model implementation
  - single markers regression (until recently)
  - multiple linked markers (1995-2000)
  - multiple QTL (Wu Zeng 2001; Wu Ma Casella 2002)
- population history
  - allow some haplotypes to be more recently related
  - assume rapid population growth, young & rare disease

basic idea of linkage disequilibrium



natural population

- historic recombination predominates
  - distant relationships between most individuals
  - assume panmixis: random mating in population
- Hardy-Weinberg equilibrium
  - genotype frequency = product of gamete frequency
  - disequilibrium: selection (e.g. affecteds)
- linkage equilibrium
  - genotype frequencies uncorrelated
    - frequency for pair of markers = product of separate frequencies
  - except at very close range or due to selection
- linkage disequilibrium
  - some correlation, usually quite local

why linkage disequilibrium?

- selection, mutation, drift, admixture
- co-segregation over multiple generations
  - physical proximity (linkage)
  - epistatic interactions (selection)
  - recent occurrence (mutation, migration)
- linkage disequilibrium decays with time
  - no LD beyond 5-10cM except due to epistasis
  - ideal for fine mapping
  - models of evolution

mechanism of LD?

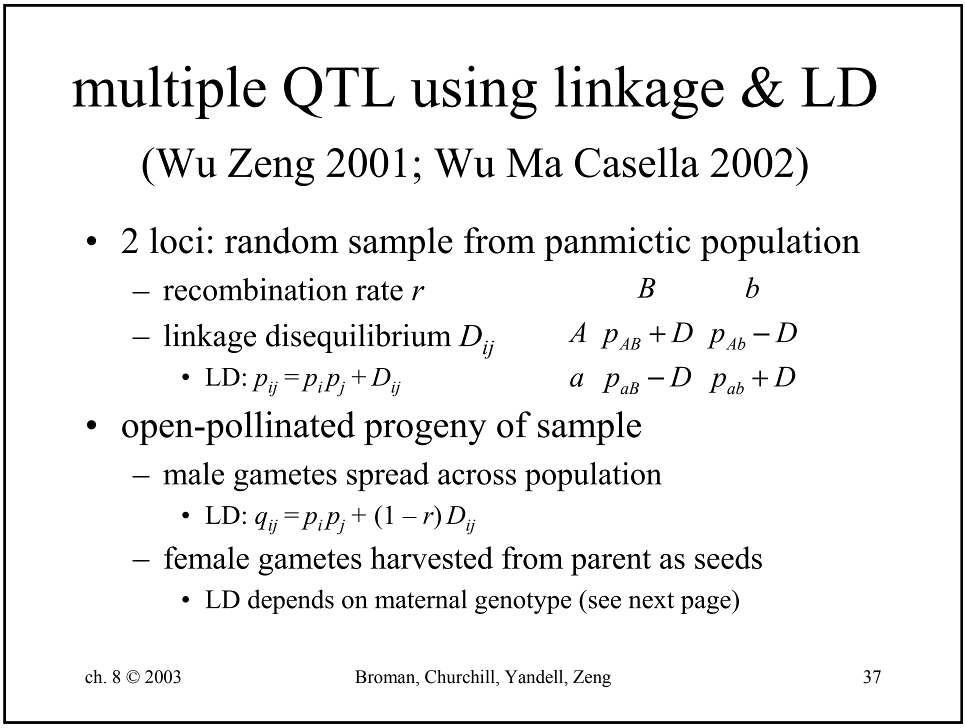
- nuclear families
  - – (e.g. humans, domestic animals)
  - – transmission/disequilibrium test (TDT)
- natural populations
  - – TDT cannot be applied
  - – dioecious vs. monoecious species
    - dioecious: animals, outbred plants
    - monoecious: inbred plants that self

transmission disequilibrium test (TDT) (Spielman et al. 1993)

- consider offspring with disease (qualitative)
- what allele did a parent transmit?
- *M,m* = alleles at a marker locus
- *a,b,c,d* = counts of families
- $E(b) = E(c)$ if no linkage
  - – $E\,(b - c) = (1 - 2r)$A
  - – r = recombination with disease locus
  - – A = constant depending on penetrance and haplotype frequencies

		not transmitted	
		M	m
transmitted	M	a	b
	m	c	d

- likelihood-based test (beyond our scope)



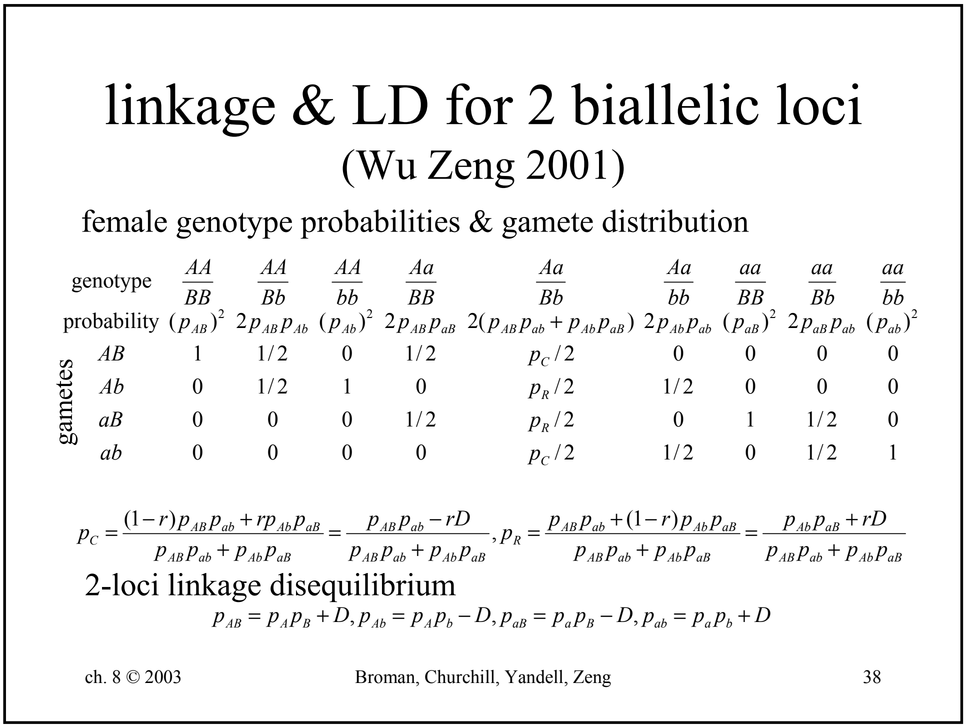
multiple QTL using linkage & LD

(Wu Zeng 2001; Wu Ma Casella 2002)

- 2 loci: random sample from panmictic population
  - recombination rate r
  - linkage disequilibrium D_{ij}
    - LD: $p_{ij} = p_i p_j + D_{ij}$

$$\begin{array}{ccc} & B & b \\ A & p_{AB}+D & p_{Ab}-D \\ a & p_{aB}-D & p_{ab}+D \end{array}$$

- open-pollinated progeny of sample
  - male gametes spread across population
    - LD: $q_{ij} = p_i p_j + (1-r) D_{ij}$
  - female gametes harvested from parent as seeds
    - LD depends on maternal genotype (see next page)



linkage & LD for 2 biallelic loci

(Wu Zeng 2001)

female genotype probabilities & gamete distribution

	genotype	$\frac{AA}{BB}$	$\frac{AA}{Bb}$	$\frac{AA}{bb}$	$\frac{Aa}{BB}$	$\frac{Aa}{Bb}$	$\frac{Aa}{bb}$	$\frac{aa}{BB}$	$\frac{aa}{Bb}$	$\frac{aa}{bb}$
	probability	$(p_{AB})^2$	$2p_{AB}p_{Ab}$	$(p_{Ab})^2$	$2p_{AB}p_{aB}$	$2(p_{AB}p_{ab}+p_{Ab}p_{aB})$	$2p_{Ab}p_{ab}$	$(p_{aB})^2$	$2p_{aB}p_{ab}$	$(p_{ab})^2$
gametes	AB	1	1/2	0	1/2	$p_C/2$	0	0	0	0
	Ab	0	1/2	1	0	$p_R/2$	1/2	0	0	0
	aB	0	0	0	1/2	$p_R/2$	0	1	1/2	0
	ab	0	0	0	0	$p_C/2$	1/2	0	1/2	1

$$p_C = \frac{(1-r)p_{AB}p_{ab} + r p_{Ab}p_{aB}}{p_{AB}p_{ab} + p_{Ab}p_{aB}} = \frac{p_{AB}p_{ab} - rD}{p_{AB}p_{ab} + p_{Ab}p_{aB}}, p_R = \frac{p_{AB}p_{ab} + (1-r)p_{Ab}p_{aB}}{p_{AB}p_{ab} + p_{Ab}p_{aB}} = \frac{p_{Ab}p_{aB} + rD}{p_{AB}p_{ab} + p_{Ab}p_{aB}}$$

2-loci linkage disequilibrium

$$p_{AB} = p_A p_B + D, p_{Ab} = p_A p_b - D, p_{aB} = p_a p_B - D, p_{ab} = p_a p_b + D$$

on to QTL with linkage & LD

- Wu Zeng (2001)
  - – extend from 2 loci to 3 to marker map
  - – consider marker order
- Wu Ma Casella (2002)
  - – use recombination model above
    - restrict to biallelic codominant loci
    - extendible to mutiallelic, missing data
  - – single QTL phenotype model
  - – simulation example

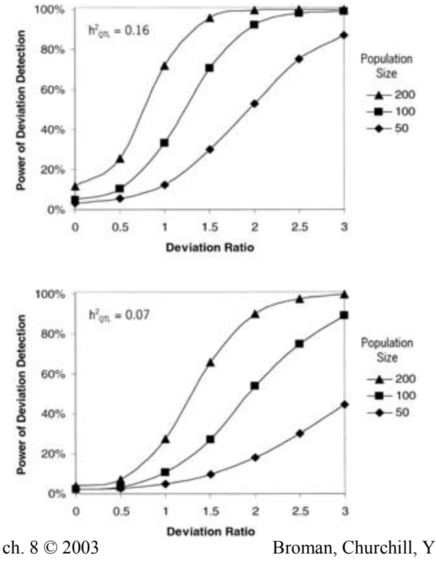
linkage & LD in general pedigree (Hoeschele 2001)

- ideas gleaned from several paper
- quantitative vs. qualitative trait
  - – location *r* and effect size *a* are confounded
  - – recall single marker regression: $(1 - 2r)\,a$
    - small close QTL ≈ large far QTL
  - – need multilocus approach (multipoint mapping)
- likelihood and/or Bayesian approach
  - – combine linkage & LD: ideas in infancy
  - – Yi Xu (2000); Sillanpaa et al. (2003)

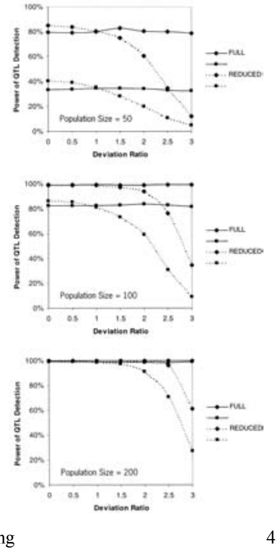
diallele cross (Jannick Jansen 2001)

- does QTL effect depend on genetic background?
  - epistatic interaction with other QTL
  - common environment eliminates QTL x environment
- diallele cross with *s* inbred parents
  - A,B,C inbred parents (actually DH lines)
    - F1s from AxB, AxC, BxC
    - DH progeny from F1s
  - CIM (=MQM) model
    - cofactor (other QTL) effects differ by cross
- test if QTL effect same or different by cross
  - scan genome to identify QTL with epistatic effects
  - follow up with 2-QTL analysis (2-step testing)

power to detect QTL deviation



Jannick Jansen (2001) Fig. 2 & 3



mixed model idea for outbreds

- model components
  - phenotype = design + QTLs + polygenes + env
  - $Y = \mu + G_Q + g + e$
  - $Y_i = \mu + G(Q_i) + g_i + e_i,\ i = 1,\ldots,n$
- QTL effects: fixed or random
- random polygenic effects
  - usually assumed normal
  - correlation depends on genetic relationship A

$$g \sim MVN(0, \sigma_P^2 A), \text{ or } \operatorname{cov}(g_1, g_2) = \sigma_P^2 A_{12}$$

design components

- individual reference $\mu_i = X_i\beta$
  - blocking & local environment
  - (fixed) treatments
    - soil amendments, diet, drugs, shade
  - covariates: individual non-genetic effects
    - sex, age, parity, historical factors
    - other phenotypic traits possibly affected by genotype
  - remove design effect & analyze residuals?
- design x genotype interactions
  - separate analysis by factor levels (e.g. sex)
  - joint analysis (next chapter)