Identity by descent in pedigrees and populations; methods for genome-wide linkage and association.

UNE Short Course: Feb 14-18, 2011

Timetable

Monday

9:00-10:30am	1: Introduction and Overview.
11:00am-12:30pm	2: Identity by Descent; relationships and relatedness
1:30-3:00pm	3: Genetic variation and allelic association.
3:30-5:00pm	4: Allelic association and population structure.

Tuesday

9:00-10:30am	5: Genetic associations for a quantitative trait
11:00am-12:30pm	6: Hidden Markov models; HMM
1:30-3:00pm	7: Haplotype blocks and the coalescent.
3:30-5:00pm	8: LD mapping via coalescent ancestry.

Wednesday a.m.

9:00-10:30am	9: The EM algorithm
11:00am-12:30pm	10: MCMC and Bayesian sampling

Wednesday p.m.

1:30-3:00pm	11: Association mapping in structured populations
3:30-5:00pm	12: Association mapping in admixed populations

Thursday

9:00-10:30am	13: Inferring *ibd* segments; two chromosomes.
11:00am-12:30pm	14: BEAGLE: Haplotype and *ibd* imputation.
1:30-3:00pm	15: *ibd* between two individuals.
3:30-5:00pm	16: *ibd* among multiple chromosomes.

Friday

9:00-10:30am	17: Pedigrees in populations.
11:00am-12:30pm	18: Lod scores within and between pedigrees.
1:30-3:00pm	19: Wrap-up and questions.

Bibliography
Software notes and links.

Introduction and Overview

1. TWO IRRITATING QUESTIONS/COMMENTS
2. MARKERS, DATA STRUCTURES, AND TRAITS
3. PROBLEMS OF WGAS
4. NOT ASSOCIATION vs LINKAGE
5. PEDIGREES AND POPULATIONS; *ibd* vs IBS
6. HOW STRONG SHOULD THIS PEDIGREE PRIOR BE??
7. QUANTITATIVE TRAITS
8. WHY POPULATION STRUCTURE?
9. OUTLINE; FIRST HALF
10. OUTLINE; SECOND HALF

1. TWO IRRITATING QUESTIONS/COMMENTS:

- (1) Do you do study linkage or association?
  Answer: YES. We'll come back to that one shortly.

- (2) You use MCMC so you must be a Bayesian.
  Response: There is nothing Bayesian about MCMC.

- MCMC is simply a method to sample from a probability known only up to a normalizing constant: $\Pr(X;\theta) = h(X;\theta)/c(\theta)$, where $c(\theta)$ cannot be computed explicitly.

- MCMC is widely used by Bayesians, to sample from posterior distributions

$$\pi(\theta \mid X) = \Pr(X;\theta)\pi(\theta) / \int_x \Pr(X;\theta)\pi(\theta)d\theta$$

since the denominator often cannot be computed.

- Using Bayes' Theorem: $\Pr(X|Y) = \Pr(Y|X)\Pr(X)/\Pr(Y)$ does NOT make one a Bayesian.

2. MARKERS, DATA STRUCTURES, AND TRAITS:

Date	Marker type	Data structure	Trait type
1970	Blood types	Nuclear families	Mendelian
1980	RFLPs	Large pedigrees	Simple traits
1990	STRs (Microsatellites)	Small pedigrees	Quantitative traits
2000	SNPs	Case/Control ("unrelated")	Complex traits
2010	1M SNPs and Sequence data	Pedigrees in Populations	Complex quantitative traits

3. PROBLEMS OF WGAS:

• Traits are complex; many genes of small effect. Leads to increasingly larger studies; 10,000 cases/controls.

• Traits are heterogeneous.
The problem of rare variants; allelic heterogeneity.
There are many ways to mess up a functional gene.

• Where is the "missing heritability"? Genes interact; epistasis.

• Data quality control (PLINK; (Purcell et al., 2007); GENEVA (Laurie et al., 2010)). Large studies ⇒ multi-centre studies. Genotyping failure is NOT random, and case-control differences in time of sampling, lab handling, ...

• Population structure and history;
Problems increase as study sizes increase.
Cases (ascertained) and controls usually differ in population history.

4. NOT ASSOCIATION vs LINKAGE:

• The objective is to map genes affecting a trait;
find where they are in the genome.

• Linkage results in the cosegregation of DNA at nearby genetic loci.

• Linkage therefore maintains associations in the allelic types of DNA on a chromosome at nearby genetic loci; these associations (linkage disequilibrium; LD) are used in association mapping.

• Linkage results in patterns of cosegregation that can be inferred in pedigrees: this dependence in segregation is used in classical linkage mapping.

• The difference is not in the objective, but in the data structures, models, and data we choose to use.

5. PEDIGREES AND POPULATIONS; *ibd* vs IBS:

• In human populations the (known) pedigrees are small, relative to dairy cattle pedigrees.

• The pedigrees I look at are large human pedigrees, but still small, relative to dairy cattle pedigrees.

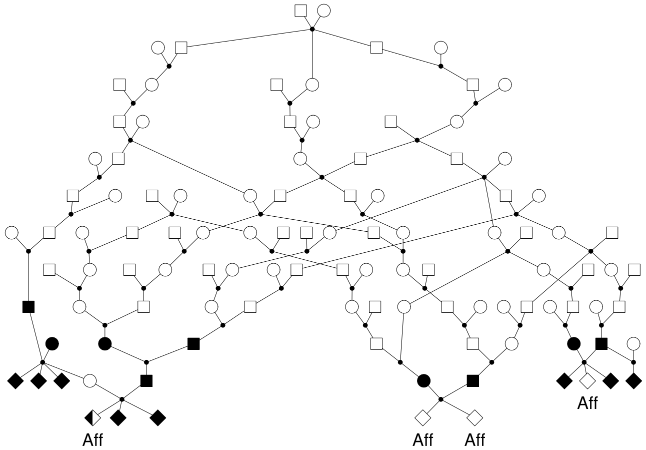
• The number of SNPs for human data is huge (1M SNP chip), vs 50K for dairy cattle – but I work at the 50K level.

• However, LD is likely much higher in cattle populations, due to breeding, history etc.

• I look at *identity by descent* (*ibd*), rather than allelic associations (*identity by state*: IBS); IBS is a reflection of *ibd* whether in pedigrees or populations.

• The only (?) difference between pedigrees and populations is that a pedigree gives a very strong prior on *ibd*.

6. HOW STRONG SHOULD THIS PEDIGREE PRIOR BE??:



• 1990s marker data, in absence of trait model, show no evidence of relatedness among families.

• Assumption of rare recessive trait drives inference of ancestry.

• Details of the ancestral pedigree are surely wrong/biased.

7. QUANTITATIVE TRAITS:

• Quantitative traits are important:
"Traditional" animal breeding quantitative traits.
Quantitative measures of human complex diseases.
Quantitative measures of gene expression (eQTL).

• Exact models specific to the problem:
Herd, year, ... fixed effects;
shared environment variance components
Diet, smoking, ...; shared environment;
missing covariates – less control of environment.
Plate effects, dye effects, batch effects, ...

• Message 1: whatever the quantitative trait model superimposed, it is better done on *ibd* and coancestry, rather than allelic effects.

• Message 2: Markers are useful for what they can tell us about *ibd*.

• Message 3; To find what markers tell us about *ibd* we need to understand the population genetics underlying the relationship.

8. WHY POPULATION STRUCTURE?:

• This is a new area for me to teach in a course; notes not polished! – note re numbering of slides!

• Course has evolved in the writing; become focused towards models for population structure, and inference of population structure and under population structure.

• Population genetics, and ideas of ancestry, gene identity by descent (*ibd*) have much to offer association studies.

• Focus on ideas underlying methods; not implementation.

• Focus on structured populations: all real populations have structure.

• Focus on model-based methods: for quantitative traits we need models, for the phenotypes, so model-based analysis of marker data makes sense.

9. OUTLINE; FIRST HALF:

• Part 1: Basics
2: Identity by Descent; relationships and relatedness
3: Genetic variation and allelic association.
4: Allelic association and population structure.
5: Genetic associations for a quantitative trait

• A bridge to Part 2:
6: Hidden Markov models; HMM

• Part 2: Haplotypes and coalescents
7: Haplotype blocks and the coalescent.
8: LD mapping via coalescent ancestry.
9: The EM algorithm.
10: MCMC and Bayesian sampling

10. OUTLINE; SECOND HALF:

• A bridge to Part 3:
11: Association mapping in structured populations
12: Association mapping in admixed populations

• Part 3: *ibd* inference in populations
13: Inferring *ibd* segments; two chromosomes.
14: BEAGLE: Haplotype and *ibd* imputation.
15: *ibd* between two individuals.
16: *ibd* among multiple chromosomes.

• and finally, *ibd* in pedigrees
17: Pedigrees in populations.
18: Lod scores within and between pedigrees.

Identity by Descent; relationships and relatedness

1. MENDEL's LAWS: THE INHERITANCE OF DNA
2. PEDIGREES IN POPULATIONS: POPULATION PEDIGREES
3. THE INHERITANCE OF CHROMOSOMES
4. *ibd* of CHROMOSOMAL SEGMENTS
5. LENGTHS OF *ibd* SEGMENTS
6. VARIATION IN *ibd* IN OFFSPRING OF FIRST COUSINS
7. RELATIONSHIPS, RELATEDNESS AND *ibd*
8. WHY ESTIMATE RELATIONSHIP/RELATEDNESS?
9. PRZEWALSKI HORSES: MIXED UP RECORDS
10. CALIFORNIA CONDORS: NO RECORDS, LITTLE DATA

1. MENDEL's LAWS; INHERITANCE OF DNA:

• **Mendel's Laws (1866)**: For diploid individuals:
1. At any given locus, each individual has two genes, one maternal and the other paternal. Each individual copies to each offspring a randomly chosen one of its two genes; independently to each offspring, independently of gene segregated by the spouse, independently of gene segregated from parent.
2. Independently for different loci. (Not true; segregation of genes at loci on the same chromosome are dependent)

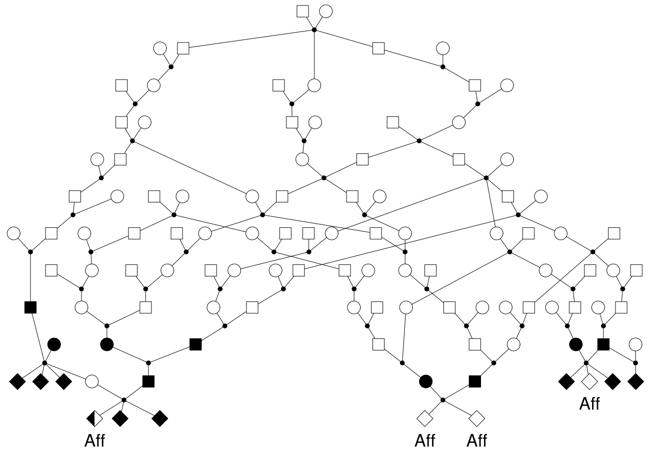
• Hence, individuals carry at a locus pieces of DNA that are copied through repeated segregations from their ancestors. Relatives who share a common ancestor may both carry copies of the same ancestral piece of DNA. Such pieces of DNA are said to be

identical by descent (*ibd*) .

• Known or unknown, members of populations are related!

• *ibd* is relative! – to some time point or founder population.

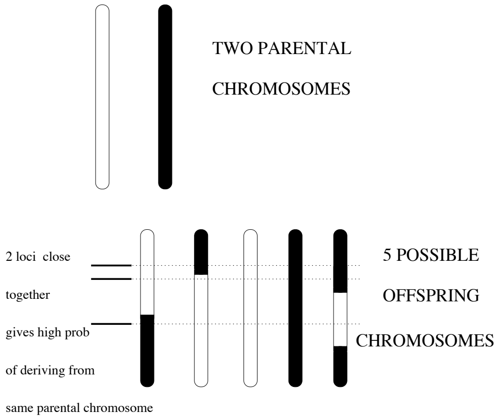
2. PEDIGREES IN POPULATIONS: POPULATION PEDIGREES:



• Details of the ancestral pedigree are surely wrong/biased.
We want to use the *ibd* information, but not the ancestral pedigree.

• 1990s data were insufficient for between-family inference of *ibd*.
With modern data, we could infer *ibd* among families.

3. THE INHERITANCE OF CHROMOSOMES:



• Chromosomes are inherited in large chunks, $\sim\ 10^8$ bp.

• Assume, no genetic interference.

• In any meiosis, crossovers occur as a Poisson process along the chromosome.

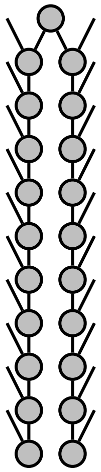
• Between any two positions (loci), in any meiosis, there is recombination if the DNA at those positions derives from different parental chromosomes (i.e. odd number of crossovers).

• Probability of recombination increases with genetic distance.

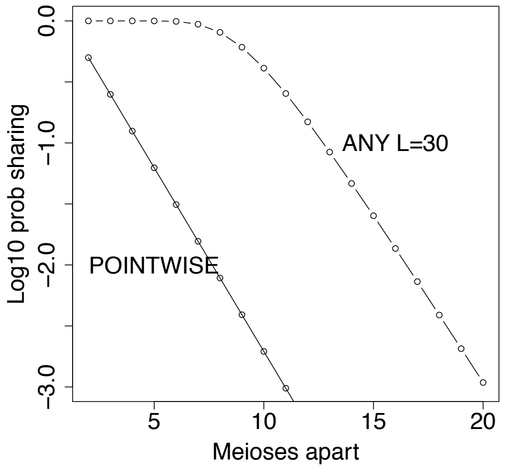
• At large distances, even and odd have equal probability; $r \approx 1/2$.

4. *ibd* PROBABILITIES IN REMOTE RELATIVES:

$\Pr$(2 kids get same) $= 1/2$
$\Pr$(descendants share)
$= 2 \times (1/2)^m$

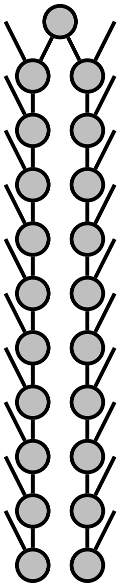


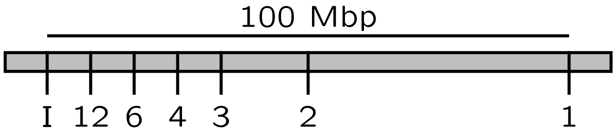
$\Pr$(share any genome length L=30 Morgans)
$= 1 - \exp(-(m-1)L/2^{m-1})$
K. P. Donnelly (1983).



	$m = 12$	$m = 20$
at point	0.0005	2×10^{-6}
any ($L = 3000$Mbp)	0.148	0.001

5. LENGTHS OF *ibd* SEGMENTS:

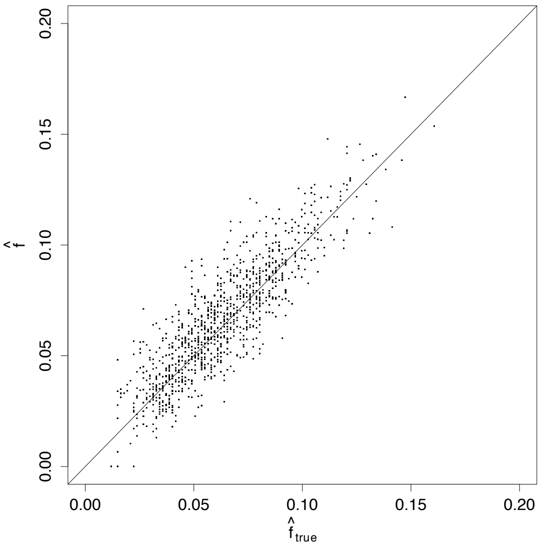




- *ibd* genome segments are few, not short K. P. Donnelly (1983).
- From the start point of the segment:
  In any meiosis, distance to next recomb. is exponential ($\mathcal{E}$) mean 1.
- Over m meioses, where is closest recomb.?
  Meioses are independent (Mendel's 1 st law)
  For m meioses, m times rate in Poisson process.
  Min of m indep. $\mathcal{E}$s mean 1 is $\mathcal{E}$ mean $1/m$.
- For $m = 20$, expected length is 1/20 Morgans or 5 Mbp.
- Human (and bovine) genomes are short!!

(The variance in proportion genome shared is high)

6. VARIATION IN *ibd* IN OFFSPRING OF FIRST COUSINS:



Leutenegger et al. (2003) simulated offspring of 1000 indep. first cousin pairs:

Estimation of f using 5cM
microsatellite map
(630 markers)

$$\overline{f} = 1/16 = 0.0625$$

At most 50 “indep” *ibd* events.
The human genome is short.

7. RELATIONSHIPS, RELATEDNESS AND *ibd*:

- Pedigree relationships?
  Validating pedigrees: human, Przewalski horses.
  Is this my half-sib? Which horse is this?
  Even a whole genome of data may not answer this,
  without prior information/specific hypotheses.

- General degree of relationship (kinship)?
  California condors, estimating population substructure.
  Are these two birds closely related ? How close?
  Do these two birds share more genome than these two?
  Can be estimated with enough genetic data,
  but it is a genome-wide ”average” answer?

- *ibd* at specific genome locations ?
  Associating genome locations with traits; linkage analysis.
  Do affected relatives tend to be *ibd* at this candidate location?
  With modern genetic data (dense SNP markers)
  we can detect even (relatively) small segments of *ibd*.

8. WHY ESTIMATE RELATIONSHIP/RELATEDNESS?:

- Forensic questions: identifying individuals from their relatives
  victims of natural or man-made disasters
- Legal questions: Identifying parents, children, siblings:
  paternity testing, adoptions, immigration cases.
- Medical Genetics: for example, sib pair studies
  Validation of stated pedigree relationships. Sample swaps.
- Conservation Genetics: establishing breeding strategy for
  severely endangered species: California condor,
  Przewalski horse, Caribbean iguanas
- Ecological Genetics: gene flow, and reproductive success
  dispersal of seed pollen and juveniles
  perennial plants, armadillos, salmon

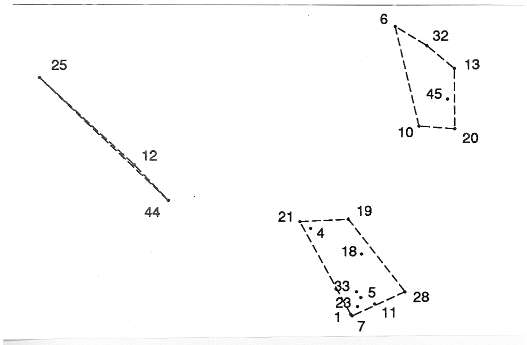
9. PRZEWALSKI HORSES: MIXED UP RECORDS:

Only “true” wild horse:
66 chromosomes (vs 64)
Captive-bred (13 founders)
1927-1997
One was known Mongolian domestic.
Askania Nova; main “pure” group,
and one more recent (1953) founder.

Many uncertainties; horses mixed up. Wrong ones shipped.
– concerns as to validity of International Stud Book.
San Diego “pure” stallion (1985), led to establishing of two groups (“pure”/“mixed”)in USA, but he was not. etc. etc.
1992: genetic marker data used to resolve many pedigree errors.

Now reintroduced in China & Mongolia, but still threatened.

10. CALIFORNIA CONDORS: NO RECORDS, LITTLE DATA:



Genetically; Three groups

1984-5: Population crash; survivors into captivity; also eggs
Condors live long, fly far; how are these related ??
Topa-Topa in LA Zoo 20 years, maybe brother to AC5 –from wild
Who should be bred? Who released? Maintain the gene pool.
Now over 200 total: 100 in SD/LA, 100 fly (semi-)free.

Genetic variation and allelic association

1. GENETIC TERMINOLOGY
2. GENE *ibd* AND ALLELIC TYPES
3. POPULATION KINSHIP and INBREEDING
4. ALLELE and GENOTYPE frequencies
5. POPULATION SUBDIVISION AND STRUCTURE
6. MEASURES OF POPULATION STRUCTURE: Wright (1951)
7. MAINTAINING VARIATION: MUTATION AND SELECTION
8. RANDOM GENETIC DRIFT
9. HOMOZYGOSITY AND POPULATION DIVERGENCE

1. GENETIC TERMINOLOGY:

• Human cell nucleus — has 46 chromosomes (each double-strand DNA): 22 pairs of autosomes, and 2 sex chromosomes, X and/or Y.

• Nuclear genome:
DNA of these (22+X+Y) chromosomes, 3×10^9 bp.

• Locus— position on a chromosome, or DNA at that position, or the piece of DNA coding for a trait.

• Allele— type of the DNA at a particular locus

• SNP — single nucleotide polymorphism; two alleles A and B.

• Genotype— (unordered) pair of alleles at a particular locus
in a particular individual. AA, AB, BB
Homozygote– a genotype with two like alleles. AA, BB
Heterozygote – a genotype with two unlike alleles. AB

• Phenotype— observable characteristics of an individual
For SNP loci we score a genotype, but there may be error.

2. GENE *ibd* AND ALLELIC TYPES:

• What is a gene?? – “a much over-used word” (R. C. Elston)
– the chunk of DNA coding for a functional protein.
– Not a locus. Not an allele.

• A simple model: (in which non-*ibd* implies Hardy Weinberg freq.)
ibd genes are of the same allelic type: ignores mutation etc.
non-*ibd* genes are of independent types: ignores popn structure...

• In a pedigree *ibd* is easily defined, relative to the founders of the pedigree as stated, but pedigrees exist within populations ....

• In a population, there is no absolute definition of *ibd*, but relative to any given time point, some individuals share more *ibd* than do others.

• More closely related individuals are (on average) more similar than are less related individuals, because they have higher probabilities of having *ibd* genes, that are copies of the same gene in a common ancestor.

3. POPULATION KINSHIP and INBREEDING:

- The simplest probabilities of gene *ibd* are between two genes.

- For genes: The population kinship ψ is the probability two homologous genes in chromosomes randomly chosen from the population are *ibd*. (See note below.)

- For individuals, the coefficients of kinship (ψ) and inbreeding (f) are

$$\begin{aligned}\psi(B,C) &= \mathrm{Pr}(\text{homologous genes segregating} \\ &\qquad \text{from } B \text{ and } C \text{ are } \textit{ibd}) \\ f(B) &= \mathrm{Pr}(\text{homologous genes in B are } \textit{ibd}) \\ &= \psi(M_B, F_B)\end{aligned}$$

where M_B and F_B are the parents of B.

For one locus, segregating $\equiv$ randomly chosen.
but for multiple loci "randomly chosen" is not well defined.

4. ALLELE and GENOTYPE frequencies:

- Suppose A has frequency p, B has frequency $q = (1-p)$.

In what population?? Technically, the one relative to which we are measuring *ibd*. In practice: a current population sample. Sensitivity to allele frequencies is an issue.

- Suppose an individual has inbreeding coefficient f:

$$\begin{aligned}\mathrm{Pr}(AA) &= (1-f)p^2 + fp &&= p^2 + fpq \\ \mathrm{Pr}(AB) &= (1-f)2pq &&= 2pq - 2fpq \\ \mathrm{Pr}(BB) &= (1-f)q^2 + fq &&= q^2 + fpq\end{aligned}$$

- If $f = 0$; $\mathrm{Pr}(AA) = p^2$, $\mathrm{Pr}(AB) = 2pq$, $\mathrm{Pr}(BB) = q^2$.

These are Hardy-Weinberg frequencies (HWE).
($f = 0$ means the individual's parents have no common ancestors, relative to ...?)

- One generation of random mating* establishes HWE, since, by definition, the two genes in an individual are copies of independently sampled parental genes. (*: Random union of gametes.)

- If widespread non-HWE, then likely there is structure.

5. POPULATION SUBDIVISION AND STRUCTURE:

• Suppose populations i, each in HWE, with p_i the freq of allele A in population i, and α_i the proportion of population i.
So $\Pr(A) = p = \sum_i \alpha_i p_i$

$$\Pr(AA) - (\Pr(A))^2 = \sum_i \alpha_i p_i^2 - p^2 = \sum_i \alpha_i (p_i - p)^2 \geq 0$$

$$\Pr(AB) - 2\Pr(A)\Pr(B) = 2\left(\sum_i \alpha_i p_i(1-p_i) - p(1-p)\right)$$

$$= -2\left(\sum_i \alpha_i p_i^2 - p^2\right) = -2\sum_i \alpha_i (p_i - p)^2$$

• Thus, population subdivision results in homozygote excess relative to HWE. This excess is known as the Wahlund variance σ_f^2.
In total, we therefore have heterozygote deficiency, but (for multiple alleles) not necessarily for each heterozygote.

6. MEASURES OF POPULATION STRUCTURE: Wright (1951):

• Let $X = 1$ is allele is A, and $X = 0$ otherwise. $\mathbf{E}(X) = p$, $\text{Var}(X) = p(1-p)$.

• For an individual (I) in a (sub)population (S), allele indicators X_1,X_2;
$\Pr(X_1 = X_2 = 1) = \Pr(AA) = p^2 + fp(1-p)$, or

$$\begin{aligned} corr.(X_1, X_2) &= (\mathbf{E}(X_1X_2) - \mathbf{E}(X_1)\mathbf{E}(X_2))/\sqrt{\text{Var}(X_1)\text{Var}(X_2)} \\ &= (\Pr(AA) - p^2)/p(1-p) = f = F_{IS} \end{aligned}$$

F_{IS} measures departure from HWE within (sub)-populations.

• Now consider subpopulations (S) making up total population (T). And correlation of alleles within subpopulations (S) relative to (T).
Now $\mathbf{E}(X_1X_2)$ is $\Pr(AA)$ of previous slide, and $corr.(X_1, X_2)$ is

$$F_{ST} = \sigma_f^2/p(1-p) = (\Pr(AA) - p^2)/p(1-p)$$

where now $\Pr(AA)$ refers to probability both alleles are A.
F_{ST} measures association due to population structure.

7. MAINTAINING VARIATION: MUTATION AND SELECTION:

- Mutation is the only source of new variation at a locus.
  Mutation rates are about 10^{-8} to 10^{-9} per bp per meiosis.
  Mutation rates are about 10^{-5} to 10^{-6} per coding functional gene.
  Mutation rates are hard to measure directly.

- Directional selection removes variation.
  In equilibrium, “loss” = “gain”.
  Hence, indirect estimates of mutation rates.
  Consider normal allele A and rare mutatnt B.

- For example: rare dominant with selection coeficient s:
  we lose a B allele, with prob, s, for each AB individual.
  we gain μ B alleles, in each of $2N$ meioses (approx.)
  So $Ns2q(1-q) = 2N\mu$, or $\mu = 2sq$. ($q = \mu/2s$.)

- For example, recessive with selection coefficient s:
  we lose 2 B alleles, with prob, s, for each BB individual.
  we gain μ B alleles, in each of $2N$ meioses (approx.)
  So $Ns2q^2 = 2N\mu$, or $\mu = sq^2$. ($q = \sqrt{\mu/s}$.)

8. RANDOM GENETIC DRIFT:

- Real populations are finite (and have structure, and history, . . .).

- Let $X(t)$ be number of A alleles at time t in popn size $2N$ genes. Suppose $(X(t)|X(t-1))$ is $Bin(2N, X(t-1)/2N)$ (Wright-Fisher model; random union of gametes). Then

$$\begin{aligned}\mathbf{E}(X(t)) &= \mathbf{E}(\mathbf{E}(X(t)|X(t-1))) = \mathbf{E}\left(2N\frac{X(t-1)}{2N}\right) \\ &= \mathbf{E}(X(t-1)) = \ldots = X(0)\end{aligned}$$

- Ultimately (without mutation) variation is lost:
  $\mathbf{E}(X(\infty)) = X(0)$ so $\Pr(X(\infty) = 2N) = X(0)/2N$.

- *ibd* increases. Consider non-*ibd*, $(1 - f(t))$:

$$(1-f(t)) = (1-(1/2N))(1-f(t-1)) = (1-(1/2N))^t$$

with *ibd* measured relative to time 0.

9. HOMOZYGOSITY AND POPULATION DIVERGENCE:

- Homozygosity increases:

$$\begin{aligned} \text{Note } \mathbf{E}(X^2) &= \text{Var}(X) \;+\; (\mathbf{E}(X))^2 \\ \text{So } \mathbf{E}(X(t)^2) &= \mathbf{E}(\mathbf{E}(X(t)^2|X(t-1))) \\ &= \mathbf{E}(\text{Var}(X(t)|X(t-1))) + \mathbf{E}(X(t-1)^2) \end{aligned}$$

- Homozygosity increases relative to time 0, because the allele frequency has increasing chance of being closer to 0 or 1, but population is still in HWE.

- Populations diverge: Let $V_t = \text{Var}(X(t))$ and $X_1(t)$ and $X_2(t)$ counts in two indep popns with same $X(0)$

$$\begin{aligned} \mathbf{E}((X_1(t) - X_2(t))^2) &= \mathbf{E}(X_1^2) - 2\mathbf{E}(X_1 X_2) + \mathbf{E}(X_2^2) \\ &= (V_t + X(0)^2) - 2X(0)^2 + (V_t + X(0)^2) \\ &= 2V_t \\ &\approx (4Nt)(X(0)/2N)(1 - X(0)/2N) \end{aligned}$$

Allelic association and population structure

1. SEGREGATION OF HAPLOTYPES
2. ALLELIC ASSOCIATION; LINKAGE DISEQULIBRIUM
3. DECAY OF LD
4. LD and POPULATION STRUCTURE
5. THE BASIC ALLELIC ASSOCIATION TEST
6. THE GENOME-WIDE DISTRIBUTION OF p-VALUES
7. GENOMIC CONTROL: Devlin and Roeder (1999)
8. DETECTING POPULATION STRUCTURE
9. WITHIN-LOCUS CONTROL: the TDT (1993)

1. SEGREGATION OF HAPLOTYPES:

- For 2 SNPs, alleles A_j,B_j at locus j there are 4 haplotypes: A_1A_2, A_1B_2, B_1A_2 and B_1B_2 frequencies q_1, q_2, q_3, q_4.

- Only the double heterozygote A_1B_1, A_2B_2 cannot be phased,

- Homozygous individuals (both loci): for example an A_1A_1, B_2B_2 individual segregates only A_1B_2 haplotypes.

- Homozygote/Heterozygote: for example, an A_1A_1, A_2B_2 individual passes on A_1A_2 or A_1B_2 each with probability $1/2$ regardless of recombinaton probability ρ.

- A double-heterozygote individual passes each of the four haplotypes A_1A_2, A_1B_2, B_1A_2 and B_1B_2, with probabilities:
$(1-\rho)/2$, $\rho/2$, $\rho/2$ and $(1-\rho)/2$ if his genotype is A_1A_2/B_1B_2,
$\rho/2$, $(1-\rho)/2$, $(1-\rho)/2$, and $\rho/2$ if his genotype is A_1B_2/B_1A_2.

- Recombination breaks up chromosomes, but we only see this directly if genotypes are heterozygous at all loci.

2. ALLELIC ASSOCIATION; LINKAGE DISEQULIBRIUM:

- A measure of allelic association between the two loci is

$$\begin{aligned} \triangle &= \Pr(A_1A_2) - \Pr(A_1)\Pr(A_2) \\ &= q_1 - (q_1+q_2)(q_1+q_3) \\ &= (q_1q_4 - q_2q_3) \end{aligned}$$

since $q_1+q_2+q_3+q_4 = 1$. This measure is known as the coefficient of *linkage disequilibrium*.

- Allelic associations between loci arise from population structure, admixture and history, or from selection.
Example of mixture/subdivision— the “nuisance” case. Vs. case of interest— original mutation on some genetic background.

- Associations are, however, maintained by tight linkage ($\rho \approx 0$).
LD blocks are the remnants of recombination; they are not caused by linkage, but they survive because of linkage.

- Contrast with HWE: Even for unlinked loci equilibrium ($\triangle = 0$) is not achieved in one generation.

3. DECAY OF LD:

• Suppose current haplotype frequencies are q_1, q_2, q_3 and q_4, and at next generation are q_1^*, q_2^*, q_3^* and q_4^*.

• Now, for example, an offspring A_1A_2 haplotype arises
from a A_1A_2/A_1A_2 parent with prob 1.
from a A_1A_2/A_1B_2 or A_1A_2/B_1A_2, with prob $1/2$,
from a A_1A_2/B_1B_2 with prob $(1-\rho)/2$
from a A_1B_2/B_1A_2 with prob $\rho/2$. Thus

$$\begin{aligned} q_1^* &= q_1^2 + 2q_1(q_2+q_3)/2 \; + \; 2q_1q_4(1-\rho)/2 + 2q_2q_3\rho/2 \\ &= q_1(q_1+q_2+q_3+q_4) \; - \; \rho(q_1q_4 - q_2q_3) \;\; = \;\; q_1 \; - \; \rho\Delta. \end{aligned}$$

Analogously, $q_2^* = q_2+\rho\Delta$, $q_3^* = q_3+\rho\Delta$ and $q_4^* = q_4-\rho\Delta$. Thus, in expectation, allele frequencies are unchanged ($q_1^*+q_2^* = q_1+q_2$):

$$\begin{aligned} \Delta^* &= q_1^*q_4^* - q_2^*q_3^* \\ &= (q_1-\rho\Delta)(q_4-\rho\Delta) \; - \; (q_2+\rho\Delta)(q_3+\rho\Delta) \\ &= \Delta - \rho\Delta(q_1+q_2+q_3+q_4) + \rho^2(\Delta^2 - \Delta^2) \\ &= (1-\rho)\Delta. \end{aligned}$$

4. LD and POPULATION STRUCTURE:

• Population stratification creates LD, even if there is no LD within subpopulations. If allele frequencies differ so will the frequency of haplotypes.

• Consider populations i in proportions α_i and allele A alleles frequencies p_{1i}, p_{2i} at two loci.
As before $\Pr(A_1) = p_1 = \sum_i \alpha_i p_{1i}$, and $\Pr(A_2) = p_2 = \sum_i \alpha_i p_{2i}$.

• Then

$$\begin{aligned} \Delta &= \Pr(A_1A_2) \; - \; \Pr(A_1)\Pr(A_2) \\ &= \sum_i \alpha_i p_{1i}p_{2i} - (\sum_i \alpha_i p_{1i})(\sum_j \alpha_j p_{2j}) \\ &= \sum_i \alpha_i (p_{1i} - p_1)(p_{2i} - p_2). \end{aligned}$$

• An association test is looking for LD between a SNP and a “causal locus”; i.e. association between case-status and the SNP alleles.

5. THE BASIC ALLELIC ASSOCIATION TEST:

• Basically, we are looking for a difference in SNP genotype or allele frequencies between cases and controls.

	AA	AB	BB	total)	A	B	total
# B	0	1	2	(fixed)			
Cases	r_0	r_1	r_2	R	$2r_0 + r_1$	$r_1 + 2r_2$	$2R$
Controls	s_0	s_1	s_2	S	$2s_0 + s_1$	$s_1 + 2s_2$	$2S$
Total	n_0	n_1	n_2	N	$2n_0 + n_1$	$n_1 + 2n_2$	$2N$

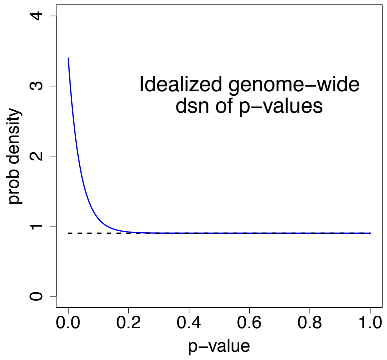
• The χ^2 statistic is

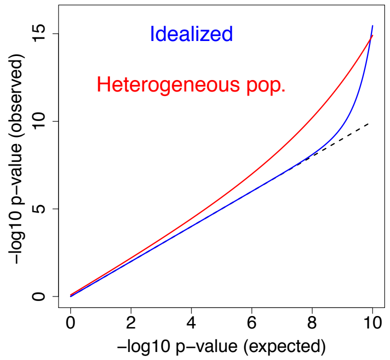
$$Y^2 = \frac{2N(2N(r_1 + 2r_2) - 2R(n_1 + 2n_2))^2}{(2R)(2S)(2N(n_1 + 2n_2) - (n_1 + 2n_2)^2)}$$

• Test using χ^2_1 dsn (or do Fisher Exact test).

• Tests are not independent: the SNPs are tightly linked, and there is LD – that is the whole point.

6. THE GENOME-WIDE DISTRIBUTION OF p-VALUES:





• Under the null hypothes, the p-values should be uniform $U(0, 1)$.

• Ignoring population structure etc. we will see a mixture of $U(0, 1)$ and associations giving rise to small p-values.

• This is basis of FDR appoaches for microarrays (Storey, 2002). Same problem of very large numbers of tests.

• The χ^2 tail probabilities are meaningless for the extreme p-values we seek $\sim 10^{-8}$ when doing 100K tests – they are simply a measure of where are the most extreme associations.

7. GENOMIC CONTROL: Devlin and Roeder (1999):

• A big problem is that there are MANY differences between cases and controls! Also, due to ascertainment, that cases are often more inter-related than controls (e.g. WTCCC).

• This causes inflation on the Y^2 statistic– see red curve above. But how should we estimate the inflation factor.

• Under the null hypothesis; $X = +\sqrt{Y^2}$ is absolute value of $N(0,1)$; Median of X is 0.675.

• Median more robust than mean: recall there will be a few true associations (large X-values). Median is not affected by these.

• Estimate inflation factor λ over the genome as

$$(\text{median}(X_1, ..., X_{100000})/0.675)^2.$$

• Then adjust all Y_j^2 by factor λ; Assume Y^2/λ is χ_1^2.

8. DETECTING POPULATION STRATIFICATION:

• "Desirable" LD caused by original causal variant mutation, maintained by tight linkage.

• "Undesirable LD" caused by admixture, population heterogeneity, ..., allele frequency differences in cases and controls.

• Whereas Devlin and Roeder (1999) use the genome-wide distribution of p-values to adjust the statistics to account for stratification, Pritchard and Rosenberg (1999) use the distribution to test for stratification.

• Test for this population stratification, by choosing ℓ unlinked marker loci over the genome. It is unlikely (??) that any are tightly linked to causal loci. So, for SNPs, these should give valid χ_1^2 distributions.

• Use $Y_S^2 = \sum_{j=1}^{\ell} Y_j^2 \sim \chi_\ell^2$.

• Choice of ℓ? Complex traits? – e.g. height Visscher et al. (2008).

9. WITHIN-LOCUS CONTROL: the TDT (1993):

Spielman et al. (1993); Spielman and Ewens (1996)

• Case-parent trios; Alleles transmitted and not transmitted to n unrelated affected kids from parents;

Transmitted	Nontransmitted allele		
allele	A	B	Total
A	a	b	$a+b$
B	c	d	$c+d$
Total	$a+c$	$b+d$	$2n$

• Assume no ambiguities
— no all-AB trios.

• Only heterozygous parents give information; a and d irrelevant.

• Test statistic $(b-c)^2/(b+c)$ is χ^2_1 under the null hypothesis of no linkage and/or no association. (McNemar's test)

• Note need both linkage AND association, to obtain non-null result.

• Population stratification control; – control is the other allele in the same parent.

Genetic associations for a quantitative trait

1. GENOTYPE ASSOCIATION IN A CASE-CONTROL STUDY
2. ASSOCIATION WITH A QUANTITATIVE TRAIT
3. TDT-TYPE TESTS FOR A QUANTITATIVE TRAIT
4. TRANSMISSION TESTS FOR A QUANTITATIVE TRAIT
5. THE QTDT; (Allison et al. (1999); Abecasis et al. (2000))
6. THE QTDT; TESTING BY PERMUTATION
7. ASSOCIATIONS USING *ibd*; Haseman and Elston (1972)
8. ESTIMATION OF *ibd* PROBABILITIES: SINGLE MARKER
9. ESTIMATION OF *ibd* PROBABILITIES: DENSE SNPs

1. GENOTYPE ASSOCIATION IN A CASE-CONTROL STUDY:

• Again, we are looking for a difference in SNP genotype frequencies between cases and controls.

	AA	AB	BB	total
# B	0	1	2	(fixed by design)
Cases	r_0	r_1	r_2	R
Controls	s_0	s_1	s_2	S
Total	n_0	n_1	n_2	N

• We can do a χ^2-test with 2 degrees of freedom.
Or, reduce to 2×2 for dominant/recessive model.

• Or, better, for additive model, Armitage trend test, with weights 0,1,2, gives χ_1^2 statistic;

$$Y_A^2 = \frac{N(N(r_1 + 2r_2) - R(n_1 + 2n_2))^2}{RS(N(n_1 + 4n_2) - (n_1 + 2n_2)^2)}$$

Compare with previous statistic (Devlin and Roeder, 1999).

2. ASSOCIATION WITH A QUANTITATIVE TRAIT:

	AA	AB	BB
# B	0	1	2
Count	n_0	n_1	n_2
Mean	X_0	X_1	X_2
SS	S_0^2	S_1^2	S_2^2

• Here, we seek an association of phenotype with the 0,1,2 number of B alleles.
• Generally, there is much more information in a quantitative trait.
• Do not have to define cases/controls.

• Can do ANOVA-type test – test for differences among the three genotypes; χ^2 with 2 degrees of freedom.

• Or, pairs of t-test;
e.g. AB vs. AA and BB vs. AA, if all three genotypes have substantial frequencies.

• Or, can do regression-type test; Regress on the number of B alleles; assumes additivity – cf. Armitage trend test.

• All such tests require a model (e.g. Normality), but t-tests are quite robust to non-normality,

3. TDT-TYPE MEANS TESTS FOR A QUANTITATIVE TRAIT:

Allison (1997) proposes tests Q1-Q5; see also Rabinowitz (1997).

• Transmission from heterozygous parent to offspring observed for quantitative trait, Y; assume only one parent is heterozygous.

• Means of individuals genotype AA,AB,BB are μ_{AA}, μ_{AB}, μ_{BB}. $T = 1$ if B was transmitted; $T = 0$ else.

• Q1: Random sampling: a t-test of differences in mean between groups with $T = 1$ and $T = 0$.

• Q2: Extreme sampling: choose offspring with $Y < Z_L$ or $Y > Z_U$. Under the null hypothesis there is no association between L/U and $T = 0/1$; do a χ^2_1 test on the 2×2 table.

• Q3: Extreme sampling: a t-test: select offspring as for Q2, and do t-test as for Q1.

• These tests are all based on Y given T.

4. TRANSMISSION TESTS FOR A QUANTITATIVE TRAIT:

• Q4: Alternatively, we can test T given Y.
H_0: $\Pr(T = 1 \mid Y > Z_U) = \Pr(T = 1 \mid Y < Z_L) = 1/2$.
Do a test of the observed binomial proportions.

• Note 1: When sample from only one tail (e.g. $Z_l = -\infty$ and Z_U is "case" threshold) this just the original TDT (Spielman et al., 1993).

• Note 2: Under segregation distortion, and H_0, $\Pr(T = 1 \mid Y > Z_U) = \Pr(T = 1 \mid Y < Z_L) \neq 1/2$, but $\Pr(T = 1 \mid Y > Z_U) + \Pr(T = 0 \mid Y < Z_L) = 1$. Hence construct test robust to segregation distortion by reversing one of the samples.

• Q5: Testing independence of Y and offspring genotype $X = 0, 1, 2$ for AA,AB,BB.
Uses all families: parent types $AB \times AA$, $AB \times AB$, and $AB \times BB$.
First: regress Y on parent mating type $1, 2, 3$ – remove stratifcation.
Then: add X and X^2 as predictors; test significance – reduction in residual sum-of-squares, using F-tests.

• Power: $Q1 < (Q2, Q4) < Q3 < Q5$.

5. THE QTDT; (Allison et al. (1999); Abecasis et al. (2000)):

- TDT-type test (??), using nuclear families with/without parents.

- Variance component framework: partitions variance due to linkage, LD and stratification.

- $X = -1, 0, 1$ for genotypes AA, AB, BB in offspring; a is additive genetic effect of the *marker* locus on the trait. LD $\leftrightarrow\ a \neq 0$.

- Families $k = 1, ..., K$, offspring $i = 1, ..., n_k$,
Simple means model: $\mathbf{E}(Y_{ki}) = \mu + \beta_a X_{ki}$; $H_0 : \beta_a = 0$.

- There is within-family dependence, both genetic and enviromental:
$\text{Var}(Y_{ki}) = \sigma_a^2 + \sigma_s^2 + \sigma_e^2, \quad \text{Cov}(Y_{ki}, Y_{ki'}) = \pi_{kii'}\sigma_a^2 + \sigma_s^2$

- Stratified means model: Each family k has its own μ_k and marker (i.e. X) probabilities.

- $\mathbf{E}(Y_{ki}) = \mu + \beta_b b_k + \beta_w w_{ki}$, where $b_k = \mathbf{E}(X_{ki})$ given family data, and $w_{ki} = X_{ki} - b_i$

6. THE QTDT; TESTING BY PERMUTATION:

- Use a multivariate Normal likelihood; that is the trait values are Normal with the means and covariance matrices of the model.

- Maximize w.r.t parameters under $H_0 : \beta_a = 0$ (Or, $\beta_w = 0$), and under unconstrained model.

- Subject to many assumptions, the $\log_e$-likelihood-ratio is χ^2, but likely these conditions are not met.

- Ascertainment of extremes (Allison, 1997) will skew the trait distribution.

- Instead, permutation tests are often used; e.g. in case-control studies permute case/control status.

- In families, permutation is harder, due to the dependence among individuals. However, in absense of LD, $\mathbf{w}_k = (w_{ki})$ and $-\mathbf{w}_k$ are equiprobable. Create new data sets by random choice of $\mathbf{w}_k$ and $-\mathbf{w}_k$ independently for each of the K families.

7. ASSOCIATIONS USING LOCUS-SPECIFIC *ibd* ; Haseman and Elston (1972):

- Regression ideas explored much earlier, in linkage sib-pair studies, using *ibd* not marker genotypes.

- Sibs share 0,1,2 genes *ibd* from parents with probabilities 1/4, 1/2, 1/4. At a causal locus, sibs sharing more *ibd* will be more similar.

- For sibs in pair i with quantitative trait values Y_{i1}, Y_{i2},
let $X_i = (Y_{i1} - Y_{i2})^2$; Regress X_i on $Z_i = 0, 1, 2$ *ibd*.
Test for significant (negative) association of X_i and Z_i.

- But Z_i is not observed:
Suppose we can estimate $\Pr(Z|\bullet)$ where ● represents marker data at test locus; suppose probabilities $(\pi_{i0}, \pi_{i1}, \pi_{i2})$.

- Mean sharing = $\mu_i \; = \; \pi_{i1} + 2\pi_{i2}$.
Regress X_i on μ_i or on π_{i2} or ...

- Basic idea: more *ibd* at a (locus linked to) a causal locus results in more phentypic similarity.

8. ESTIMATION OF *ibd* PROBABILITIES: SINGLE MARKER:

- Single-marker estimation: sibs only:
  $\Pr(A) = p$. $\Pr(B) = q = (1-p)$.
  $\Pr(Z) = (1/4, 1/2, 1/4)$, for $Z = 0, 1, 2$.

Sibs	$\Pr(\bullet\|Z)$			$\propto\ \Pr(Z\|\bullet)$		
genos	$Z=0$	$Z=1$	$Z=2$	$Z=0$	$Z=1$	$Z=2$
AA, AA	p^4	p^3	p^2	p^2	$2p$	1
AA, AB	$2p^3q$	p^2q	0	p	1	0
AA, BB	p^2q^2	0	0	1	0	0
AB, AB	$4p^2q^2$	pq	$2pq$	$2pq$	1	1

- Parents remove dependence on p and can add information:
- Example 1: sibs AA, AA;
If parents $AA \times AA$; no information; $\Pr(Z|\bullet) = (1/4, 1/2, 1/4)$
If parents $AA \times AB$: $\Pr(Z|\bullet) = (0, 1/2, 1/2)$
- Example 2: sibs AB,AB;
If parents $AA \times AB$: $\Pr(Z|\bullet) = (0, 1/2, 1/2)$
If parents $AA \times BB$; no information; $\Pr(Z|\bullet) = (1/4, 1/2, 1/4)$
If parents $AB \times AB$: $\Pr(Z|\bullet) = (1/2, 0, 1/2)$

9. ESTIMATION OF *ibd* PROBABILITIES: DENSE SNPs:

• For sibs, and modern marker data, segments of 0,1,2 sharing are easy to determine;
Share 2 *ibd*: genotypes the same over many SNP markers.
Share 1 *ibd*: at all loci share at least 1, but at some loci share only 1
Share 0 *ibd*: at some loci, one sib is AA other is BB.

• $1 \leftrightarrow 2, 1 \leftrightarrow 0$ occur on average at 50 Mbp, so with good markers have little problem (except maybe at edges).

• Use informative SNPs, not in high LD, and allow for error.

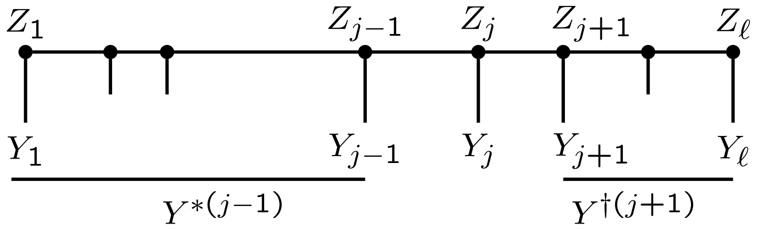
• Note: QTDT and Haseman-Elston uses estimated locus-specific *ibd*. This is in contrast to regressing on estimated proportions of genome shared; cf Visscher et al. (2008). But both use the idea of variation in *ibd*; not all sib pairs are equally "related".

• For more remote relatives, shorter segments, and/or missing data, uncertain allele frequencies, ... *ibd* may be less clear. We need a probability model, and method to estmate $\Pr(Z)$ given marker data.

Hidden Markov models: HMM

1. HIDDEN MARKOV MODELS: HMMs
2. EXAMPLE OF SIB-PAIR *ibd*
3. EXAMPLE OF SIB-PAIR DATA PROBABILITIES
4. THE PROBABILITY OF DATA
5. THE BAUM FORWARD ALGORITH (Baum, 1972)
6. COMPUTING PROBABILITIES OF LATENT STATE
7. EXAMPLE OF SIB-PAIR DATA
8. NUMERICAL EXAMPLE OF SIB-PAIR DATA

1. HIDDEN MARKOV MODELS: HMMs:



- Hidden state is Z_j, $j = 1, 2, ..., \ell$, assumed Markov:

$$\Pr(\mathbf{Z}) \;=\; \Pr(Z_1) \prod_{j=2}^{\ell} \Pr(Z_j \mid Z_{j-1})$$

- Data Y_j, $j = 1, 2, ..., \ell$, depends only on each Z_j:

$$\Pr(\mathbf{Y} \mid \mathbf{Z}) \;=\; \prod_{j=1}^{\ell} \Pr(Y_j \mid Z_j).$$

- Given Z_j, $Y^{*(j-1)}$, Y_j, and $Y^{\dagger(j+1)}$ are mutually independent.
  Also, given Z_j, $Y^{*(j-1)}$, Y_j, and Z_{j+1} are independent.
  Also, given Z_j, $Y^{\dagger(j+1)}$ Y_j, and Z_{j-1} are independent.

2. EXAMPLE OF SIB-PAIR *ibd*:

- Hidden state is $Z_j = 0, 1, 2$ shared *ibd* at locus j, $j = 1, 2, ..., \ell$. $\Pr(Z_1) = 1/4, 1/2, 1/4$ for $Z_1 = 0, 1, 2$.

- Mendel's first law: the 4 meioses from parents to the two sibs are independent.
Maternal *ibd* is independent of paternal *ibd*.
At recombination ρ probability maternal (or paternal) sharing changes is $R = 2\rho(1-\rho) \;\approx\; 2\rho$ for small ρ.

- In absence of genetic interference (crossovers ocurring as Poisson process), Z_j is Markov; transitions depend on j. At recombination ρ_j let $R_j = 1 - (\rho_j^2 + (1-\rho_j)^2)$. ($R_j$ could differ paternal/maternal.)

- Thus we have the transition matrix $\Pr(Z_j \mid Z_{j-1})$:

	0	1	2
0	$(1-R_j)^2$	$2R_j(1-R_j)$	R_j^2
1	$R_j(1-R_j)$	$1-2R_j(1-R_j)$	$R_j(1-R_j)$
2	R_j^2	$2R_j(1-R_j)$	$(1-R_j)^2$

3. EXAMPLE OF SIB-PAIR DATA PROBABILITIES:

- Data Y_j, $j = 1, 2, ..., \ell$, are the genotypes of the two sibs at locus j. Each Y_j depends only on each Z_j, provided there is no LD (??).

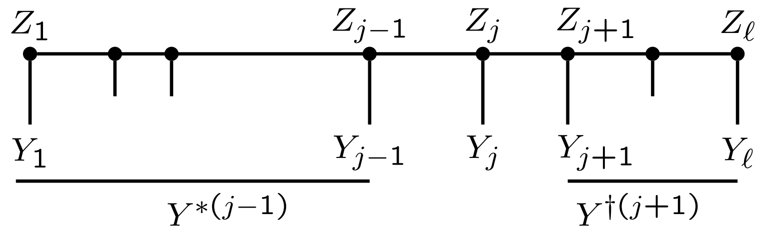
- Suppose at marker j, the frequency of A allele is p_j.
  Recall the basic model: if *ibd* same allelic type, if not *ibd* then of independent allelic types.

	0	1	2
AA, AA	p_j^4	p_j^3	p_j^2
AA, AB	$2p_j^3(1-p_j)$	$p_j^2(1-p_j)$	0
AA, BB	$p_j^2(1-p_j)^2$	0	0
AB, AB	$4p_j^2(1-p_j)^2$	$p_j(1-p_j)^*$	$2p_j(1-p_j)$

And similarly with $A \leftrightarrow B$ and $p_j \leftrightarrow (1-p_j)$.
$*$: $p_j(1-p_j) = p_j^2(1-p_j) + p_j(1-p_j)^2$; $2A, 1B$ or $1A, 2B$.

- It is also possible to incorporate an error model; see Leutn session.
  It is also possible to incorporate LD: see BEAGLE session.

4. THE PROBABILITY OF DATA:



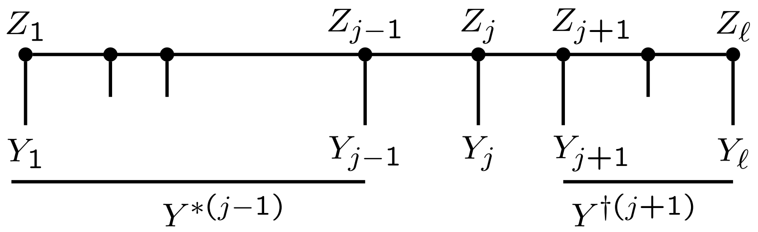
- For data observations $\mathbf{Y} = (Y_j, j = 1, \ldots, \ell)$, we want to compute $\Pr(\mathbf{Y})$. By the first-order Markov dependence of the Z_j,

$$\Pr(\mathbf{Y}) = \sum_{\mathbf{Z}} \Pr(\mathbf{Z}, \mathbf{Y}) = \sum_{\mathbf{Z}} \Pr(\mathbf{Y} \mid \mathbf{Z}) \Pr(\mathbf{Z})$$

$$= \sum_{\mathbf{Z}} (\Pr(Z_1) \prod_{j=2}^{\ell} \Pr(Z_j \mid Z_{j-1}) \prod_{j=1}^{\ell} \Pr(Y_j \mid Z_j)).$$

- Let $Y^{*(j)} = (Y_1, \ldots, Y_j)$, the data along the chromosome up to and including locus j.

5. THE BAUM FORWARD ALGORITH (Baum, 1972):



- With $Y^{*(j)} = (Y_1, \ldots, Y_j)$, note $\mathbf{Y} = Y^{*(\ell)}$. Now define

$$R_j^*(z) \;=\; \Pr(Y_k, k = 1, \ldots, j,\ Z_j = z) \;=\; \Pr(Y^{*(j)},\ Z_j = z)$$

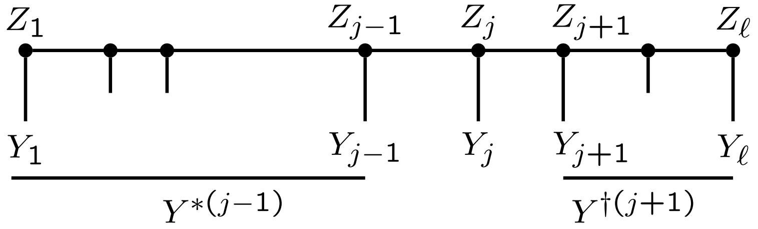
with $R_1^*(z) = \Pr(Y_1|Z_1)\Pr(Z_1 = z)$. Then

$$R_j^*(z) \;=\; \Pr(Y_j \mid Z_j = z) \sum_{z^*} \left[\Pr(Z_j = z \mid Z_{j-1} = z^*) R_{j-1}^*(z^*)\right]$$

for $j = 2, \ldots, l$, with $\Pr(\mathbf{Y}) \;=\; \sum_{z^*} R_\ell^*(z^*)$.

- For each locus $j = 1, ..., \ell$ along the chromosome, for each of M values of $Z_j = z$ we must sum over each of M values of $Z_{j-1} = z^*$. For M states and ℓ loci, compution is order ℓM^2.

6. COMPUTING PROBABILITIES OF LATENT STATE:



- Now also define

$$R_j^\dagger(z) = \Pr(Y_k, k = j+1, \ldots, \ell \mid Z_j = z) = \Pr(Y^{\dagger(j+1)} \mid Z_j = z).$$

- $$\begin{aligned} R_{j-1}^\dagger(z) &= \Pr(Y_k, k = j, \ldots, \ell \mid Z_{j-1} = z) \\ &= \sum_{z^*} \Pr(Y_k, k = j, \ldots, \ell, Z_j = z^* \mid Z_{j-1} = z) \\ &= \sum_{z^*} \Pr(Y_j \mid Z_j = z^*) R_j^\dagger(z^*) \Pr(Z_j = z^* \mid Z_{j-1} = z) \end{aligned}$$

- Then

$$\Pr(Z_j = z \mid \mathbf{Y}) \;=\; \frac{\Pr(\mathbf{Y}, Z_j = z)}{\Pr(\mathbf{Y})} \;=\; \frac{R_j^*(z) R_j^\dagger(z)}{\Pr(\mathbf{Y})}$$

7. EXAMPLE OF SIB-PAIR DATA:

• In this case, $M = 3$ ($Z_j = 0, 1, 2$), so computation is fast and storage of all $R_j^*(z)$ and $R_j^\dagger(z)$ is feasible.

• We obtain $\pi_j(k) = \Pr(Z_j = k \mid \mathbf{Y})$ for $k = 0, 1, 2$ and all j.

• Note, these probabilities $\pi_j(k)$ are often called multipoint *ibd* probabilities. They are multipoint in sense that they use all the data $\mathbf{Y}$ at multiple markers. However, they are the marginal probabilities at each locus j.

• We could also obtain joint probabilities at pairs of loci:

$$\begin{aligned}\Pr(Z_{j-1} = z*, Z_j = z \mid \mathbf{Y}) &= \Pr(\mathbf{Y}, Z_{j=1} = z^*, Z_j = z) \; / \; \Pr(\mathbf{Y}) \\ &= R_{j-1}^*(z^*) \Pr(Y_j | Z_j = z) P(Z_j = z \mid Z_{j-1} = z^*) R_j^\dagger(z) \; / \; \Pr(\mathbf{Y})\end{aligned}$$

• However, even for only 3 states, more than pairs of adjacent loci would get tedious. The *ibd* states at nearby loci are quite dependent. The marginal distributions $\Pr(Z_j = k \mid \mathbf{Y})$ are not everything.

8. NUMERICAL EXAMPLE OF SIB-PAIR DATA:

• Consider 2 sibs, and genotypes at 3 linked loci.
AA, AB; AB, AB; AA, AB.

• Suppose $\Pr(A)$ is 0.9, 0.5, 0.1 at the 3 loci. Then single locus probabilities of 0,1,2 *ibd* are (0.474, 0.526, 0), (0.2, 0.4, 0.4), and (0.09, 0.91, 0).

• Recombination probability between adjacent loci is 0.05. Gives marker-to-marker transition matrix in 0,1,2 *ibd* as

$Z =$	0	1	2
0	0.819025	0.17195	0.009025
1	0.085975	0.82805	0.085975
2	0.009025	0.17195	0.819025

• $R_2^* = \Pr(Y_1, Y_2, Z_2) = (0.0083, 0.0099, 0.0019)$,
$\Pr(\mathbf{Y}) = \Pr(Y_1, Y_2, Y_3) = 0.0001$, and
$R_2^\dagger = \Pr(Y_3 \mid Z_2) = (0.0030, 0.0076, 0.0016)$ giving
$\Pr(Z_2 \mid \mathbf{Y}) = R_2^* R_2^\dagger / P(\mathbf{Y}) = (0.24, 0.73, 0.03)$.

Haplotype blocks and the coalescent

1. HAPLOTYPE BLOCKS MAINTAINED BY LINKAGE
2. SURVIVAL OF A JUNCTION OR RARE VARIANT
3. SURVIVAL OF A JUNCTION OR RARE VARIANT; GRAPHS
4. HAPLOTYPES WITHOUT COANCESTRY
5. THE COALESCENT: IDEALIZED
6. THE COALESCENT: REALITY STRIKES
7. MORE REALITIES AND THE ANCESTRY OF A CHROMOSOME SEGMENT

1. HAPLOTYPE BLOCKS MAINTAINED BY LINKAGE:

• Tight linkage maintains not just pairwise LD but haplotype blocks.

• Using only pairwise LD loses information; there is dependence across the loci within a haplotype.

• The block structure of LD is particularly well captured by BEAGLE (Browning, 2006) model for LD (see later).

• There is heterogeneity of recombination, and there are recombination hot-spots. These lead to breaks in LD. Sperm-typing provides confirmation of some of these inferred hotspots (Li and Stephens, 2003).

• However, most of the "block" structure we see derives not (only) from recombination heterogeneity, but from the shared inheritance of recombination breakpoints, or *junctions* (Fisher, 1954).

• Each new junction operates like a new (rare) variant allele, in terms of its survival, population frequency, and haplotypic background.

2. SURVIVAL OF A JUNCTION OR RARE VARIANT:

• Any new mutation or recombination breakpoint has low probability of long-term survival.

• Hence, conditional on survival, expected counts are high. Expected counts are super-exponential in early generations.

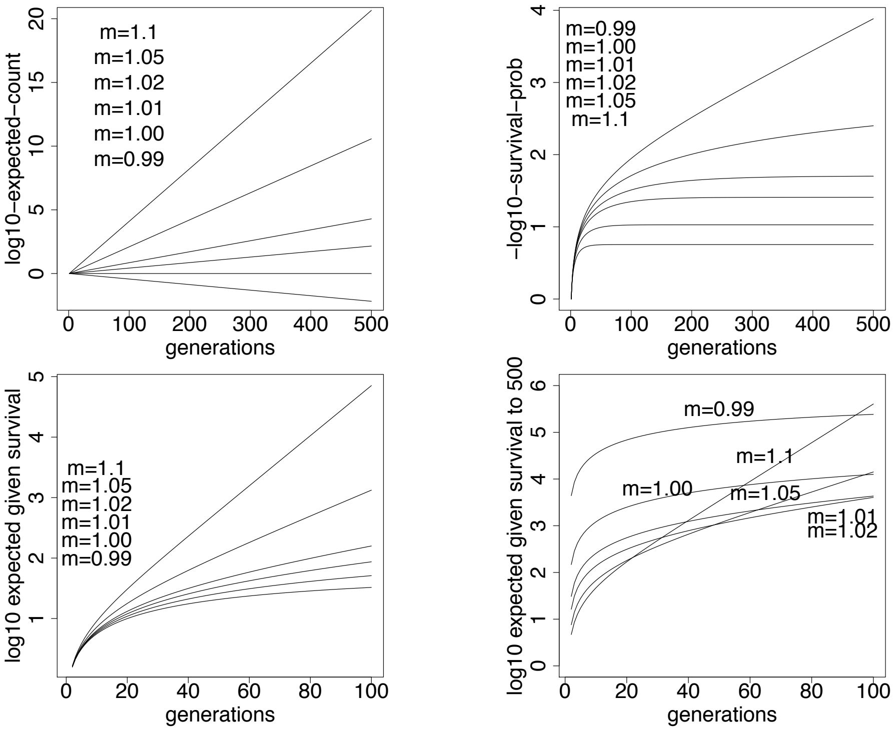
• Hence, conditional on long-term survival, expected counts are even higher, and can be highest for neutral or disadvatageous variants.

• These high-count recombination breakpoints can account for a lot of “block structure” in LD.

• Most rare variants we find are young. For an older, but still rare, variant, the rapid early expansion should be taken into account in considering its ancestry.

• Many variants underlying complex traits are rare – fine-scale mapping of rare variants is still an important problem,

3. SURVIVAL OF A JUNCTION OR RARE VARIANT;GRAPHS:



4. HAPLOTYPES WITHOUT COANCESTRY:

• Earlier papers doing LD mapping did not allow for dependence among the chomosomes due to coancestry. This is well addressed by methods of Graham and Thompson (1998).

• Earlier papers doing LD mapping did not allow for dependence over the chromosome due to descent of segments. This is main focus of McPeek and Strahs (1999).

• McPeek and Strahs (1999) is a long and complex paper. They also present a way to take coancestry dependence into account, but only through pairwise covariance structure.

• For excellent summary review of these earlier LD mapping papers see introduction of McPeek and Strahs (1999)

5. THE COALESCENT: IDEALIZED:

• Samples are smaller than populations! Rather than consider gene descent in a population, it can be useful to consider just the ancestry of an observed sample.

• For sample size k (haploid) in a constant population size $2N$ (N diploids) , the probability any pair share a parent is $1/(2N)$. So the time to common ancestry is exponential with rate $1/(2N)$.

• There are $k(k-1)/2$ pairs, and the minimum of independent exponentials is exponential. $(K << N)$. So the time to the next coalescent event is exponential with rate $k(k-1)/(4N)$, and successive events are independent.

• Rates decrease quadratic in remaining lineages: many recent coalescences, and a few deep branches.

• Expected total time depth: $\sum_2^K 4N/(k(k-1)) \approx 4N$.

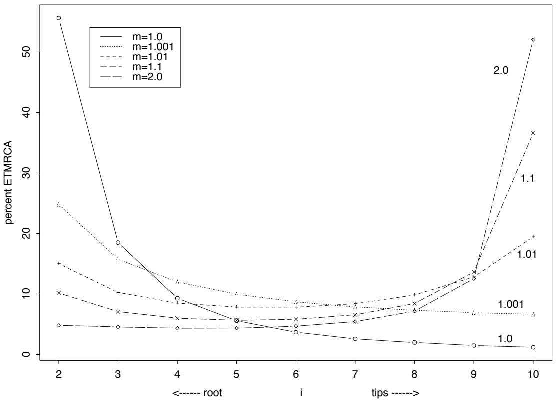
• Expected total branch length: $\sum_2^K 4N/(k-1) \approx 4N \log(K)$

6. THE COALESCENT: REALITY STRIKES:

• Biggest factor is expanding populations. At any point in time this just provides a scaling of time; coalescence rates are inversely proportional to effective population size.

• This leads to negative correlations in coalescent times. If, by chance, a coalescence time is longer, the population at that time will be smaller, so, on average, the next coalescence time will be shorter.

• In expanding populations, earlier inter-coalescence times are relatively shorter, but this does NOT mean phylogeny will be star-shaped.



7. MORE REALITIES AND THE ANCESTRY OF A CHROMOSOME SEGMENT:

• Recall conditional on survival there is rapid early expansion. For a coalescent of a rare variant allele within a larger population it is the “population size” of the variant that counts.

• Ascertained rare variants do not have the general population coalescent. If disease-associated, there may be selection.

• In general, also migration, and population structure (subdivision).

• Along the chromosome recombination events change the coalescent ancestry, leading to the *ancestral recombination graph*.

• This is too complex for us! Instead, we will just consider the segment until broken by a recombination event.

LD mapping via coalescent ancestry

1. LD AS A REFLECTION OF ANCESTRAL HAPLOTYPES
2. RECOMBINATION ON A COALESCENT ANCESTRY
3. THE LATENT RECOMBINANT CLASSES
4. LD MAPPING OF A RARE ALLELE
5. INTERVAL MAPPING LIKELIHOOD
6. FINE-MAPPING BY HAPLOTYPE DECAY
7. THE LIKELIHOOD GIVEN LATENT ANCESTRAL STATE
8. AND MORE POSSIBILITIES

1. LD AS A REFLECTION OF ANCESTRAL HAPLOTYPES:

• Suppose we have a collection of haplotypes believe to carry a rare causal variant, and variant is already mapped to some small region. We wish to fine-scale map the causal locus.

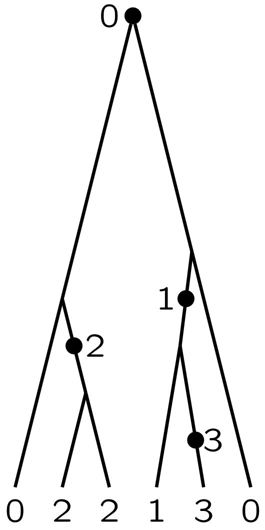
• Example: Graham and Thompson (1998) 50 "disease" haplotypes.
4 markers each with 4 "alleles" (e.g clusters of 3 SNPs)
In control population,
each marker allele at each marker has frequency 0.25.

Allele	Marker				
	Control freq.	M1	M2	M3	M4
A	0.25	9	38	6	5
B	0.25	12	6	2	28
C	0.25	24	4	41	7
D	0.25	5	2	1	10
		50	50	50	50

Clearly there is LD!
Where is the gene?
What is the ancestral haplotype?

2. RECOMBINATION ON A COALESCENT ANCESTRY:

• Suppose we have a good method to generate realizations of the coalescent of the causal variant – will depend on demography, age, frequency, ... the haplotypes are not a random sample from the population! The coalescent history is shorter than for a random sample.



• Consider a marker in the region; say a cluster of SNPs so close that no recombination among them (say at a few Kbp), and at a distance order few 100Kbp from disease locus.

• Consider recombination events on the coalescent between the marker and the causal variant.

• At each recombination event, the disease haplotype picks up a random marker allele from the control population. The descendant lineages carry that allele.

3. THE LATENT RECOMBINANT CLASSES:

• The current haplotypes fall into *recombinant classes* each descended from a recombination event without further recombination.

• In the 6-haplotype figure, there are 4 classes, 2 of size 2, 2 of size 1: $\mathbf{X} = (2, 2, 0, .., 0)$; $X_i =$ number of classes size i.

• Within a class, the haplotypes have same marker allele. In different classes, they are independent.

• $\mathbf{A}$ is the coalescent ancestry, demographic history parameters J. R are the recombination events on $\mathbf{A}$, depends on ρ, and $\mathbf{X}$ is a function of $(\mathbf{A}, R)$.

• Data $\mathbf{Y}$, allelic counts at marker, among the disease haplotypes. $\Pr(\mathbf{Y} \mid \mathbf{X})$ depends on control-population marker allele freqencies $\mathbf{q}$.

• $P_{\mathbf{q}}(\mathbf{Y} \mid \mathbf{X}) = \sum_C P(C \mid \mathbf{X})$ where $C = (c_{ij})$ is a configuration of recombinant classes consistent with $\mathbf{Y}$ and $\mathbf{X}$, such that c_{ij} classes size i are assigned allele j.

4. LD MAPPING OF A RARE ALLELE: due to Jinko Graham:

(and introduction to Monte Carlo Likelihood).

- Now we have a likelihood for ρ, for given J, $\mathbf{q}$;

$$\begin{aligned} L(\rho) &= P_{\mathbf{q},\rho,J}(\mathbf{Y}) = \sum_{\mathbf{A}} \sum_{R} P_{\mathbf{q}}(\mathbf{Y} \mid \mathbf{X}(\mathbf{A}, R)) P_{\rho}(R|\mathbf{A}) P_J(\mathbf{A}) \\ &= \mathbf{E}(P_{\mathbf{q}}(\mathbf{Y} \mid \mathbf{X}(\mathbf{A}, R))) \end{aligned}$$

where here R and $\mathbf{A}$ are random with the appropriate prob dsn.

- How to compute this???; an example of Monte Carlo likelihood.
  Generate $\mathbf{A}$, and then R on $\mathbf{A}$ by Monte Carlo.
  Then we have $\mathbf{X}$ a function of $\mathbf{A}$ and R.

- For each realized $\mathbf{X}$ compute $P_{\mathbf{q}}(\mathbf{Y} \mid \mathbf{X})$ by exact network algorithm, due to Jinko Graham.

- Averaging these probs gives a Monte Carlo estmate of $L(\rho)$.

5. INTERVAL MAPPING LIKELIHOOD:

- Extend to interval mapping:

Markers at distance s (known): trait location at $(\rho, s-\rho)$.
Data $\mathbf{Y} = (Y_1, Y_2)$ at the two flanking markers, M_1 and M_2, with control-population allele frequencies $\mathbf{q}_1$ and $\mathbf{q}_2$.

- $\mathbf{A}$ is the coalescent ancestry at the disease locus, and R_1 and R_2 are the recombinations to left-marker (M_1) and right marker M_2.

- Note recombinations to left and right are independent. There is no assumption here – these are rare events in different meioses of the history.

$$\begin{aligned} L(\rho) &= P_{\mathbf{q},\rho,J}(\mathbf{Y}) = \sum_{\mathbf{A}} \left(\sum_{R_1} P_{\mathbf{q}_1}(\mathbf{Y}_1 \mid \mathbf{X}(\mathbf{A}, R_1)) P_{\rho}(R_1|\mathbf{A}) \right) \times \\ & \left(\sum_{R_2} P_{\mathbf{q}_2}(\mathbf{Y}_2 \mid \mathbf{X}(\mathbf{A}, R_2)) P_{s-\rho}(R_2 \mid \mathbf{A}) \right) \Pr_J(\mathbf{A}) \end{aligned}$$

- But interval mapping still does not consider the whole haplotype.

6. FINE-MAPPING BY HAPLOTYPE DECAY:

McPeek and Strahs (1999)

• Considering markers pairwise loses information; instead consider whole haplotype. Ideally, consider dependence among haplotypes (as per Graham and Thompson (1998)) and across haplotypes.

• As before, assume variant is localized to region.
Suppose haplotype frequencies in a control population are known.

• Label the position of the disease locus as 0. In reality, this is unknown, but likelihood is computed for each hypothesized position.

• Relative to an ancestral haplotype τ meioses ago, the probability a haplotype length d Morgans remains intact is $\exp(-\tau d)$. The length of intact haplotype (left+right) is Gamma$(2, \tau)$, with mean $2\tau^{-1}$.

• First consider 1 haplotype h_{obs}, and suppose we we could see the range of the intact ancestral haplotype.

7. THE LIKELIHOOD GIVEN LATENT ANCESTRAL STATE:

• If, relative to location 0, we observe intact haplotype from $-k$ to j:

$$\begin{aligned} L(\tau^{-1}) &\propto g(\tau^{-1}, -k, j) \\ &= \exp(-\tau d_{-k,j})(1 - \exp(-\tau d_{-k-1,-k}))(1 - \exp(-\tau d_{j,j+1})) \end{aligned}$$

• Latent state $Z = 1$ ancestral, $Z = 0$ non-ancestral.

$$\begin{aligned} \Pr(Z(x+d) = 1 \mid Z(x) = 1) &= \exp(-\tau|d|) \\ \Pr(Z(x+d) = 0 \mid Z(x) = 1) &= (1 - \exp(-\tau|d|)) \\ \Pr(Z(x+d) = 0 \mid Z(x) = 0) &= 1 \end{aligned}$$

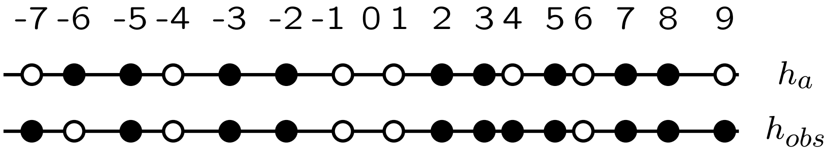
• If breakpoints not observed, use HMM for $Z(x)$ in each direction from position 0.

• With ancestral haplotype h_a as parameter, and data haplotype h_{obs}.

$$\begin{aligned} L(h_a, \tau^{-1}) &= g(\tau^{-1}, -k, j) \\ &\quad P_0(h_{obs}(j+1, j+2, ...))P_0(h_{obs}(-k-1, -k-2, ...)) \end{aligned}$$

where $P_0()$ is the null (population) probability.

8. Summing over latent possibilities:



- Agreement over (-5,3): ancestral segment $k = -5$ to $j = 3$?
  Maybe (-5,-4) is common; agreement by chance?
  Maybe (5,6,7,8) is rare; mutation/error at 4?

- By chance agreement, breakpoints could be closer to 0 than $(-k, j)$

$$L(h_a, \tau^{-1}) \; = \; \sum_{j'=0}^{j} \sum_{k'=0}^{k} \big(\, g(\tau^{-1}, -k', j') \\ P_0(h_{obs}(j'+1, j'+2, ...)) P_0(h_{obs}(-k'-1, -k'-2, ...)) \big)$$

- Allow mutation; haplotype may be ancestral but of different allelic type. Include mutation probabilities, and sum over markers to edge of data set.

8. AND MORE POSSIBILITIES:

- Missing marker data, or missing phase information from genotypes.

- Allow variant ancestral haplotypes. (Multiple origins.);

$$L \; = \; (1-p) L(h_a, \tau^{-1}) + p P_0(h_{obs})$$

- Dependence among haplotypes– model pairwise dependence only– quasi-likelihood approach, but better than assuming independence.

....

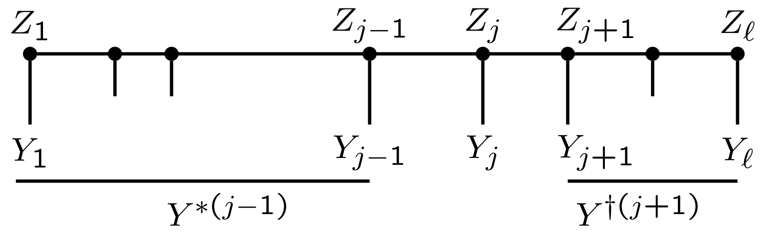
- How to estimate all the parameters introduced in this model: h_a, τ^{-1}, mutation parameters, ...

- Baum algorithm and EM algorithm.

The EM algorithm

1. THE LIKELIHOOD GIVEN HMM DATA
2. THE COMPLETE-DATA LOG-LIKELIHOOD
3. THE EM ALGORITHM FOR PARAMETER ESTIMATION
4. A NON-HMM EXAMPLE
5. EM EXAMPLE contd.
6. BACK TO HMM EXAMPLE
7. THE McPeek and Strahs (1999) EXAMPLE

1. THE LIKELIHOOD GIVEN HMM DATA:



- Hidden state is Z_j, $j = 1, 2, ..., \ell$, assumed Markov:

$$\Pr(\mathbf{Z}) = \Pr(Z_1) \prod_{j=2}^{l} \Pr(Z_j \mid Z_{j-1})$$

- Data Y_j, $j = 1, 2, ..., \ell$, depends only on each Z_j:

$$\Pr(\mathbf{Y} \mid \mathbf{Z}) = \prod_{j=1}^{\ell} \Pr(Y_j \mid Z_j).$$

- $$\Pr(\mathbf{Y}) = \sum_{\mathbf{Z}} \Pr(\mathbf{Z}, \mathbf{Y}) = \sum_{\mathbf{Z}} \Pr(\mathbf{Y} \mid \mathbf{Z}) \Pr(\mathbf{Z})$$

The Baum (1972) algorithm enables us to compute $\Pr(\mathbf{Y})$.

2. THE COMPLETE-DATA LOG-LIKELIHOOD:

• Often, data observations $\Pr(Y_j \mid Z_j)$ and latent transitions $P(Z_j \mid Z_{j-1})$ will have parameters to be estimated.

• Often, if we could observe $\mathbf{Z}$ estimation (MLE) would be easy.

• Then the likelihood based on $\mathbf{Y}$ and $\mathbf{Z}$ would be

$$\log \Pr(\mathbf{Y}, \mathbf{Z}) \;=\; \log \Pr(\mathbf{Y} \mid \mathbf{Z}) \;+\; \log \Pr(\mathbf{Z})$$

$$=\;\; \log(\Pr(Z_1)) \;+\; \sum_{j=2}^{\ell} \log(\Pr(Z_j \mid Z_{j-1})) \;+\; \sum_{j=1}^{\ell} \log(\Pr(Y_j \mid Z_j))$$

• The ECDLL is $\mathbf{E}(\log \Pr(\mathbf{Z}, \mathbf{Y}) \mid \mathbf{Y})$. It is a function of data $\Pr(\mathbf{Y})$ and the parameters.

• The ECDLL is a tool to help us maximize the likelihood $\Pr(\mathbf{Y})$ with respect to parameters. It is NOT the thing we want to maximize.

3. THE EM ALGORITHM FOR PARAMETER ESTIMATION:

• A procedure for maximum likelihood estimation of parameters in latent variable problems.

• Data $\mathbf{Y}$; likelihood $\Pr(\mathbf{Y})$ hard to compute and/or hard to maximize.

• E-step: At current parameter values compute the ECDLL.
M-step Maximize the ECDLL, to obtain new parameter estimates.

• Repeat, alternating E-steps and M-steps until convergence.

• At each round of E-step,M-step, the likelihood $\Pr(\mathbf{Y})$ cannot decrease (and generally increases, except at stationary points of $\Pr(\mathbf{Y})$).

• BEWARE:
1) Goal is to maximise $\Pr(\mathbf{Y})$ NOT the ECDLL. Do not confuse.
2) Local maxima, saddle points, etc. esp. in high dimensions.
3) $\log \Pr(\mathbf{Z}, \mathbf{Y})$ may not be a simple function of $\mathbf{Z}$:
we want the ECDLL NOT $\mathbf{E}(\mathbf{Z} \mid \mathbf{Y})$ (see later).

4. A NON-HMM EXAMPLE:

- Data $Y_1, ..., Y_n$ a sample from a mixture of Poisson distributions:
$$\Pr(Y_i = y) = (p \exp(-\lambda)\lambda^y + (1-p)\exp(-\mu)\mu^y)/y!$$

- $L(p, \lambda, \mu) = \prod_{i=1}^{n} \Pr(Y_i = y_i)$ is easy to evaluate for given data y_i, but hard to maximize w.r.t (p, λ, μ).

- Let $Z_i = 1$ if Y_i is from $\mathcal{P}o(\lambda)$ and $Z_i = 0$ if Y_i is from $\mathcal{P}o(\mu)$.

- Note, if $Z_i = z_i$ observed, estimation is easy: $\widehat{p} = \sum_i z_i/n$,
$\widehat{\lambda} = \sum_{i \in S_\lambda} y_i \;/\; (\#i \in S_\lambda)$, $\quad \widehat{\mu} = \sum_{i \in S_\mu} y_i \;/\; (\#i \in S_\mu)$

- $\Pr(Z_i, Y_i) \;\propto\; (p \exp(-\lambda)\lambda^{Y_i})^{Z_i}((1-p)\exp(-\mu)\mu^{Y_i}))^{1-Z_i}$
$$\log \Pr(Z_i, Y_i) = (Z_i \log p + (1-Z_i)\log(1-p)) \\ +(-\lambda Z_i + Y_i Z_i \log \lambda) + (-\mu(1-Z_i) + Y_i(1-Z_i)\log \mu)$$

- ECDLL: $\mathbf{E}(\log \Pr(\mathbf{Z}, \mathbf{Y})|\mathbf{Y}) \;=\; \sum_{i=1}^{n} \mathbf{E}(\log \Pr(Z_i, Y_i|Y_i))$

and we require only $\mathbf{E}(Z_i \mid Y_i) \;=\; \Pr(Z_i = 1 \mid Y_i)$.

5. EM EXAMPLE contd.:

- E-step: Given parameters, p_0, λ_0, μ_0,
$$\eta_i \;=\; \Pr(Z_i = 1 \mid Y_i = y) \;=\; \Pr(Y_i = y \mid Z_i = 1)\Pr(Z_i) \;/\; \Pr(Y_i = y) \\ \;=\; \exp(-\lambda_0)\lambda_0^y p_0/(p_0 \exp(-\lambda_0)\lambda_0^y + (1-p_0)\exp(-\mu_0)\mu_0^y)$$

- ECDLL is $(\sum_i \eta_i \log p + (n - \sum_i \eta_i)\log(1-p)) \;+$
$(-\lambda \sum_i \eta_i + \sum_i Y_i \eta_i \log \lambda) + (-\mu(n - \sum_i \eta_i) + (\sum_i Y_i - \sum_i Y_i \eta_i)\log \mu)$

- M-step: $p_1 = \sum_i \eta_i / n$, $\lambda_1 = \sum_i \eta_i y_i \;/\; \sum_i \eta_i$,
$\mu_1 = (\sum_i y_i - \sum_i y_i \eta_i) \;/\; (n - \sum_i \eta_i)$.

- $\mathbf{Y} = (0, 1, 2, 6, 7)$, $n = 5$; $\sum_i Y_i = 16$.
One-sample; $\widehat{\lambda} = 16/5$; $\sum_i \log \Pr(Y_i; \lambda = 16/5) = -13.18$.

p	λ	μ	η_3	$\sum_i \eta_i$	$\sum_i \eta_i y_i$	ECDLL	$\log \Pr(\mathbf{Y})$
0.5	1	5	0.686	2.59	2.31	-4.74	-3.72
0.518	0.89	5.67	0.760	2.71	2.49	-4.36	-3.56
0.542	0.92	5.89	0.806	2.77	2.59	-4.23	-3.53
0.553	0.94	6.00	0.827	2.79	2.64	-4.18	-3.53
0.559	0.95	6.05	0.836	2.80	2.66	-4.16	-3.53

6. BACK TO HMM EXAMPLE:

- The complete-data log-likelihood is

$$\log(\Pr(Z_1)) \; + \; \sum_{j=2}^{\ell} \log(\Pr(Z_j \mid Z_{j-1})) \; + \; \sum_{j=1}^{\ell} \log(\Pr(Y_j \mid Z_j))$$

- The ECDLL is $\mathbf{E}(\log \Pr(\mathbf{Z}, \mathbf{Y}) \mid \mathbf{Y})$; often this is a simple function of $\Pr(Z_j \mid \mathbf{Y})$ and $\Pr(Z_{j-1} = z, Z_j = z^* \mid \mathbf{Y})$
$= R^*_{j-1}(z)\Pr(Y_j|Z_j = z^*)\Pr(Z_j = z^*|Z_{j-1} = z)R^\dagger_j(z^*)/\Pr(\mathbf{Y})$.
- Recall for HMM we can compute:

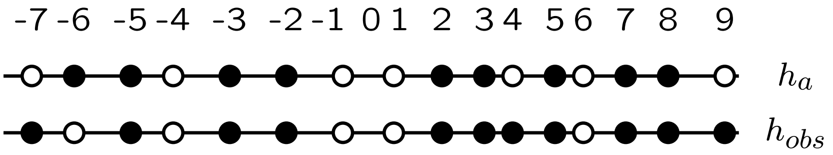
$$\Pr(Z_j = z \mid \mathbf{Y}) \; = \; \frac{\Pr(\mathbf{Y}, Z_j = z)}{\Pr(\mathbf{Y})} \; = \; \frac{R^*_j(z) R^\dagger_j(z)}{\Pr(\mathbf{Y})}$$

and

$$\Pr(Z_{j-1} = z^*, Z_j = z \mid \mathbf{Y}) \; = \; \Pr(\mathbf{Y}, Z_{j=1} = z^*, Z_j = z) \;/\; \Pr(\mathbf{Y})$$
$$= \; R^*_{j-1}(z^*)\Pr(Y_j|Z_j = z)P(Z_j = z \mid Z_{j-1} = z^*)R^\dagger_j(z) \;/\; \Pr(\mathbf{Y})$$

- For indicator Z_i this is often enough to compute the ECDLL.

7. THE McPeek and Strahs (1999) EXAMPLE:



- Recall here $Z_j = 1$ if h_{obs} is ancestral at marker j, and $Z_j = 0$ if not. Ancestral segment is $(-k, j)$.

- Note, if $\mathbf{Z}$ known, e.g.$Z = 1$ from -3 to $+8$, we would know the chance matches at $-5, -4$ and mutation/errors at $+4$, and could estimate parameters.

- Recall have simple HMM for $\mathbf{Z} = (..., Z_{-1}, Z_1, Z_2, ...)$.
$Z_0 = 1$, and can only switch $1 \to 0$ going from $-k$ left or j right, at a rate depending on parameter τ^{-1}.

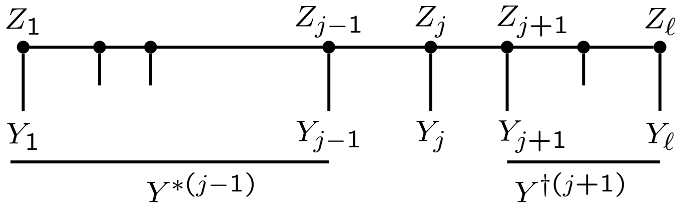
- Thus for given parameters, can compute $\Pr(Z_j, Z_{j+1} \mid h_{obs})$ and $\Pr(Z_{-k}, Z_{-k-1} \mid h_{obs})$ (E-step).

- Give (probabilities of) $\mathbf{Z}$ for each haplotype, can estimate h_a, τ and mutation/error probabilities (M-step).

MCMC and Bayesian sampling

1. SAMPLING THE LATENT HMM STATE (iid)
2. MCMC: THE GIBBS SAMPLER
3. SAMPLING THE LATENT HMM STATE (MCMC)
4. MCEM ESTIMATION OF PARAMETERS
5. BAYESIAN SAMPLING OF PARAMETERS
6. COALESCENT MCMC FOR FINE-SCALE MAPPING
7. FRAMEWORK FOR INFERENCE
8. SAMPLING THE COALESCENT, & MARKERS, GIVEN G
9. THE TRAIT LIKELIHOOD ON THE LOCAL COALESCENT

1. SAMPLING THE LATENT HMM STATE (iid):



- Sometimes we would like to know $\mathbf{Z}$ jointly across loci, not just pairwise (Z_j, Z_{j-1}). If ℓ large, cannot compute, but can sample.
- Compute $R_j^*(z) = \Pr(Y^{*(j)}, Z_j = z)$, $j = 1, 2, 3, ...\ell$ as before.
- First, Z_ℓ is sampled from $\propto R_\ell^*(z)$.
  (All sampling probabilities will be normalized over M z-values.)
- Then, given a realization of $(Z_j = z^*, Z_{j+1}, \ldots, Z_\ell)$,

$$\Pr(Z_{j-1} = z \mid Z_j = z^*, Z_{j+1}, \ldots, Z_\ell, \mathbf{Y}) \;=$$
$$\Pr(Z_{j-1} = z \mid Z_j = z^*, Y^{(j-1)}) \;\propto\; \Pr(Z_j = z^* \mid Z_{j-1} = z) R_{j-1}^*(z)$$

- Normalizing these probabilities, we realize each Z_{j-1}, for $j = \ell, \ell-1, \ldots, 4, 3, 2$ in turn, providing an overall realization $\mathbf{Z} = (Z_1, \ldots, Z_\ell)$ from $\Pr(\mathbf{Z} \mid \mathbf{Y})$.

2. MCMC: The GIBBS SAMPLER:

• Sometimes M is too large to compute $R_j^*(z)$ and $\Pr(\mathbf{Y})$. Then, we also cannot do the i.i.d. sampling of previous slide.

• MCMC is a way of sampling from $\Pr(\mathbf{Z} \mid \mathbf{Y})$ when $\Pr(\mathbf{Y}, \mathbf{Z})$ is easy, but we cannot compute $\Pr(\mathbf{Y})$.

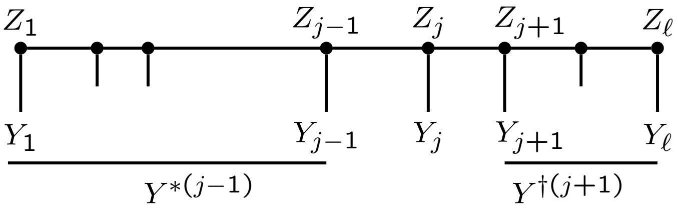
• The Gibbs sampler is a special case of MCMC; at each step, a subset $\mathbf{Z}_u$ of the components of $\mathbf{Z}$ are "updated" from the "full conditionals" $\Pr(\mathbf{Z}_u \mid \mathbf{Y}, \mathbf{Z}_f)$ where $\mathbf{Z}_f$ are the "fixed" components.

• Suppose the current $\mathbf{Z}$ is from $\Pr(\mathbf{Z} \mid \mathbf{Y})$, and $\mathbf{Z}^* = (\mathbf{Z}_u^*, \mathbf{Z}_f)$ is result of resampling $\mathbf{Z}_u$. Then $P^*(\mathbf{Z}^*) = P^*(\mathbf{Z}_u^* \mid \mathbf{Z}_f) P^*(\mathbf{Z}_f)$ $= \Pr(\mathbf{Z}_u^* \mid \mathbf{Z}_f, \mathbf{Y}) \Pr(\mathbf{Z}_f \mid \mathbf{Y}) = \Pr(\mathbf{Z}^* \mid \mathbf{Y})$. That is, the required distribution is equlibrium distribution of this Markov process.

• Subject to various conditions, the average of $g(\mathbf{Z})$ over the chain is $\mathbf{E}(g(\mathbf{Z}) \mid \mathbf{Y})$.

• Do NOT confuse the Markov chain of MCMC sampling with the Markov chain of the HMM.

3. SAMPLING THE LATENT HMM STATE (MCMC):



• One usually easy option is to update a single Z_j at each step:

$$\begin{aligned}\Pr(Z_j \mid Z_i \ (i \neq j), \mathbf{Y}) &= \Pr(Z_j \mid Z_{j-1}, Z_{j+1}, Y_j) \\ &\propto \Pr(Z_{j+1} \mid Z_j) \Pr(Z_j \mid Z_{j-1}) \Pr(Y_j \mid Z_j)\end{aligned}$$

• Compute this for each of M possible values of Z_j, and normalize the probabilities for sampling.

• For example: take a random permutation of $\{1, 2, ..., \ell\}$, and update each Z_j in order of the permutation: this is a *random scan*.

• Note, changes in Z_j may be small if successive Z_j are highly dependent. Then we have "poor MCMC mixing". Better samplers update a block of contiguous Z_j together, but computations get harder.

• In any event, we get a collection of realizations of $\mathbf{Z}$, from $\Pr(\mathbf{Z} \mid \mathbf{Y})$ (at least approximately).

4. MCEM ESTIMATION OF PARAMETERS:

• Back to EM: computing the ECDLL or even $\Pr(Z_j, Z_{j-1} \mid \mathbf{Y})$, may be hard or impossible if M is too large to compute $R_j^*(z)$ etc.

• But we can always do MCMC at current parameter values to get our realizations of $\mathbf{Z}$, and use then to estimate the ECDLL.

• In the McPeek and Strahs (1999) example, we needed $\Pr(Z_j = 1, Z_{j+1} = 0 \mid \mathbf{Y} = h_{obs})$ and $\Pr(Z_{-k-1} = 0, Z_{-k} = 1 \mid \mathbf{Y} = h_{obs})$ to estimate breakpoints $(-k, j)$ from the ancestral haplotype to each current haplotype h_{obs}.

• Given a set of realizations of $\mathbf{Z}$ from $\Pr(\mathbf{Z} \mid \mathbf{Y})$, we can just count the proportion that have the breakpoints at each particular $(-k, j)$. This is an estimate of the required probabilities, for every pair $(-k, j)$.

• The M-step, estimating h_a,τ and any mutation parameters is as before.

• Generally MCEM behaves just as well as EM, but care and large samples needed near convergence.

5. BAYESIAN SAMPLING OF PARAMETERS:

• Suppose θ is a parameter of $\Pr(\mathbf{Z})$ or $\Pr(Z_j \mid Z_{j-1})$, and γ is a parameter of $\Pr(\mathbf{Y}|\mathbf{Z})$ or $\Pr(Y_j \mid Z_j)$.

• MCMC is NOT “Bayesian”; nothing so far is Bayesian.

• However, in very complex multi-parameter problems, a Bayesian approach is useful. That is our parameters (θ and γ) have prior distributions $\pi(\theta)$ and $\pi(\gamma)$.

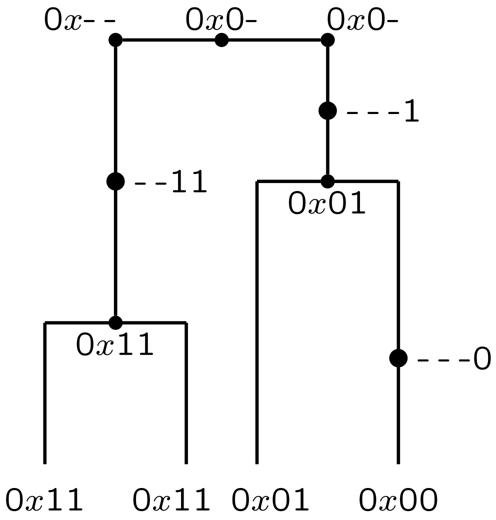
• Extend the MCMC:
Resample $\mathbf{Z}$ given $\mathbf{Y}$, θ, γ (one scan)
Resample θ given $\mathbf{Z}$: $\pi(\theta \mid \mathbf{Z}) \propto \pi(\theta)\Pr(\mathbf{Z} \mid \theta)$.
Resample γ, give $\mathbf{Y}$ and $\mathbf{Z}$: $\pi(\gamma \mid \mathbf{Y}, \mathbf{Z}) \propto \pi(\gamma)\Pr(\mathbf{Y} \mid \mathbf{Z}, \gamma)$.

• Usually more general MCMC methods (Metropolis-Hastings samplers) are needed.

• We obtain a large set of realizations of θ and γ; these provide an estimate of the *posterior distribution* of these parameters.

6. COALESCENT MCMC FOR FINE-SCALE MAPPING:

Zöllner and Pritchard (2005); implemented *TreeDL*.



- Samples coalescent T_x at position x by MCMC, given marker haplotypes G.
- Not full ARG, but local coalescent at x with extent of haplotype present in current sample of haplotypes.
- Under the null hypothesis (no trait association) phenotypes are randomly distributed to the tips of the tree.
- The method looks for groups of similarity of phenotype clustered within the coalescent at specific locations; integrates across allelic heterogeneity.

7. FRAMEWORK FOR INFERENCE:

- Because they sample coalescent of all chromosomes,

—- binary trait: cases and controls; more information extracted.
—- quantitative trait, can be accommodated; population samples.

- $\Phi = (\phi_i)$ vector of phenotypes, individuals i.
  $G = (G_{ij})$ = set of multilocus marker genotypes of i, locus j.

- Adopts a "standard" linkage LR perspective:

$$\text{LR} = \frac{\Pr(\Phi, G; x, P_x)}{\Pr(\Phi, G; P_0)} = \frac{\Pr(\Phi \mid G; x, P_x)}{\Pr(\Phi; P_0)}$$

where P_x and P_0 are penetrances under x and unlinked ”null”.

- Adopts a Bayesian sampling view w.r.t x (also P_x or P_0);

$$\Pr(x|\Phi, G) = \frac{\Pr(\Phi, G \mid x)\pi(x)}{\int_{x'} \Pr(\Phi, G \mid x')\pi(x')} \propto \Pr(\Phi, G \mid x)\pi(x)$$

(However, priors are uniform, so posterior $\propto$ likelihood.)

8. SAMPLING THE COALESCENT, & MARKERS, GIVEN G:

- Estimating $\Pr(\Phi \mid G; x)$;

$$\Pr(\Phi \mid G; x) \ \propto \ \sum_{T_x} \Pr(\Phi \mid x, T_x) \Pr(T_x | G)$$

(G provides information about T_x; given T_x, G and Φ independent.)

- At positions x, sample trees T_x given G. Focus is on shared coancestry of haplotypes (backwards) rather than decay of ancestral haplotype (forwards) (cf. McPeek and Strahs (1999)).

- Marker model incorporates mutation per marker per unit coal.time. and recombination per unit distance per unit coalescent time.

- Samples recombination and mutation events, given G, T_x and hence haplotypes at internal and external nodes.

9. THE TRAIT LIKELIHOOD ON THE LOCAL COALESCENT:

- Hypothetical disease alleles A and B. Single chromosome probabilities:

$$P_A(\phi) = \text{Pr}(A|\phi), \qquad P_B(\phi) = \Pr(B \mid \phi)$$

- Sum over assignment of A and B:
Under T_x: by placing mutations (rate ν) on the tree $\Rightarrow$ alleles cluster.
Under the null; by random (independent) assignment according to mutation model.

- M is branches of coalescent T_x that contain mutations, and γ is collection of current chromosomes carrying A (M, γ unobservable).

$$\Pr(\Phi \mid x, T_x, \nu) = \sum_M \left((\prod_{i \in \gamma} P_A(\phi_i))(\prod_{i \in \gamma^c} \phi_B(\phi_i)) \Pr(M \mid x, T_x, \nu) \right)$$

- Parameters of penetrances $\phi_B()$, $\phi_A()$ are sampled, or for binary data computed on grid, but summation over M is by peeling (as in pedigrees).

Association mapping in structured populations

1. CONTROL FOR POPULATION STRUCTURE
2. INFERRING POPULATION STRUCTURE
3. MCMC SAMPLING: ADDING ADMIXTURE
4. MCMC SAMPLING: INFERENCE
5. ASSOCIATION MAPPING IN STRUCTURED POPULATIONS
6. THE TEST STATISTIC AND TEST
7. COMPARISONS: χ^2, STRAT, and TDT
8. EIGENSTRAT AND MORE

1. CONTROL FOR POPULATION STRUCTURE:

• Recall LD results from population heterogeneity and structure.

• For subdivided population, proportions α_i and frequency p_i in subpopulation i, $p = \sum_i \alpha_i p_i$ and

$$\sigma_F^2 = \Pr(AA) - p^2 = \sum_i \alpha_i (p_i - p)^2$$

• Recall $F_{ST} = \sigma_F^2 / p(1-p)$ is correlation between alleles within subpopulations, and standard measure of structure.

• For 2 populations (e.g. case and control) we can use standard 2×2 table χ_1^2 to test allele frequency differences.

• We can sum many of these χ_1^2 to test genome-wide for structure Pritchard and Rosenberg (1999).

• Two general approaches: correct for structure (genomic control) or model the structure.

2. INFERRING POPULATION STRUCTURE:

• Model-based clustering of individuals on basis of genome-wide unlinked markers. Pritchard et al. (2000a); program STRUCTURE.

• Assume K (sub)populations. $\mathbf{Y}$ are genotypes of individuals.
$\mathbf{Z}$ are latent indicators of population of origin of individuals.
P are (unknown) allele frequencies in all populations..

• Assume HWE and absence of LD within subpopulations;
so $\Pr(Y_j^{(i,a)} \mid Z_i = k, P_{k,j})$ for allele $a = 1, 2$ of indivdual i, is allele frequency in subpopulation $Z_i = k$ of relevant allele at locus j.

• Bayesian approach: $\Pr(\mathbf{Z}, P \mid \mathbf{Y}) \propto \pi(\mathbf{Z})\pi(P)\Pr(\mathbf{Y} \mid \mathbf{Z}, P)$.

• Choice of priors: simplest case:
$\pi(Z_i = k) = 1/K, k = 1, 2, ..., K.$
$\pi(P_{k,jl})$ uniform on $\sum_l P_{k,jl} = 1$ for each locus j in each subpopulation k (For SNPs, minor allele frequency $U(0, 0.5)$ for each locus in each subpopulation.)

3. MCMC SAMPLING: ADDING ADMIXTURE:

• MCMC approach for new realization $(\mathbf{Z}', P')$:
Step 1: Sample P' from distribution given $\mathbf{Y}$, $\mathbf{Z}$.
Step 2: Sample $\mathbf{Z}'$ from distribution given $\mathbf{Y}$, P'.

• Once we sample $Z^{(i,a)}$, then we can make its value k locus-specific: $Z_j^{(i,a)} = k$ if allele $a = 1, 2$ of individual i at locus j comes from population k.

• Additional latent B (note Q in papers): $\beta_k^{(i)}$ proportion of i's genome that is from population k.

• As before: $\Pr(Y_j^{(i,a)} \mid Z_j^{(i,a)} = k, P, B)$ is allele frequency in subpopulation k of relevant allele at locus j,
but now $\Pr(Z_j^{(i,a)} = k \mid B) = \beta_k^{(i)}$.

• Prior for $(\beta_1, ..., \beta_K)$ for each locus and each individual is Dirichlet $\mathcal{D}(\alpha, \alpha, ..., \alpha)$;
α large: each individual an equal mix of populations
α small: each individual from one population, all equally prob.

4. MCMC SAMPLING: INFERENCE:

- MCMC Sampling of $(P', B', \mathbf{Z}', \alpha')$:
  Step 1: Sample (P', B') from $\Pr(P, B \mid \mathbf{Y}, \mathbf{Z})$,
  Step 2: Sample $\mathbf{Z}'$ from $\Pr(\mathbf{Z} \mid \mathbf{Y}, P', B')$.
  Step 3: Update α given $\mathbf{Z}'$.

- Inference for B is primary interest: the labeling problem.
  With K populations, there will be $K!$ symmetric modes.
  By symmetry, overall average is uniform regardless of data!

- Fortunately, the MCMC normally explores one mode
  – but care is required:
  The labeling of the populations is irrelevant;
  we do not want to average over these labelings!

- Inference for K:
  In theory, we could put a prior on K and sample also.
  Pritchard et al. (2000a) propose an *ad hoc* estimate.
  In practice, users run *structure* program with different K.

5. ASSOCIATION MAPPING IN STRUCTURED POPULATIONS:

Pritchard and Donnelly (2001) Pritchard et al. (2000b) ; STRAT

- Recall Pritchard and Rosenberg (1999) provided test for structure using genome-wide markers not in LD; but what if structure is found?

- Pritchard et al. (2000a) uses inferred ancestry to subdivide population (as in Pritchard et al. (2000a)); then do association testing within subpopulations.

- Use the admixed version: $\beta_k^{(i)}$ is proportion of i's genome from population k. Estimate $\widehat{B}$, for all individuals, cases and controls.

- H_0 : no within-population association: i.e. subpopulation allele frequencies at candidate locus are independent of phenotype.
  H_1: there is within-subpopulation association with phenotype.

- Assess significance of test statistics by simulation under H_0.

6. THE TEST STATISTIC AND TEST:

- LR statistic: $\Lambda = \Pr(Y; \widehat{P_1}, \widehat{B}) / \Pr(Y; \widehat{P_0}, \widehat{B})$ where $Y = (y^{(i,a)}, a = 1, 2)$ are observed genotypes at candidate marker locus.

- At candidate locus:
Under H_0: $P_0 = (p_{kj})$ frequencies of alleles j in subpop k.
Under H_1: $P_1 = (p_{kj}^{(\phi)})$ frequencies of alleles j in subpop k among individuals of phenotype ϕ.

$$\Pr(y^{(i,a)} = j; P_0, B, \Phi) \;=\; \sum_k \beta_k^{(i)} p_{kj} \quad \text{regardless of } \Phi$$

$$\Pr(y^{(i,a)} = j; P_1, B, \Phi) \;=\; \sum_k \beta_k^{(i)} p_{kj}^{(\phi(i))}$$

- Use *structure* (Pritchard et al., 2000a) to estimate B.
Use EM to estimate P_0 and P_1; recall latent variables $Z_j^{(i,a)}$ with probabilities $\Pr(Y_j^{(i,a)} \mid Z_j^{(i,a)} = k, P, B) \;=\; p_{kj}$.

7. COMPARISONS: χ^2, STRAT, and TDT:

	Nominal size 0.01			Power	
	STRAT	TDT	χ^2	STRAT	TDT
2 Discrete Populations				1 model **	
0.1, 0.1	0.009	0.010	0.009	0.16	0.06
0.5, 0.1	0.009	0.010	0.260	0.39	0.22
0.9, 0.1	0.010	0.009	0.649	0.07	0.03
Admix of 2 Populations					
0.1, 0.1	0.010	0.009	0.010	0.47	0.41
0.5, 0.1	0.008	0.010	0.370	0.98	1.00
0.9, 0.1	0.005	0.010	0.979	0.76	1.00

- STRAT and TDT give approx correct type-1 error; (χ^2 not!) STRAT conservative in admixed populations; estimation of B not great.

- Same allele associated in both populations; For given number cases TDT wins over STRAT, but not if account for genotyping of parents.

- (Not shown) Different alleles associated in the two populations; STRAT retains power, but TDT does not.

- ** Association in one population only; STRAT often wins over TDT, but not in admixed populations where allele freqs differ widely.

8. EIGENSTRAT AND MORE:

- *structure* is a model-based clustering procedure; can be slow.
- Other methods (EIGENSTRAT) rely on Principal Component Analysis of individuals, to cluster into subpopulations. $\beta_k^{(i)}$ replaced by coefficients on the k first principal components.
- This is much faster, but does not yield estimates of individual admixture. However, it is popular, because much faster.
- Recently, Alexander et al. (2009) have proposed a new model-based algorithm, that is comparable to EIGENSTRAT for speed, using the model of *structure*.
- Uses same model as *structure*, but focuses on maximizing likelihood $\Pr(\mathbf{Y} \mid B, P)$ w.r.t B and P, not on sampling from posterior.
- Uses optimization techniques to update B and P iteratively: much faster than EM (and probably more reliable for a high-dimensional multimodal likelihood).

Association mapping in admixed populations

1. THE BASIS OF LD ADMIXTURE MAPPING
2. BASICS OF ADMIXTURE MAPPING
3. MARKOV MODEL FOR ADMIXTURE
4. HMM FRAMEWORK: FOR EACH INDIVIDUAL i
5. HMM COMPUTATIONS AND MCMC
6. ANCESTRY INFORMATIVE MARKERS (AIMs)
7. ADMIXTURE MAPPING: LOCUS-GENOME TEST
8. ADMIXTURE MAPPING: CASE-CONTROL TEST

1 THE BASIS OF LD ADMIXTURE MAPPING:

McKeigue (2005) and references therein.

• Some traits are more frequent in some ethnic groups (or breeds of cattle?). So at a trait locus, the allele frequencies will differ. At marker loci, the allele frequencies likely differ.

• LD is created by admixture, if allele frequencies differ:

$$\Delta_0 \;=\; m(1-m)\delta_1\delta_2$$

where m is the mixing proportion, and δ_j is the difference in allele frequency at locus j. (Recall the effects of admixture on LD is one reason for Genomic Control!)

• LD is maintained by linkage: in absence of further mixing

$$\Delta_t \;=\; (1-\rho)^t\Delta_0 \text{ (see “DECAY OF LD”).}$$

• In this framework, admixture LD mapping is simply LD mapping, using the substantial LD that can be caused by admixture, and finding regions where this is high.

• This approach does not use all the available information.

2 BASICS OF ADMIXTURE MAPPING:

• Associate the trait with the degree of ancestry at a locus (2,1,0 copies from high-risk population, say).

• Individuals vary in their degree of admixture; we need some form of “genomic control”. Condition on “parental admixture proportions” – parents not observed – in effect, on the genome-wide proportion in the individuals.

• Only linkage results in locus-ancestry associations that are independent of parental admixture, regardless of degree/continuation of admixture. (see “structured association’; Pritchard and Donnelly (2001)).

• Most single loci convey little infomation; these are not fixed differences. So we use an HMM to combine information over the chromosome to estimate the local ancestry.

• STRAT uses correlations within subpopulations at independent markers. ADMIX uses correlations in ancestry over linked loci; inheritance of segments.

3 MARKOV MODEL FOR ADMIXTURE:

Adapted from Patterson et al. (2004).

- For individual i; $\beta_i = \mathrm{Pr}($allele from $Pop1)$ and $\alpha_i =$ rate of ancestry “break-points” (per Morgan) along chromosome. (# generations since admixture).

- Latent state $Z_{ij} = 0, 1, 2$ is number of alleles at locus j deriving from Pop1.

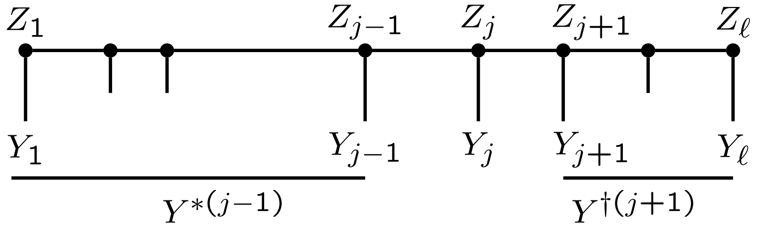
- Prior:

$$\begin{aligned} \eta_{i,0} &= \mathrm{Pr}(Z_{ij} = 0) = (1-\beta_i)^2, \\ \eta_{i,1} &= \mathrm{Pr}(Z_{ij} = 1) = 2\beta_i(1-\beta_i), \\ \eta_{i,2} &= \mathrm{Pr}(Z_{ij} = 2) = \beta_i^2. \end{aligned}$$

- For markers at distance d Morgans;

0 breaks :	$\exp(-2\alpha_i d)$:	$Z_{j+1} = Z_j$.
2 breaks :	$(1-\exp(-\alpha_i d))^2$;	Z_{j+1} from prior
1 break :	$2\exp(-\alpha_i d)(1-\exp(-\alpha_i d))$;	“average”

4 HMM FRAMEWORK: FOR EACH INDIVIDUAL i:



- Hidden state is Z_j, $j = 1, 2, ..., \ell$, $Z_j = 0, 1, 2$.

- Data Y_j is genotype of individual i at locus j.

- Allele frequencies for SNP allele A q_{1j} and q_{2j} in Pops 1 and 2 (assumed estimated from parental populations).

- For given individual, at given locus (drop i, j subscripts):

	$Z = 0$	$Z = 1$	$Z = 2$
$Y = AA$	q_2^2	$q_1 q_2$	q_1^2
$Y = AB$	$2q_2(1-q_2)$	$q_1(1-q_2) + (1-q_1)q_2$	$2q_1(1-q_1)$
$Y = BB$	$(1-q_2)^2$	$(1-q_1)(1-q_2)$	$(1-q_1)^2$

5 HMM COMPUTATIONS AND MCMC:

• For given parameter values we can compute

$$\gamma_{i,k}(j) \;=\; \Pr(Z_{ij} = k \mid (Y_{i1}, ..., Y_{i\ell})),\; k = 0, 1, 2$$

• MCMC is used to sample all parameters:
α_i, β_i for all i; q_{1j}, q_{2j} for all j.

• Initial values: $\alpha_i = 6$, β_i estimate from data on i treating loci as unlinked, allele freqs. from "parent" populations.

• Prior on β_i; Beta dsn with mean/spread 0.2 ± 0.12.
Prior on α_i; Gamma dsn with mean/spread 6 ± 2.
Allele frequencies; centred on modern "parental" populations, with dispersion hyperparameter τ.

• MCMC provides posterior realizations of all parameters.
Estimates used are posterior means.

6 ANCESTRY INFORMATIVE MARKERS (AIMs):

• Instead of correcting for structure, we *use* structure.

• We want markers with substantial differences among populations, i.e. geographic differentiation (but we may not want strong selection).

• Informativeness for ancestry (Rosenberg et al., 2003):
Suppose SNP has allele frequencies p_k and $(1 - p_k)$ in population $k = 1, 2, ..., K$, and $p = (1/K)\sum_k p_k$. Information is $-p \log p - (1-p)\log(1-p) + (1/K)\sum_k (p_k \log p_k + (1 - p_k)\log(1-p_k))$. If all $p_k = p$, information is 0; otherwise positive.

• Example 1: African-American Smith et al. (2004);
Require few missing data, and HWE within parent populations.
Require high informativeness, and homogeneity within parent populations. Remove markers within 50kbp or in LD.
Result is 3,011 SNP markers across the human autosomal genome.

Example 2: European populations; Price et al. (2008); Panel of 300 markers sufficient to correct for stratification of European populations.

7. ADMIXTURE MAPPING: LOCUS-GENOME TEST:

• Compare locus-specific estimates of proportion of genome from Pop1, with genome-wide average.

• ψ_k is increase in disease risk due to having k alleles from Pop1, relative to $\psi_0 = 1$. Note, population risks are smaller than allelic risks, since it averages over alleles.

• Use a LR test to compare H_1: disease locus associated with locus j, vs H_0: no disease locus near j.

• The locus-genome statistic (for case individual i at locus j):

$$L_{ij} = \frac{\Pr(\text{case}; H_1)}{\Pr(\text{case}; H_0)} = \frac{\gamma_{i,0}(j) + \gamma_{i,1}(j)\psi_1 + \gamma_{i,2}(j)\psi_2}{\eta_{i,0} + \eta_{i,1}\psi_1 + \eta_{i,2}\psi_2}$$

• Robustness to choice of ψ_1, ψ_2.

• Pointwise, or average, or maximum over the genome.
Significance thresholds detemined by parametric simulation.

8. ADMIXTURE MAPPING: CASE-CONTROL TEST:

• Compare cases with controls at each locus j in the genome.

• Protects against stratification; a deviation from genome-wide average of population ancestry seen in cases but not controls provides evidence.

• For individual i at locus j: $y_i(j) = 2\beta_i - (2\gamma_{i,2}(j) + \gamma_{i,1}(j))$.
Use t-statistic T_j to test differences in $y_i(j)$ between cases and controls.

• Advantages: t-test is robust to heterogeneity of variance (over i).
No specific trait model (i.e. ψ_1,ψ_2) required.
No simulation to determine significance thresholds required.

• Disadvantage: “Randomness” in controls contributes to uncertainty.

• Advantage of admixture mapping in general:
Segments are much larger than LD – many fewer markers required.
Problems of multiple testing reduced – 1,000 tests vs 10^6.

Inferring *ibd* segments; two chromosomes

1. THE AIM OF *ibd* MAPPING (SB *ibd* slide)
2. APPLICATIONS OF *ibd* MAPPING (SB *ibd* slide)
3. *ibd* MAPPING: WILL IT BE USEFUL? (SB *ibd* slide)
4. LATENT *ibd* MARKOV MODEL FOR 2 CHROMOSOMES
5. THE DATA MODEL
6. *ibd* AND PARAMETER ESTIMATION
7. LEUTENEGGER (2003) RESULTS: ESTIMATING f OR β
8. LEUTENEGGER (2003) DETECTION OF INDIVIDUAL *ibd*
9. ERROR MODELLING IN HBD SEGMENTS (SB *ibd* slide)
10. GENOMIC CONTROL

Slides from Sharon Browning are removed for the web version of these lecture notes. I am very grateful to Sharon Browning for sending me these slides for use in giving the lectures.

1,2,3: Three Browning IBD-talk slides:

These three slides removed for web version as requested by Sharon Browning.

4. LATENT *ibd* MARKOV MODEL FOR 2 CHROMOSOMES:

(Leutenegger et al., 2003)

- Two-parameter Markov model: marginal prob β, rate change α. In reality, *ibd* is not Markov and expected segment length depends on # meioses to the common ancestor.

- Markov rate matrix between non-*ibd* (0) and *ibd* (1) is

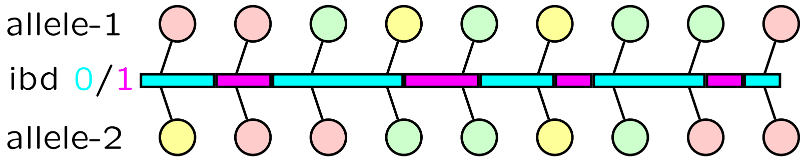
$$Q = \begin{pmatrix} -\alpha\beta & \alpha\beta \\ \alpha(1-\beta) & -\alpha(1-\beta) \end{pmatrix} = \alpha(-I + \begin{pmatrix} 1 \\ 1 \end{pmatrix}(1-\beta, \beta))$$

Model of "segments" of exponential length (mean α^{-1}) each independently of type 1 (*ibd*) with probability β.

- Thus *ibd* segments are exponential with expected length $(\alpha(1-\beta))^{-1}$ and the equilibrium marginal probability of *ibd* is β. The relative rate of gain vs loss of *ibd* is $\beta/(1-\beta)$.

- Estimation of β and α (see later) or *ad hoc* choice depending on overall *ibd* level (β), and typical segment length (e.g $\alpha = 1 \times 10^{-6}$bp).

5. THE DATA MODEL:



- Allele frequencies q_i of alleles a_i assumed known: in reality they can be well estimated from genotypic samples.
- *ibd* ⇒ same allele; non-*ibd* ⇒ independent alleles.
  Allow error so different alleles can still be *ibd*.

	non-*ibd*	*ibd*
$a_i,\ a_i$	q_i^2	$(1-\varepsilon)q_i + \varepsilon q_i^2$
$a_i,\ a_j (i < j)$	$2q_iq_j$	$\varepsilon 2q_iq_j$

- Given a model, a standard HMM forward-backward algorithm gives $\Pr(\textit{ibd}(j) \mid \mathbf{Y})$, at each positions j where $\mathbf{Y}$ are allele types on the chromosomes over all loci.

6. *ibd* AND PARAMETER ESTIMATION:

• Suppose $Z_{c,j} =$ *ibd*(j) at locus j on chrom. c; we can compute $\Pr(Z_{j-1}, Z_j \mid \mathbf{Y})$. If $Z_{c,j}$ "observed"; the log-likelihood would be

$$\sum_{c,j} \log \Pr(Y_{c,j} \mid Z_{c,j}; \varepsilon, \mathbf{q}_{c,j}) + \sum_c \Big(\Pr(Z_{c,1}; \beta) + \sum_j \log \Pr(Z_{c,j} \mid Z_{c,j-1}; \alpha, \beta, d_j)\Big)$$

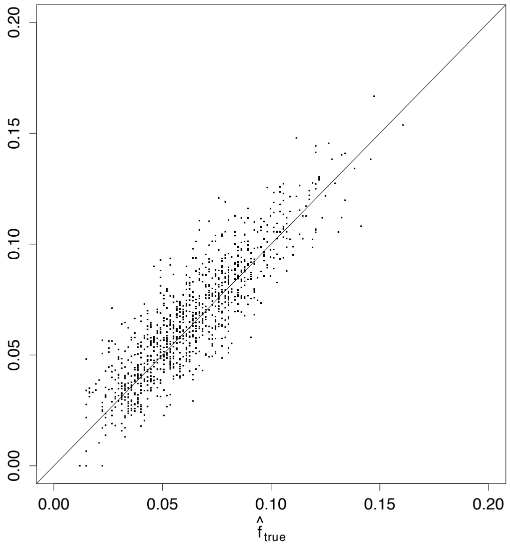
• First term estimates ε, from those $Z_{c,j} = 1$.
Next, $\{Z_{c,1}\}$ is binomal sample $\Pr(Z_{c,1} = 1) = \beta$ (ignore).

• For simplicity, suppose markers equidistant, and counts of transitions are $T_{0,1}$, $T_{0,0}$, $T_{1,0}$ and $T_{1,1}$; Let $h = (1 - \exp(-\alpha d))$. So probs from $Z_{c,j-1}$ to $Z_{c,j}$ are $\begin{pmatrix} 1 - h\beta & h\beta \\ h(1-\beta) & (1 - h(1-\beta)) \end{pmatrix}$.

• $\hat{h}\hat{\beta} = T_{0,1}/(T_{0,0} + T_{0,1})$ and $\hat{h}(1 - \hat{\beta}) = T_{1,0}/(T_{1,1} + T_{1,0})$. With $N_0 = (T_{0,0} + T_{0,1})$, $N_1 = (T_{1,1} + T_{1,0})$, $\hat{h} = T_{0,1}/N_0 + T_{1,0}/N_1$ and $\hat{\beta} = T_{0,1}/\hat{h}N_0$.

• EM algorithm;
E-step: compute $\mathbf{E}(T_{i,k}|\mathbf{Y})$ from $\Pr(Z_{c,j-1} = i, Z_{c,j} = k \mid \mathbf{Y}_c)$.
M-step: restimate parameters from current $\mathbf{E}(T_{i,k}|\mathbf{Y})$.

7. LEUTENEGGER (2003) RESULTS: ESTIMATING f (OR β) :

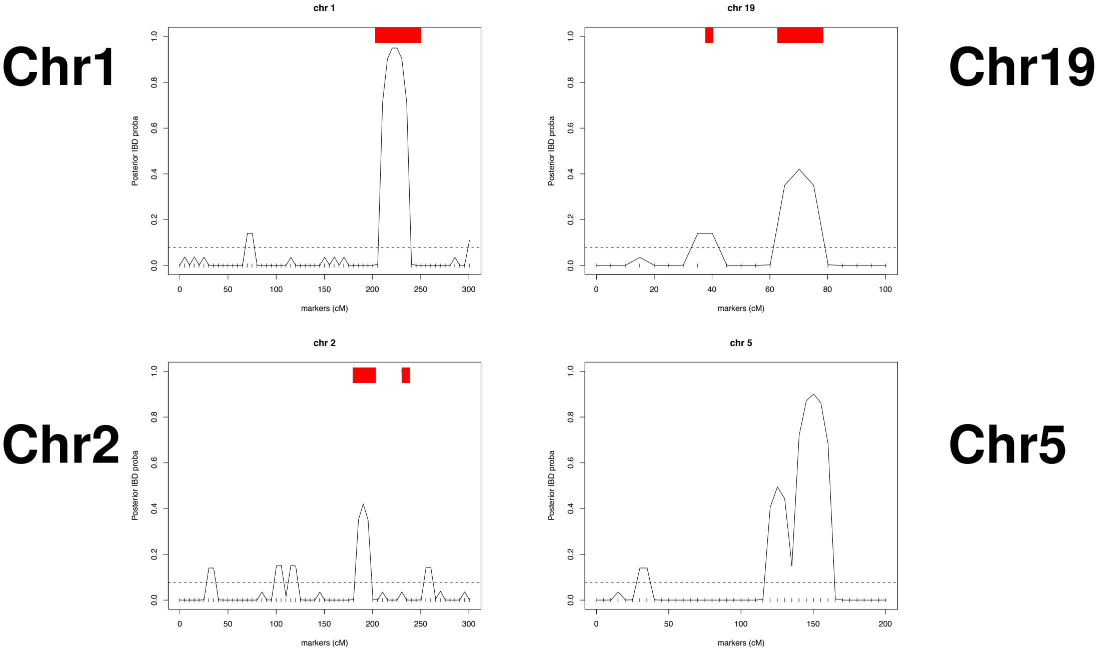


Offspring of 1000 first cousin pairs:
Estimation of f using 5cM microsatellite map (630 markers)

$\overline{f} = 1/16 = 0.0625$

At most 50 "indep" *ibd* events.
The human genome is short.

8. LEUTENEGGER (2003) DETECTION OF INDIVIDUAL *ibd*:



9. Another Browning IBD-talk slide:

This slide removed for web version as requested by Sharon Browning.

10. GENOMIC CONTROL:

- Individuals vary in their degree of inbreeding; we need some form of "genomic control".

- Standard MLE methods also give an individual-based confidence interval for the genome-wide estimate of β_i for individual i.

- To assess significance of a region, within an individual, compare the estimated conditional $\Pr(Z_{c,j} = 1 \mid \mathbf{Y}_c)$, with the genome-wide confidence interval for this individual.

- Unequal spacing etc.; need more numerical methods to implement the EM algorith, but idea is same.

- Method implemented in *FEstim* program (Leutenegger et al., 2006).

- Methods were first used (Leutenegger et al., 2006) to enhance homozygosity mapping when there is cryptic relatedness.

- Now being used in mapping of rare recessives, where case parents are not known to be related. (See IGES abstracts 2009, 2010).

BEAGLE: Haplotype and *ibd* imputation

1. SHOULD WE MODEL LD IN *ibd* INFERENCE?
2. PUTTING LD INTO THE HMM
3-11. THE BEAGLE LD MODEL (SB hap slides)
12-14. THE BEAGLE LD MODEL (SB imput slides)
15. ADVANTAGES OF BEAGLE MODEL (SB hap slides)
16-17. *ibd* INFERENCE WITH LD (SB ibd slides)

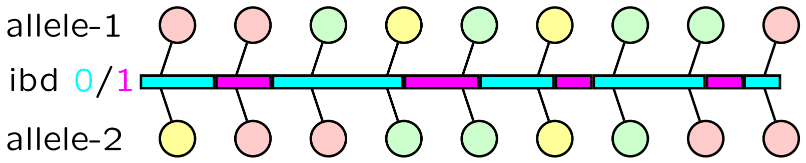
Slides from Sharon Browning are removed for the web version of these lecture notes. I am very grateful to Sharon Browning for sending me these slides for use in giving the lectures.

1. SHOULD WE MODEL LD IN *ibd* INFERENCE?:

● Do we want to? Recall LD is a reflection of coancestry.
By conditioning out the LD we are conditioning out this coancestry.
But if we do not, we get many “false-positive” *ibd* signals when using dense markers.

● LD is also caused by stratification/admixture. This can cause long-range LD. However, this is not much of a problem in inferring *ibd* segments, so long as there is plenty of variation within subpopulations.

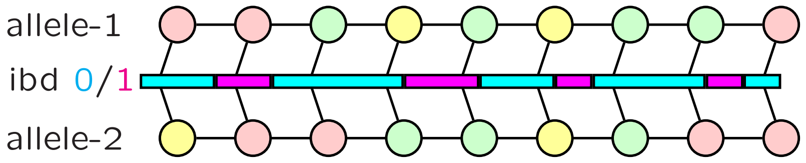
● If we do want to, then how?



● We want a Markov model for the alleles along a haplotype, to superimpose on our *ibd* HMM.

2. PUTTING LD INTO THE HMM:

● Combined HMM for alleles and *ibd*.



$$\Pr(Y_j, Z_j \mid Y_{j-1}, Z_{j-1}) = \Pr(Y_j \mid Z_j, Y_{j-1}) \Pr(Z_j \mid Z_{j-1})$$

● Can compute under this augmented HMM, but the simple 1st-order HMM LD model does not fit; see e.g. Fu and Thompson (2007).

● Condition each Y_j on preceeding SNP with highest LD,
may be better. See e.g. Albrechtsen et al. (2009).

● Variable length Markov chains of BEAGLE (Browning, 2008)– works well. Note, for *ibd* Browning (2008) deals only with pair of chromosomes ($Z = 0, 1$). Browning and Browning (2010) deals with pair of individuals, but still only uses 2-state *ibd*

3-11. Nine Browning slides:

These slides from Sharon Browning's haplotyping talk.

These slides removed for web version as requested by Sharon Browning.

12,13,14. Three Browning imputation slides:

These slides from Sharon Browning's imputation talk.

These slides removed for web version as requested by Sharon Browning.

15. Browning haplotyping slide:

This slides from Sharon Browning's haplotype talk.

This slide removed for web version as requested by Sharon Browning.

16,17. Browning *ibd* slide:

These two slides from Sharon Browning's ibd talk.

These slides removed for web version as requested by Sharon Browning.

ibd between two individuals

1. THE FOUR GENES OF TWO INDIVIDUALS
2. *ibd* OF TWO NON-INBRED RELATIVES
3. FROM KINSHIP TO *ibd* PROBABILITIES
4. HMM FOR *ibd* BETWEEN NON-INBRED INDIVIDUALS
5. THE PLINK MODEL FOR GENOTPYES
6. *ibd* ONLY BETWEEN INDIVIDUALS IS OVER-SIMPLIFICATION
7. MODELS FOR GENOMIC IBD ESTIMATION
8. THE FOUR GENES OF TWO INDIVIDUALS– AGAIN
9. MODEL FOR POPULATION *ibd* AT ONE LOCUS
10. POPULATION *ibd* OVER THE GENOME (Thompson, 2008)
11. INFERRING *ibd* FROM POPULATION DATA
12. INFERRING *ibd* FROM POPULATION DATA: RESULTS

1. THE FOUR GENES OF TWO INDIVIDUALS:

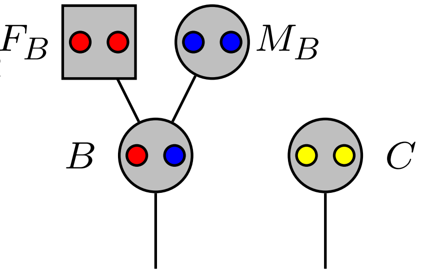
ibd pattern		*ibd* label	*ibd* group	state description	
B_1	B_2			individuals	genes
$p\ m$	$p\ m$			*ibd* within	shared
● ●	● ●	1 1 1 1	1 1 1 1	B_1, B_2	4 genes *ibd*
● ●	● ○	1 1 1 2	1 1 1 2	B_1	3 genes *ibd*
● ●	○ ●	1 1 2 1			
● ○	● ●	1 2 1 1	1 2 1 1	B_2	3 genes *ibd*
● ○	○ ○	1 2 2 2			
● ●	○ ○	1 1 2 2	1 1 2 2	B_1, B_2	none
● ●	○ †	1 1 2 3	1 1 2 3	B_1	none
● ○	† †	1 2 3 3	1 2 3 3	B_2	none
● ○	● ○	1 2 1 2	1 2 1 2	none	2 genes shared
● ○	○ ●	1 2 2 1			
● ○	● †	1 2 1 3	1 2 1 3	none	1 gene shared
● ○	† ●	1 2 3 1			
● ○	○ †	1 2 2 3			
● ○	† ○	1 2 3 2			
● ○	† ⋆	1 2 3 4	1 2 3 4	none	none

2. *ibd* OF TWO NON-INBRED RELATIVES:

- For two non-inbred relatives, 7 states, 3 classes, 2 probs $\kappa_i = \Pr(i \text{ genes } ibd)$, $\kappa_2 + \kappa_1 + \kappa_0 = 1$. Also $\psi = \frac{1}{2}\kappa_2 + \frac{1}{4}\kappa_1 + 0\kappa_0 \;=\; \frac{1}{4}(2\kappa_2 + \kappa_1)$. If $\kappa_2 = 0$, $\kappa_1 = 4\psi$.
- Computing kinship ψ in known pedigrees:

Provided B is not C nor an ancestor of C

$$\begin{aligned}\psi(B,C) &= (\psi(M_B,C) + \psi(F_B,C))/2\\ \psi(B,B) &= (1 + f_B)/2\\ &= (1 + \psi(M_B,F_B))/2\end{aligned}$$



Boundary conditions:
If A is a founder, and not an ancestor of C,

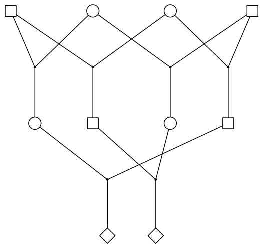
$$\psi(A,A) \;=\; 1/2 \text{ and } \psi(A,C) \;=\; 0$$

3. FROM KINSHIP TO *ibd* PROBABILITIES:

The following equations relate ψ and κ_i, $i = 0, 1, 2$.

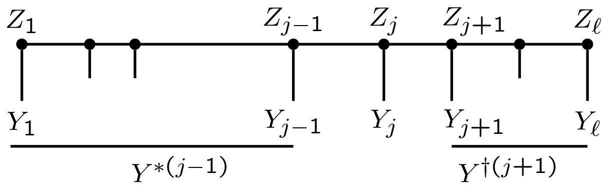
$$\begin{aligned}\psi &= (1/2)\kappa_2 + (1/4)\kappa_1 \;=\; (1/4)(1 + \kappa_2 - \kappa_0)\\ \psi(B_1,B_2) &= (1/4)(\psi(M_1,M_2) + \psi(M_1,F_2)\\ &\quad + \psi(F_1,M_2) + \psi(F_1,F_2))\\ \kappa_2(B_1,B_2) &= \psi(M_1,M_2)\psi(F_1,F_2) + \psi(M_1,F_2)\psi(F_1,M_2)\\ \kappa_1(B_1,B_2) &= 4\psi(B_1,B_2) - 2\kappa_2(B_1,B_2)\\ \kappa_0(B_1,B_2) &= 1 - \kappa_1(B_1,B_2) - \kappa_2(B_1,B_2)\end{aligned}$$

- Example: Quadruple-half-first-cousins.

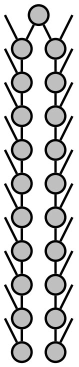


Then all four of $\psi(M_1,M_2)$, $\psi(F_1,F_2)$, $\psi(M_1,F_2)$ and $\psi(F_1,M_2)$ are non-zero without the children being inbred.
$\psi(M_1,M_2) = \psi(F_1,F_2) =$
$\psi(M_1,F_2) = \psi(F_1,M_2) = 1/8$
so $\kappa_2 = 1/32$, $\psi = 1/8$,
$\kappa_1 = 4\psi - 2\kappa_2 = 7/16$,
$\kappa_0 = 1 - \kappa_2 - \kappa_1 = 17/32$

4. HMM FOR *ibd* BETWEEN NON-INBRED INDIVIDUALS:



- Purcell et al. (2007) use an HMM to estimate locus-specific probabilities of sharing $Z_j = 0, 1, 2$ *ibd* at locus j.



- For one pair haplotypes a: $\Pr($*ibd*$) = (1/2)^{m-1}$.
Recall half-sib *ibd*; $R = 1 - (\rho^2 + (1-\rho)^2)$
$\Pr(1 \to 0) = a_{10} = 1 - (1-\rho)^{m-2}(1-R)$.
$\Pr(0 \to 1) = a_{01} = a_{10}/(2^{m-1} - 1)$.

For 2 pairs haplotypes a,b:

$$\Pr(Z_{j+1} = 0 \mid Z_j) \;=\; \begin{pmatrix} a_{00}b_{00} \\ (a_{00}b_{10} + a_{10}b_{00})/2 \\ a_{10}b_{10} \end{pmatrix}$$

and similarly for other transitions.

5. THE PLINK MODEL FOR GENOTPYES:

- Now need a model for $\Pr(Y_j \mid Z_j)$ where Y_j is pair of genotypes at locus j.

- Instead of the usual allele-frequency model, Purcell et al. (2007) treat alleles as sampled *without replacement* from the set of alleles of the individuals in the study.

- This has consequence that one allele being of type A *decreases* the probability that another (non-*ibd*) allele is type A.

- This might make sense if the sample was the whole population:
Example 1: in an endangered species, what is the allele frequency?
Example 2: in the pedigree of a genetic isolate; when infer some copies of an allele are *ibd*, this reduces the number of “independent copies” of this allele.

- It does not make sense (to me) in context of a sample from a large population.

6. *ibd* ONLY BETWEEN INDIVIDUALS IS OVER-SIMPLIFICATION:

- BEAGLE: Divides into state of any-*ibd* or no-*ibd* between the individuals.

- PLINK: Allows only 0,1,2 *ibd*-between. No-*ibd* within.

- Am I as related to any of you
  as my parents are related to eachother?
  ibd within is at least as great as between
  Prior screening for *ibd* within??– and cnv/deletions?

- PLINK: m, the "minimum number of meioses to coancestry". is estimated from considering genotypes as two "haplotype" pairs with independent *ibd*, and fitting genome-wide sharing estimates of $Z = 0, 1, 2$ obtained from SNP-by-SNP method of moments??

7. MODELS FOR GENOMIC IBD ESTIMATION:

	Leut.	PLINK	Brow	Thom
Data structure	1 ind	2 ind	2 chr/ind	more!
Phased chromosomes	No	No	Yes/No	Yes
Genotypic data	Yes	Yes	No/Yes	Yes
Error allowed	Yes	No	No/Yes	Yes
Linkage disequilibrium	No	No	Yes	No?
Multiple chromosomes jointly	No	No*	No**	Yes

- Leut is Leutenegger et al. (2003); less dense data.
  ibd model and error model of Leut can be extended to multiple genomes. HMM structure can be extended to include LD, but ....
- PLINK is Purcell et al. (2007)
  *Models 2 individuals, but allows only 2,1,0 *ibd* between, not within.
- Brow is Browning (2008) and Browning and Browning (2010).
  **: 2 *ibd* states only. Includes LD, but LD reflects coancestry;
  Do we want to condition out this LD ??
- Thom is Thompson (2008) and Thompson (2009)

8. THE FOUR GENES OF TWO INDIVIDUALS– AGAIN:

ibd patt^n		*ibd*	*ibd*	state descriptions		
B_1	B_2	label	group	partition	*ibd* within	“Ewens”
1 2	3 4			of genes	& between	(a_1, a_2, a_3, a_4)
● ●	● ●	1 1 1 1	1 1 1 1	(1,2,3,4)	1, 1, 1	(0,0,0,1)
● ●	● ○	1 1 1 2	1 1 1 2	(1,2,3)(4)	1, 0, 1	(1,0,1,0)
● ●	○ ●	1 1 2 1		(1,2,4)(3)		(1,0,1,0)
● ○	● ●	1 2 1 1	1 2 1 1	(1,3,4)(2)	0, 1, 1	(1,0,1,0)
● ○	○ ○	1 2 2 2		(1)(2,3,4)		(1,0,1,0)
● ●	○ ○	1 1 2 2	1 1 2 2	(1,2)(3,4)	1, 1, 0	(0,2,0,0)
● ●	○ †	1 1 2 3	1 1 2 3	(1,2)(3)(4)	1, 0, 0	(2,1,0,0)
● ○	† †	1 2 3 3	1 2 3 3	(1)(2)(3,4)	0, 1, 0	(2,1,0,0)
● ○	● ○	1 2 1 2	1 2 1 2	(1,3)(2,4)	0, 0, 2	(0,2,0,0)
● ○	○ ●	1 2 2 1		(1,4)(2,3)		(0,2,0,0)
● ○	● †	1 2 1 3	1 2 1 3	(1,3)(2)(4)	0, 0, 1	(2,1,0,0)
● ○	† ●	1 2 3 1		(1,4)(2)(3)		(2,1,0,0)
● ○	○ †	1 2 2 3		(1)(2,3)(4)		(2,1,0,0)
● ○	† ○	1 2 3 2		(1)(2,4)(3)		(2,1,0,0)
● ○	† ⋆	1 2 3 4	1 2 3 4	(1)(2)(3)(4)	0, 0, 0	(4,0,0,0)

9. MODEL FOR POPULATION *ibd* AT ONE LOCUS:

- At least, we must consider the genotypes of two individuals, and hence *ibd* among 4 genomes.

- Marginal multigene *ibd* probabilities from Ewens’ sampling formula. (Balding and Nichols, 1994). Model for allelic variation, in which genes descended from the same mutation event are same allele: the mutation defines the *ibd* set of genes.

- In a sample size n, let a_i be the number of *ibd* groups of size i. Then the number of *ibd* groups is $k = \sum a_i$, $n = \sum i a_i$

$$\pi_n(a_1, ..., a_n) \;=\; \frac{n!\beta^{n-k}(1-\beta)^{k-1}}{(1+\beta)(1+2\beta)....(1+(n-2)\beta)} \prod_{j=1}^{n} (j^{a_j} a_j!)^{-1}$$

- Here $\pi_2(a_2 = 1) = P(2 \text{ genes } ibd) = \beta = 1/(1+\theta)$: “Ewens’ θ”.

10. POPULATION *ibd* OVER THE GENOME (Thompson, 2008):

• Generalization to n chromosomes of Leutenegger et al. (2003) model. (Without details, as we will see a better model soon.)

• For a given chromosome, in relation to others, let rate of gain of *ibd* be g and rate of loss be h, as proceed along the chromosome.
Let $g/h = \beta/(1-\beta)$.

• Permitted transitions limited; for $n = 4$, 15×15 'transition rate matrix Q has many 0s

• Allow all transitions by combining with ”random changes” model:

$$Q^{\dagger} = (1-\delta)Q + \delta\alpha(-I + \mathbf{1}\pi')$$

• If Q has the equilibrium dsn π, so does $Q^{\dagger}$. Allowing any transition, but with small probability, will let the data speak. Want to approx real process, but real process is complex.

• Model is still Markov when reduced to 9×9 matrix for genotypic states.

11. INFERRING *ibd* FROM POPULATION DATA
A simulation study: Glazner et al. (2010):

• Simulated descent of founder chromosomes in $\sim$ random mating population of 7000 individuals over 200 generations.

• Output: FGL segments in current individuals over a 2×10^8bp chromosome. Gives many small ($\sim$ 0.5Mbp) segments of *ibd* among current individuals, and larger ones among closer relatives.

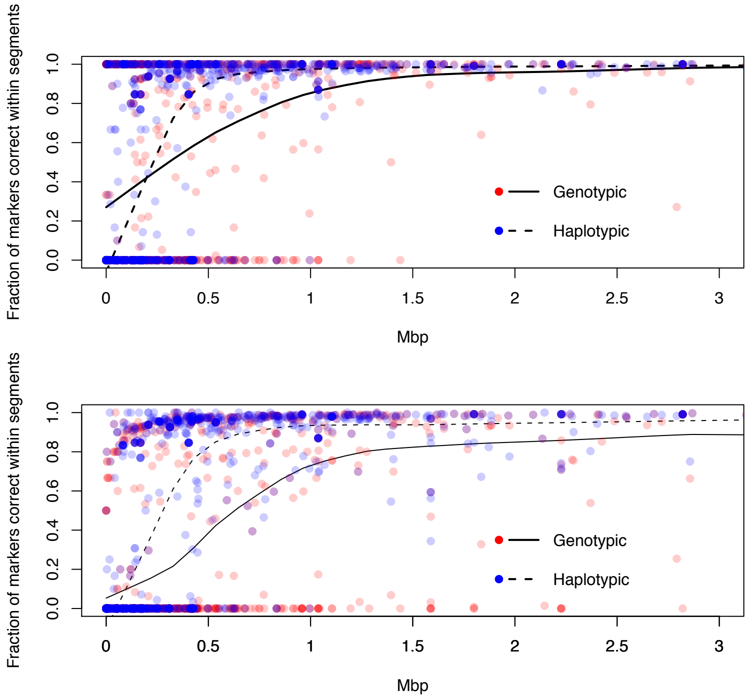
• Real data: 1900 unrelated male X-chromosomes of Framingham Heart Study (FHS). Naturally phased, Good size chromosome.

• Take out SNPs with MAF ≤ 0.05. Take out 3Mbp around centromere. Result: 7000 SNPs over 140 Mbp (avg. 50 per Mbp), real LD, real freq., real locations.

• For sample of current individuals, assign a different random FHS X-chromosome to each FGL appearing in the sample.

• Run the *ibd* estimation program on sets of 4 chromosomes (2 individuals). Output *ibd* probs by SNP; Call criterion ≥ 0.9 for one state.

12. INFERRING *ibd* FROM POPULATION DATA: RESULTS:



10 individuals;
45 pairs.
≈900 *ibd* segs.

By length of true *ibd* segment in simulated population, the proportion of markers within each segment that:
(a) Detect any *ibd* among the 4 haplotypes.
(b) Detect correct state of *ibd*.

ibd among multiple chromosomes

1. MODELS FOR GENOMIC IBD ESTIMATION
2. MODEL FOR POPULATION *ibd* AT ONE LOCUS
3. SPECIFICATION OF *ibd* STATES AS PARTITIONS
4. A MODEL FOR PARTITIONS ALONG A CHROMOSOME
5. *ibd* via COALESCENT OF THE SAMPLE
6. COMPARISON: COALESCENT vs *ibd* PROCESS
7. MCMC OVER PARTITIONS

1. MODELS FOR GENOMIC IBD ESTIMATION:

	Leut.	PLINK	Brow	Thom
Data structure	1 ind	2 ind	2 chr/ind	more!
Phased chromosomes	No	No	Yes/No	Yes
Genotypic data	Yes	Yes	No/Yes	Yes
Error allowed	Yes	No	No/Yes	Yes
Linkage disequilibrium	No	No	Yes	No
Multiple chromosomes jointly	No	No*	No**	Yes

- Leut is Leutenegger et al. (2003); less dense data.
  ibd model and error model of Leut can be extended to multiple genomes. HMM structure can be extended to include LD, but ....
- PLINK is Purcell et al. (2007)
  *Models 2 individuals, but allows only 2,1,0 *ibd* between, not within.
- Brow is Browning (2008) and Browning and Browning (2010).
  **: 2 haplotypes only. Includes LD, but LD reflects coancestry;
  Do we want to condition out this LD ??
- Thom is Thompson (2008) and Thompson (2009)

2. MODEL FOR POPULATION *ibd* AT ONE LOCUS:

- At least, we must consider the genotypes of two individuals, and hence *ibd* among 4 genomes.

- Marginal multigene *ibd* probabilities from Ewens' sampling formula. (Balding and Nichols, 1994). Model for allelic variation, in which genes descended from the same mutation event are same allele: the mutation defines the *ibd* set of genes.

- In a sample size n, let a_i be the number of *ibd* groups of size i.
  Then the number of *ibd* groups is $k = \sum a_i$, $n = \sum i a_i$

$$\pi_n(a_1, ..., a_n) = \frac{n!\beta^{n-k}(1-\beta)^{k-1}}{(1+\beta)(1+2\beta)....(1+(n-2)\beta)} \prod_{j=1}^{n} (j^{a_j} a_j!)^{-1}$$

$$= \frac{\theta^k}{(1+\theta)(2+\theta)...(n-1+\theta)} \prod_{j=1}^{n} (j^{a_j} a_j!)^{-1}$$

- Here $\pi_2(a_2 = 1) = P(2 \text{ genes } ibd) = \beta = 1/(1+\theta)$: "Ewens' θ".

3. SPECIFICATION OF *ibd* STATES AS PARTITIONS:

• For n chromosomes, simpler to specify as partitions:
$Z = \{1, 2, 4\}, \{3, 7\}, \{5, 9, 10\}, \{6\}, \{8\}$

• Since for a given $(a_1, ..., a_n)$, the a_j groups of size j may be permuted and the j elements of each of the a_j groups of size j may be permuted, the number of unordered labeled partitions with given $(a_1, ..., a_n)$ is $n!/\prod_j (j!)^{a_j} a_j!$. Hence the probability of each unordered labeled partition z of the n chromosomes is Ewens (1972)

$$\begin{aligned}
\pi_n(z) &= \pi_n(a_1, ..., a_n) \prod_j (j!)^{a_j} a_j!/n! \\
&= (\Gamma(\theta)\theta^k/\Gamma(\theta + n)) \prod_j \Gamma(j)^{a_j} \\
&= \frac{\theta^k}{(1+\theta)(2+\theta)...(n-1+\theta)} \prod_j ((j-1)!)^{a_j}
\end{aligned}$$

4. A MODEL FOR PARTITIONS ALONG A CHROMOSOME:

• Changing *ibd* along a chromosome: model of Chaozhi Zheng (a version of the Chinese Restaurant Problem Tavare and Ewens (1997)): need to maintain the constant n chromosomes.

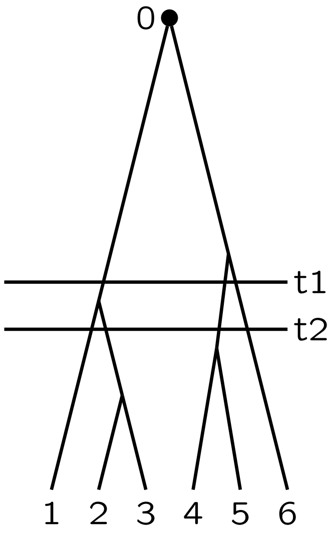
• Changes in *ibd* occur at some rate per bp along the chromosome – a normalized recombination rate ρ.

• First, a *supplementary* chromosome is proposed as a singleton with probability $\theta/(\theta + n)$, and to join each group of size j with probability $j/(\theta + n)$.

• Next, one of the $n + 1$ chromosomes is selected for deletion, and, if not deleted, the supplementary chromosome is given the identity of the deleted chromosome.

• A wider class of transitions.
Maintains the equilibrium distribution.
Remains Markov when reduced to genotypic states.

5. *ibd* via COALESCENT OF THE SAMPLE:



• An alternative (more exact) way to view *ibd* of a sample of chromosomes is through the coalescent Hudson (1991).

• At any point in the genome the coalescent is simply the ancestral tree of the chromosomes, and we can measure *ibd* relative to some past time.

At t1: $\{\{1,2,3\},\{4,5\},\{6\}\}$

At t2 $\{\{1\},\{2,3\},\{4,5\},\{6\}\}$

Varying time, changes the *ibd* groups.

• Along a chromosome, the coalescent changes due to recombination events, and we have the *ancestral recombination graph* (ARG); this has been approximated by a Markov process (McVean and Cardin, 2005).

6. COMPARISON: COALESCENT vs *ibd* PROCESS:

• Given the number of *ibd* groups, the partition among groups in coalescent and Ewens' sampling formulae are the same.

• The number of *ibd* groups, at any past time, has smaller variance than the number given by Ewens' sampling formula; ok as prior?

• Partitions under *ibd* model:

If $\{\{1,4,5\},\{2\},\{3,6,7\},\{8,11,12\},\{9,10\}\}$ is state 1. Then
$\{\{1,4,5\},\{2,3,6,7\},\{8,11\},\{9,10,12\}\}$ is 2-step change,

Either of these steps could occur as changes in the ARG.

• Consider $\{\{1,2,6,7\},\{3,4,9\},\{5,8,10\}\}$ and
$\{\{1,2,6,7\},\{3,4,5,8,9,10\}\}$.

These are only one coalescent event away, but are result of three chromosomes moving (3 steps) in *ibd* process.

• If *ibd* levels low, not so much a problem; for high levels of *ibd* ... ??

7. MCMC OVER PARTITIONS:

• Data are SNP alleles along n haplotypes (phased).

• For larger numbers of chromosomes (e.g. $n = 20$) cannot do HMM computations; far too many *ibd* states!

• Chaozhi Zheng has implemented MCMC over partitions over (small regions) of chromosome (200 SNPs), updating *ibd* state at 5-SNP blocks of SNP positions, conditional on flanking states.

• He uses same error model as Leutenegger et al. (2003): an *ibd* group of chromosomes will have some allelic type. Each chromosome in the group may be observed as of independent allelic type with probability ε.

• Priors on θ $(= (1 - \beta)/\beta)$, ε, and "recombination rate" ρ.

• Update *ibd* partition Z over chromosome, in blocks, given ρ,θ,ε.
Update ρ,θ,ε given Z over chromosome and data.

Pedigrees in populations

1. WHY PEDIGREES?
2. *ibd* IN PEDIGREES
3. ONE LARGE PEDIGREE OR THREE FAMILIES?
4. THE *ibd* GRAPH ON A PEDIGREE: ONE LOCUS
5. CHANGES IN *ibd* GRAPH ALONG A CHROMOSOME
6. *ibd* GRAPHS WITHIN AND BETWEEN FAMILIES
7. MARKER AND TRAIT DATA ON PEDIGREES
8. UKNOWN ANCESTRY IN PEDIGREES
9. *ibd* IN PEDIGREES: MEIOSIS INDICATORS
10. *ibd* IN PEDIGREES: THE MARKOV MODEL
11. *ibd* IN PEDIGREES: THE HMM
12. *ibd* IN PEDIGREES: THE MARKER MODEL
13. THE FACTORED HMM: INDEP MEIOSES
14. *ibd* IN LARGE PEDIGREES: MCMC

1. WHY PEDIGREES?:

• The issue is not pedigrees vs population, but whether the pedigree is known.

*-2mm • A pedgree simply provides a prior for the *ibd* among a set of individuals.

• Compared to our population priors on *ibd*, it is a very informative prior, but a very constraining prior.

• Where the pedigree is known, and individuals observed, we should use it.

• If there are multiple generations unobserved, it is likely more effective to use a population model (even if we think we know the pedigree).

• Pedigrees can provide useful phase information, and so improve population-based *ibd* inferences.

2. *ibd* IN PEDIGREES:

• In a pedigree: *ibd* is well-defined, relative to the founders, and can be inferred by pedigree analysis methods.

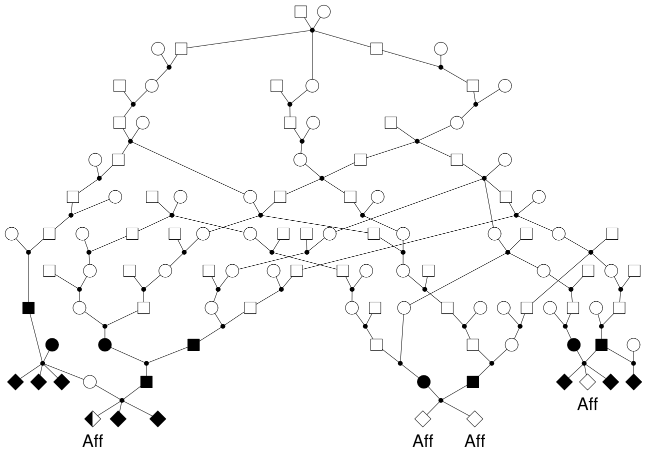
• In a population: *ibd* is defined, relative to some founder population or time-point (??) and can be inferred using a population model for changing *ibd* along a chromosome.

• Whether in pedigrees or populations. allelic similarity is a reflection of *ibd*.

• Whether in a pedigree or a population, (closer) relatives are similar because they have (more) *ibd* genome

• Pedigree-based *ibd* inferred within pedigrees can be combined with population-based *ibd* inferred between pedigrees,

3. ONE LARGE PEDIGREE OR THREE FAMILIES?:



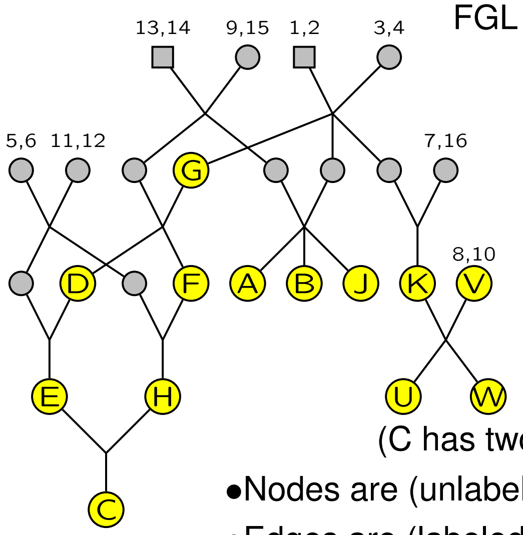
- Details of the ancestral pedigree are surely wrong/biased.

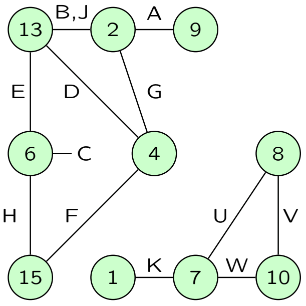
We want to use the *ibd* information, but not the ancestral pedigree.

- 1990s data were insufficient for between-family inference of *ibd*. With modern data, we could infer *ibd* among families

4. THE *ibd* GRAPH ON A PEDIGREE: ONE LOCUS:

FGL = founder genome label.





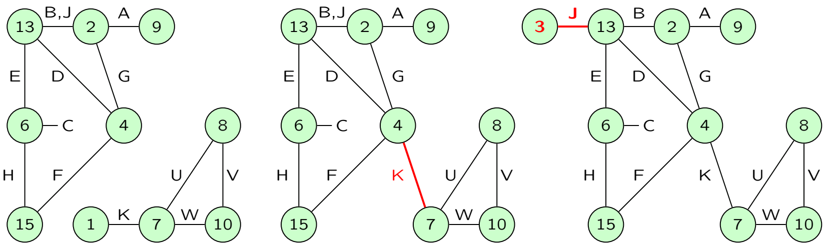
(C has two copies of FGL ”6”)

•Nodes are (unlabeled) *ibd* genome.

•Edges are (labeled) observed individuals.

•Only *ibd* matters, not (labeled) founder origins (FGL), and no longer the pedigree once *ibd* is known/inferred from marker data!

5. CHANGES IN *ibd* GRAPH ALONG A CHROMOSOME:



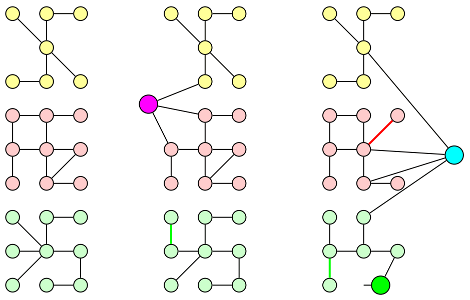
Recomb in meiosis to K. Recomb in meiosis to J.

- Recombination events change the nodes present in observed individuals, and hence the structure of the *ibd* graph. The edges are the same, but may connect different nodes. Nodes may appear/disappear. (Nodes labeled for convenience only.)

- Changes are few (on bp scale); recall in any 1 meiosis, crossovers occurs at $\sim 10^8$ bp, or once per 100 Mbp per meiosis.

- Components of the *ibd* graph tend to be small, when only current generation(s) observed for trait.

6. *ibd* GRAPHS WITHIN AND BETWEEN FAMILIES:



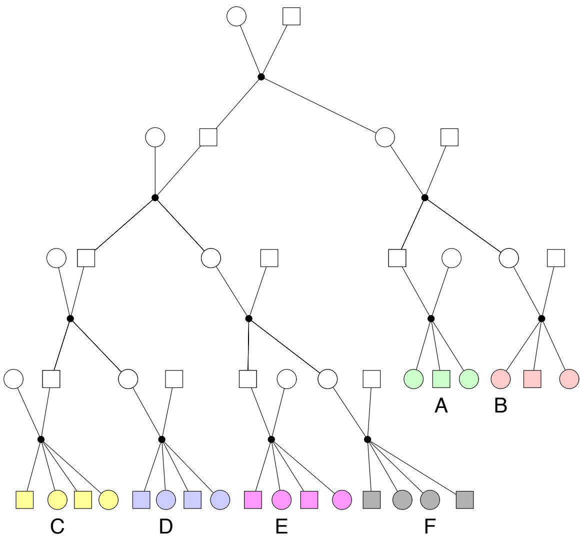
- Within families, recombinations change the gene ibd graph along a chromosome.
- There may be ibd between founders in a given family,
- ... and/or between founders of different families.

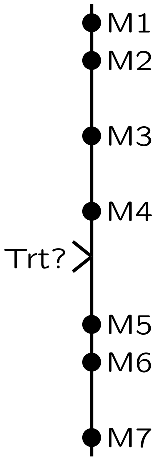
- Generally, such links will be few and sparse, but, with ascertainment, several families might share *ibd* at some points.

- Again components of these graphs are not large/complex.

- Again, the component graphs are slowly varying (on bp scale).

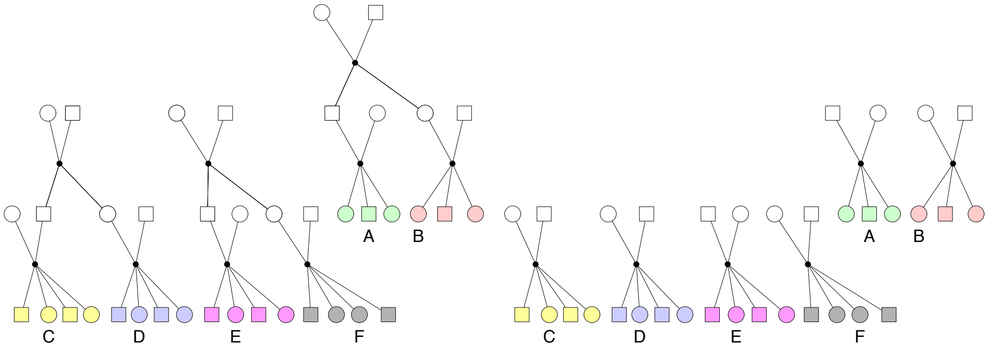
7. MARKER AND TRAIT DATA ON PEDIGREES:





- 22 kids in 6 sibships, observed for markers and trait.
- Markers at known locations. Where is DNA affecting trait?
- Simulate *ibd*, trait data, and marker sets (all with same *ibd*).

8. UKNOWN ANCESTRY IN PEDIGREES:



- Instead of knowing the whole pedigree, we might know only the three pairs of cousinships, or maybe even only the six sibships.
- How much information is lost knowing only the subpedigrees?
- Can we regain lost information by inferring the *ibd* between sub-pedigrees?
- How dense/informative do we need the markers to be to do this?

9. *ibd* IN PEDIGREES: MEIOSIS INDICATORS:

For multiple loci, $j,\ j = 1, \ldots, l$, it is hard to work directly with *ibd*. Instead we define:

$$\begin{aligned} S_{i,j} &= 0 \quad \text{if gene at meiosis } i \text{ locus } j \text{ is parent's maternal} \\ &= 1 \quad \text{if gene at meiosis } i \text{ locus } j \text{ is parent's paternal.} \end{aligned}$$

For convenience, we define

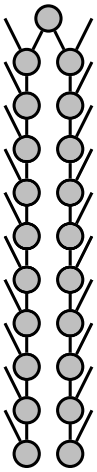
$$\begin{aligned} S_{\bullet,j} &= \{S_{i,j}; i = 1, \ldots, m\},\ \ j = 1, \ldots, l \\ S_{i,\bullet} &= \{S_{i,j}; j = 1, \ldots, l\},\ \ i = 1, \ldots, m \end{aligned}$$

where m is the number of meioses in the pedigree, and l the number of loci along the chromosome.

Dependence of the $\{S_{i,j}\}$
$S_{i,\bullet}$ are independent over i, $i = 1, ..., m$.
$S_{i,j}$ are independent for loci on different chromosome pairs
$S_{\bullet,j}$ are dependent among loci j on the same chromosome pair

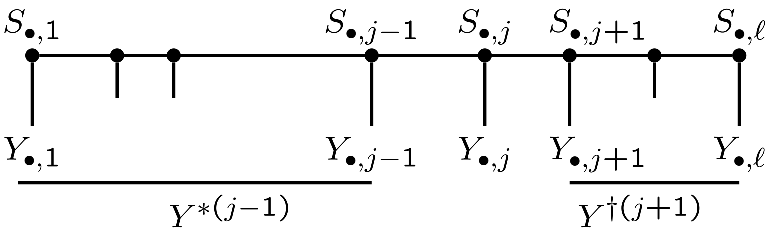
10. *ibd* IN PEDIGREES: THE MARKOV MODEL:

- Note $S_{\bullet,j}$ are Markov (approx); *ibd* is not.



- When we switch from *ibd* we are still ”close” to a configuration that gives *ibd*.
- When we have been non-*ibd* over many markers, likely we have several recombination switches that must get reset to the lineage in order to regain *ibd*.
- *ibd* segments are clustered (Donnelly, 1983).

11. *ibd* IN PEDIGREES: THE HMM:



$$\Pr(\mathbf{S}) = \Pr(S_{\bullet,1}) \prod_{j=2}^{l} \Pr(S_{\bullet,j} \mid S_{\bullet,j-1})$$

$$\Pr(\mathbf{Y} \mid \mathbf{S}) = \prod_{j=1}^{\ell} \Pr(Y_{\bullet,j} \mid S_{\bullet,j}).$$

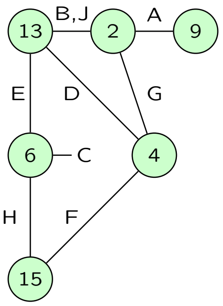
Note that, given $S_{\bullet,j}$,
$Y^{*(j-1)}$, $Y_{\bullet,j}$, and $Y^{\dagger(j+1)}$ are mutually independent.

Also, given $S_{\bullet,j}$, $Y^{*(j-1)}$, $Y_{\bullet,j}$, and $S_{\bullet,j+1}$ are independent.
Also, given $S_{\bullet,j}$, $Y^{\dagger(j+1)}$ $Y_{\bullet,j}$, and $S_{\bullet,j-1}$ are independent.

12. *ibd* IN PEDIGREES: THE MARKER MODEL:

Sobel and Lange (1996); Kruglyak et al. (1996)

- $\Pr(\mathbf{Y} \mid \mathbf{J}(\mathbf{S}))$ is the sum over all possible assignments $\mathcal{A}$ of allelic types to genes of the product of allele frequencies $q_{a(k)}$ of assigned alleles $a(k)$: $\Pr(\mathbf{Y} \mid \mathbf{J}(\mathbf{S})) = \sum_{\mathcal{A}} \prod_k q_{a(k)}$.
- $\Pr(Y_{\bullet,j} \mid \mathbf{S}) = \sum \Pr(\mathcal{A}_j)$: sum over all $\mathcal{A}_j$ consistent with $Y_{\bullet,j}$.



- Suppose $A,\ B,\ J$ are all a_1a_4, G is a_1a_6, D is a_4a_6, E is a_4a_2, C is a_2a_2, F is a_3a_6, and H is a_2a_3.
- Then 2 is a_1; 9, 13 are a_4; 4 is a_6; 6 is a_2; 15 is a_3. The probability is $q_1\ q_2\ q_3\ q_4^2\ q_6$.
- There are always 2, 1, or 0 possible $\mathcal{A}_j$.
- Probabilities multiply over disjoint components.

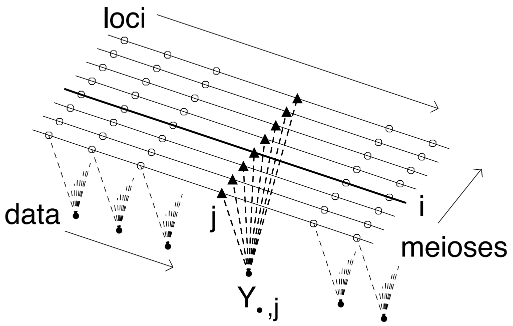
13. THE FACTORED HMM: INDEP MEIOSES:

If there are m meioses on the pedigree, then $S_{\bullet,j}$ can take 2^m values. Computations involve, for each locus, transitions from the 2^m values of $S_{\bullet,j}$ to the 2^m values of $S_{\bullet,j+1}$. Computation is order $L2^{2m}$. For Genehunter, for a pedigree with n individuals, f of whom are founders, $m = 2n - 3f$, and $m \leq 16$. Additionally, for each locus and for each value of $S_{\bullet,j}$, we must compute $\Pr(Y_{\bullet,j} \mid \mathbf{J}(S_{\bullet,j}))$. Although this is easy for given $S_{\bullet,j}$, this limits size of pedigree.

Actually better algorithms using independence of meioses give us a *factored HMM* which means we can get an algorithm of order $mL2^m$ but is is still exponential in pedigree size.

14. *ibd* IN LARGE PEDIGREES: MCMC:

- The independence of meioses $S_{i,\bullet}$ and Markov dependence of inheritance vectors $S_{\bullet,j}$ provide good MCMC block Gibbs samplers:



L-sampler: resample $S_{\bullet,j}$ given $\mathbf{Y}$ and $S_{\bullet,j'}, j \neq j'$ Heath (1997). Uses pedigree peeling.

M- (or MM-) sampler: resample $\{S_{i,\bullet}; i \in I^*\}$ given $\mathbf{Y}$ and $\{S_{i',\bullet}; i' \notin I^*\}$ Uses HMM peeling. Tong and Thompson (2008)

- Computations practical if assume $S_{\bullet,j}$ Markov over chromosome. $\Pr(Y_{\bullet,j} \mid S_{\bullet,j})$ is trivial for markers observed without error.

- L-sampler irreducible; M-sampler mixes with tight linkage: use in combination!!.

Lod scores within and between pedigrees

1. THE LINKAGE LOD SCORE VIA *ibd*
2. MONTE CARLO LOD SCORES
3. LOD SCORES FOR A QUANTITATIVE TRAIT
4. THE LOD SCORES FROM 1000 REALIZATIONS OF $\mathbf{S} \mid \mathbf{Y}_M$
5. UNKNOWN COANCESTRY IN PEDIGREES
6. 1000 LOD SCORES CONTRIBUTIONS ON SUBPEDIGREES
7. PED-CD SWITCHING AT MARKERS 53-62
8. WITHIN- VS BETWEEN-PEDIGREE INFORMATION
9. BETWEEN-PEDIGREE CONTRIBUTIONS TO THE LOD SCORE
10. BETWEEN COUSINSHIP true AND INFERRED inferred *ibd*
11. PUTTING THE *ibd* TOGETHER
12. *ibd* GRAPHS AND LOD SCORES AT MARKER 100
13. *ibd* GRAPHS AND LOD SCORES AT MARKER 160
14. CONCLUSIONS

1. THE LINKAGE LOD SCORE VIA *ibd*:

- The lod score is a tool to map the genes affecting a trait against a know genetic marker map. For trait data $\mathbf{Y}_T$ and marker data $\mathbf{Y}_M$

$$\text{lod} = \log_{10} \frac{\Pr(\mathbf{Y}_T, \mathbf{Y}_M; \Gamma)}{\Pr(\mathbf{Y}_T, \mathbf{Y}_M; \Gamma_0)}$$

where Γ_0 is Γ with no T/M linkage

$$= \log_{10} \frac{\Pr(\mathbf{Y}_T, \mathbf{Y}_M; \Gamma)}{\Pr(\mathbf{Y}_T; \Gamma)\Pr(\mathbf{Y}_M; \Gamma)} = \log_{10} \frac{\Pr(\mathbf{Y}_T \mid \mathbf{Y}_M; \Gamma)}{\Pr(\mathbf{Y}_T; \Gamma)}$$

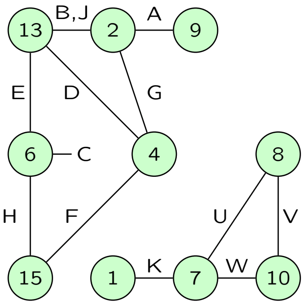
- Compute by summing over *ibd* pattern among individuals observed for trait at each hypothesized trait location:

$$\Pr(\mathbf{Y}_T \mid \mathbf{Y}_M; \Gamma) = \sum_{ibd_j} \Pr(\mathbf{Y}_T \mid ibd_j; \Gamma_T)\Pr(ibd_j \mid \mathbf{Y}_M; \Gamma_M)$$

- Note *ibd* is inferred at location (or locations) hypothesized as affecting the trait, but conditional on marker data jointly at all locations.

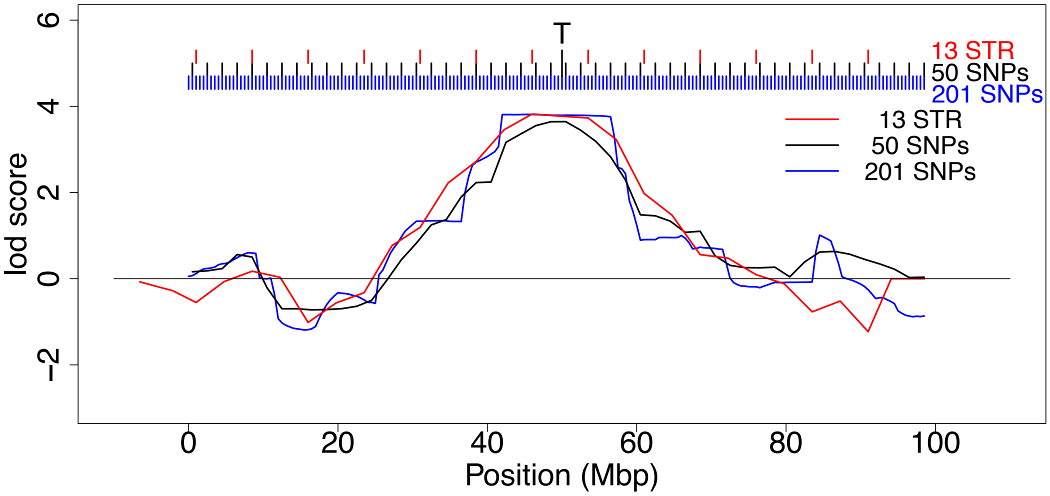
2. MONTE CARLO LOD SCORES:

- Exact computation infeasible: instead, sample multiple realizations of *ibd* across all locations j, given all marker data $\mathbf{Y}_M$.
- Estimate of $\Pr(\mathbf{Y}_T \mid \mathbf{Y}_M; \Gamma)$ is given by averaging $\Pr(\mathbf{Y}_T \mid$ *ibd*$)$ over the sampled realizations of *ibd*.



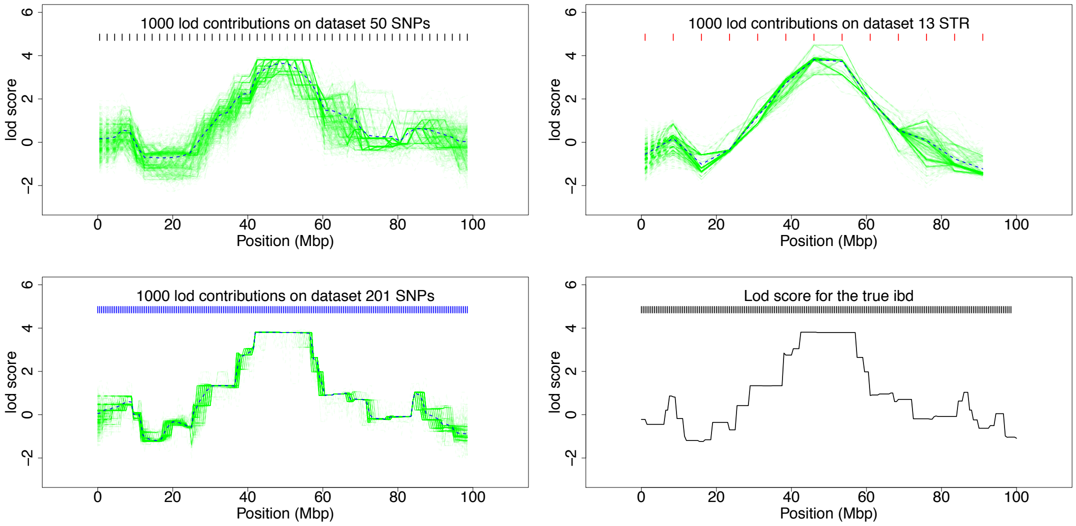
- Computation of $\Pr(\mathbf{Y}_T \mid$ *ibd*$)$ is simple using the *ibd*-graph.
- Components of *ibd* graphs are small, relative to pedigrees.
- In fact, we can compute $\Pr(\mathbf{Y}_T \mid$ *ibd*$)$ over joint *ibd*-graphs for several genome locations– complex trait models.

3. LOD SCORES FOR A QUANTITATIVE TRAIT:



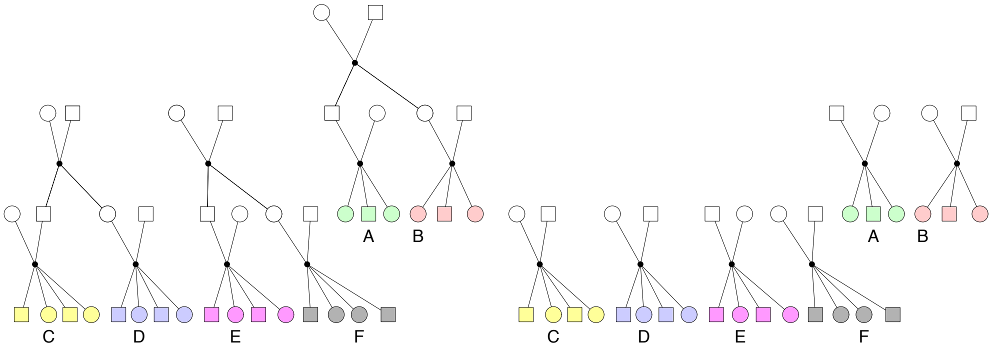
- Each based on 1000 realizations of chromosomal *ibd* sampled at spacing 30 MCMC scans.

4. THE LOD SCORES FROM 1000 REALIZATIONS OF $\mathbf{S} \mid \mathbf{Y}_M$:



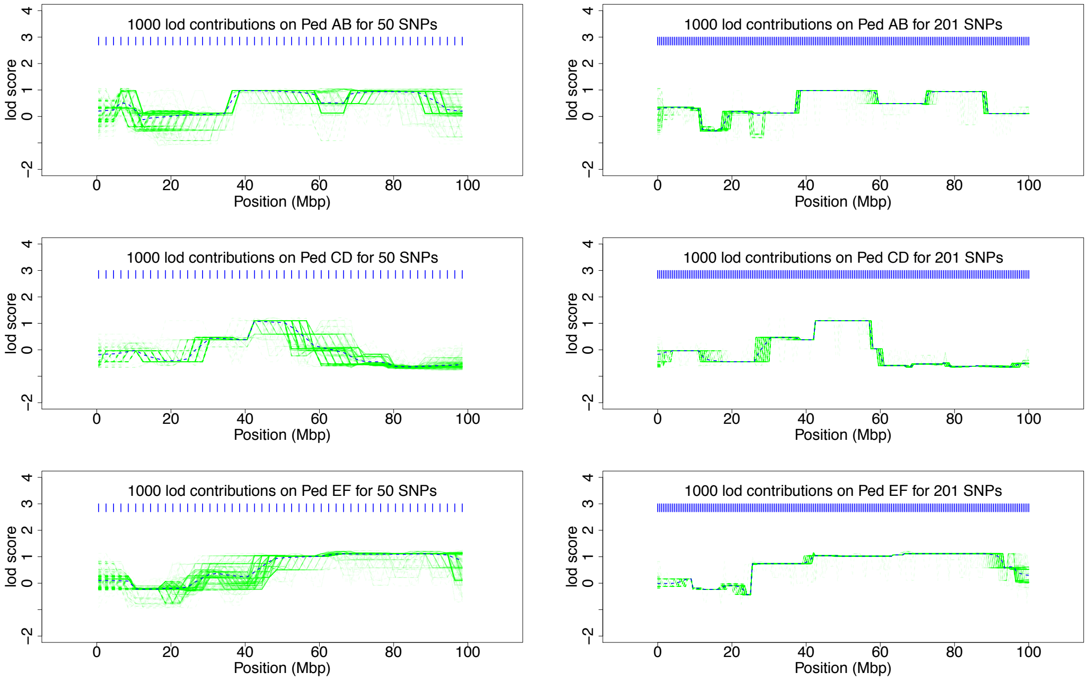
- 50 SNPs; not enough precision
- 13 STR; not enough resolution
- 201 SNPs still some uncertainty, but close to ”true” lod score.

5. UNKNOWN COANCESTRY IN PEDIGREES:



- Instead of knowing the whole pedigree, we might know only the three pairs of cousinships, or maybe even only the six sibships.

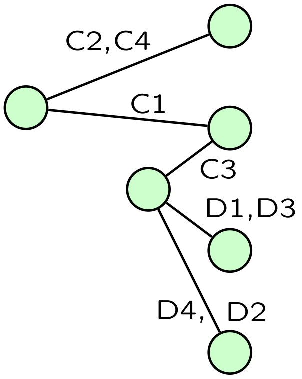
6. 1000 LOD SCORES CONTRIBUTIONS ON SUBPEDIGREES:



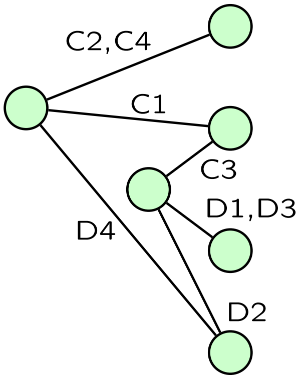
7. PED-CD SWITCHING AT MARKERS 53-62:

Marker 34 to 58.

Marker 59 to 80

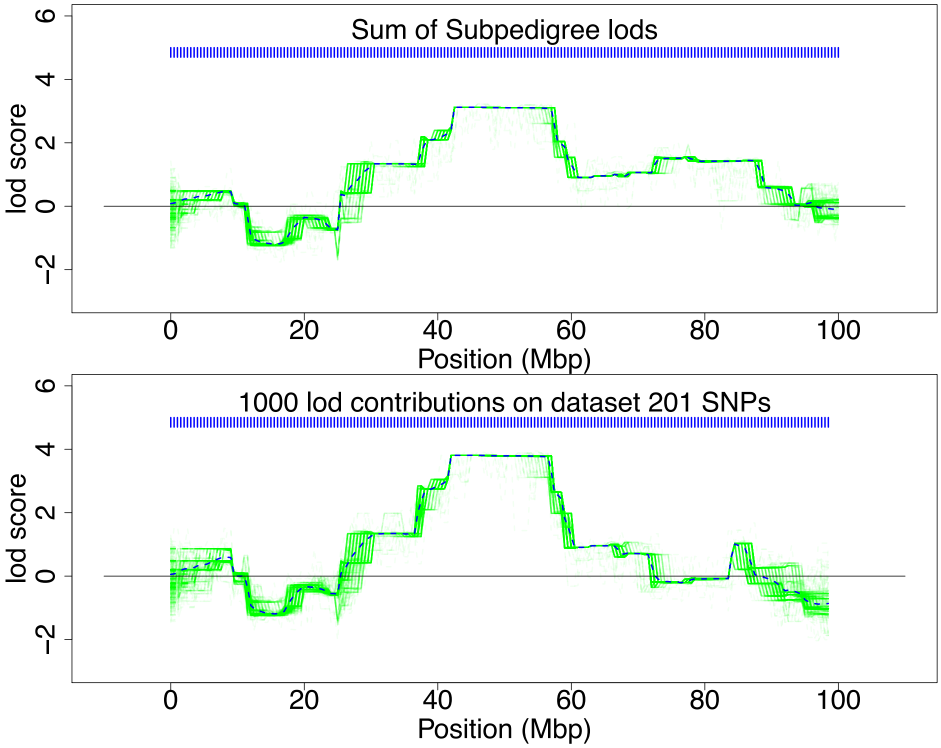


Single switch in D4 causes lod score change from -0.45 to +0.47; D4 has low trait value, as do cousins C1, and lowish C2, C4. But C3, D2, and D3 have high values. D1 intermediate.



- Left *ibd* is clear to marker 52. Right *ibd* is clear from marker 62.
- SNP markers 53 to 61 uninformative about this *ibd*

8. WITHIN- VS BETWEEN-PEDIGREE INFORMATION:



9. BETWEEN-PEDIGREE CONTRIBUTIONS TO THE LOD SCORE:

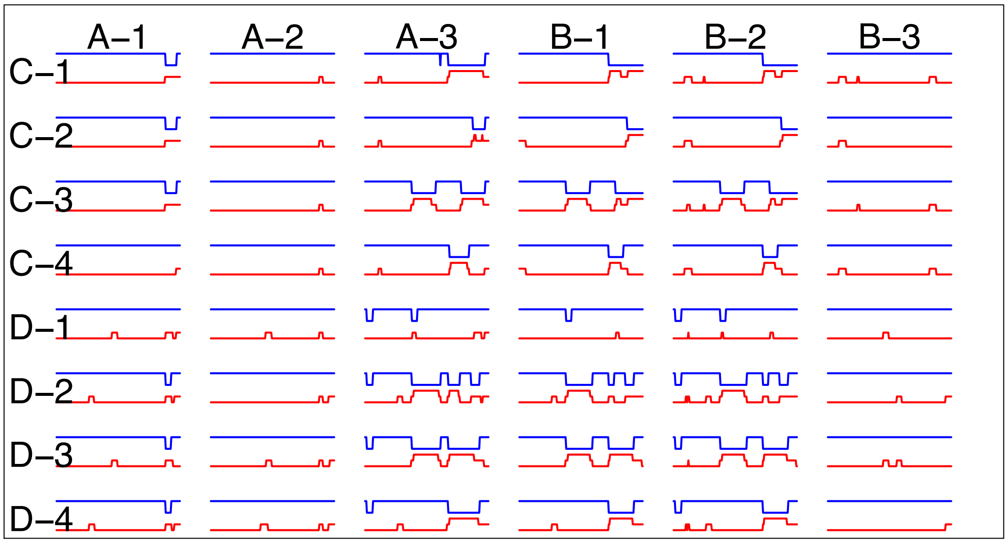
Consider 4 positions at which there is almost no uncertainty in *ibd*:

- SNPs 65, 100, 125, 160
- Positions 32, 50, 62, 80 Mbp.

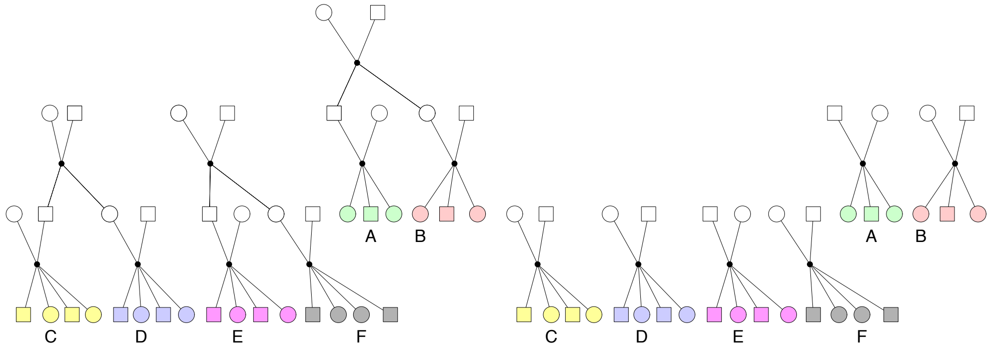
marker	Overall	Sum of cousinships	
SNP-65	1.342	1.336	$\sim$ no *ibd*
SNP-100	3.793	3.103	*ibd* concordant with trait
SNP-125	0.908	0.907	$\sim$ no *ibd*
SNP-160	-0.089	1.416	*ibd* discordant with trait

10. BETWEEN COUSINSHIP true AND INFERRED inferred *ibd*:

- Using the 201-SNP marker data to infer *ibd* between cousinships.
- Require a 0.9 probability of *ibd* state to call the *ibd*.

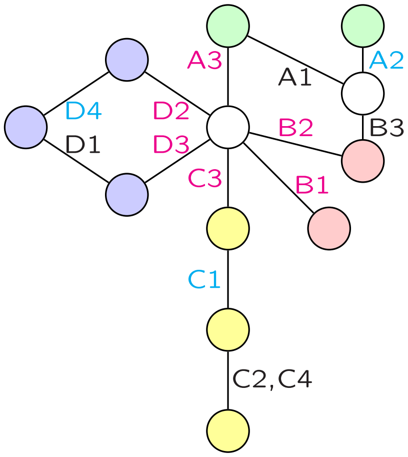


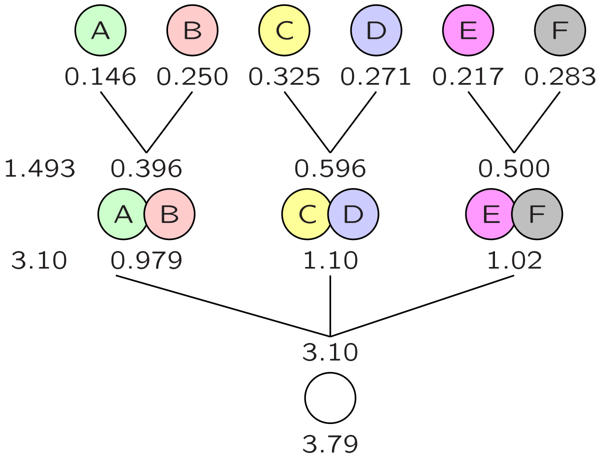
11. PUTTING THE *ibd* TOGETHER:



marker	Overall Pedigree	Sum of 3 cousinships	Sum of 6 sibships	Combined using inferred *ibd*
SNP-65	1.342	1.336	0.5774	1.340
SNP-100	3.793	3.103	1.4934	3.794
SNP-125	0.908	0.907	0.8603	0.911
SNP-160	-0.089	1.416	0.2425	-0.080

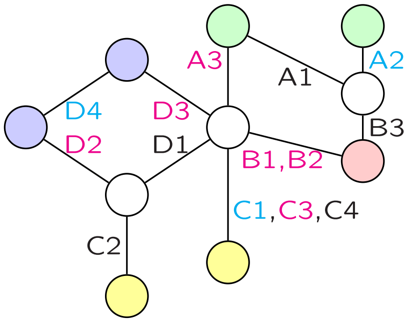
12. *ibd* GRAPHS AND LOD SCORES AT MARKER 100:

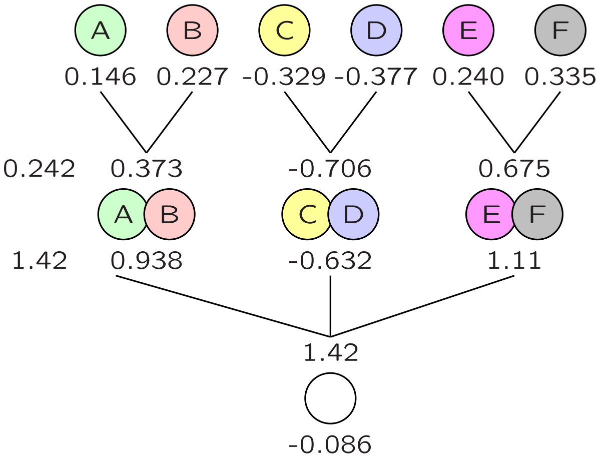




e.g. A3 is kid-3 in sibship A
high/normal/low trait values

13. *ibd* GRAPHS AND LOD SCORES AT MARKER 160:





e.g. A3 is kid-3 in sibship A
high/normal/low trait values

14. CONCLUSIONS:

• In analyzing the genetics of a trait only the *ibd* matters.

• *ibd* is *ibd* whether in pedigrees or in populations.

• *ibd* in pedigrees and in populations can be inferred from marker data.

• SNPs at a density of 0.5 Mbp leave little uncertainty in *ibd* in pedigrees, and permit inference of *ibd* between pedigrees.

• For more remote coancestry (smaller *ibd* segments) we need more SNPs (e.g. 50 per Mbp), but still 10 times less than what exists.

• *ibd* in populations is NOT a nuisance (cf. association studies).

• Inferred *ibd* can be used to give increased power for linkage detection, and increased resolution of causal loci.

References

Abecasis GR, Cardon L, Cookson WOC (2000) A general test of association for quantitative traits in nuclear families. American Journal of Human Genetics 66:279–292

Albrechtsen A, Korneliussen TS, Moltke I, van Overseem Hansen T, Nielsen FC, Nielsen R (2009) Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. Genetic Epidemiology 33:266–274

Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Research 19:1655–1664

Allison DB (1997) Transmission-disequilibrium for quantitative traits. American Journal of Human Genetics 60:676–690. (See also erratum: P.1571)

Allison DB, Heo M, Kaplan N, Martin ER (1999) Sibling-based tests of linkage and association for quantitative trait. American Journal of Human Genetics 64:1754–1763

Balding DJ, Nichols RA (1994) DNA profile match probability calculations: How to allow for population stratification, relatedness, database selection, and single bands. Forensic Science Int 64:125–140

Baum LE (1972) An inequality and associated maximization technique in statistical estimation for probabilistic functions on Markov processes. In O Shisha, ed., *Inequalities-III; Proceedings of the Third Symposium on Inequalities. University of California Los Angeles, 1969*, 1–8. Academic Press, New York

Browning SR (2006) Multilocus association mapping using variable-length Markov chains. American Journal of Human Genetics 78:903–913

— (2008) Estimation of pairwise identity by descent from dense genetic marker data in a population sample of haplotypes. Genetics 178:2123–2132

Browning SR, Browning BL (2010) High-resolution detection of identity by descent in unrelated individuals. American Journal of Human Genetics 86:526–539

Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55:997–1004

Donnelly KP (1983) The probability that related individuals share some section of genome identical by descent. Theoretical Population Biology 23:34–63

Ewens WJ (1972) The sampling theory of selectively neutral alleles. Theoretical Population Biology 3:87–112

Fisher RA (1954) A fuller theory of junctions in inbreeding. Heredity 8:187–197

Fu AQ, Thompson EA (2007) Inference of Identity-by-Descent in Sib Pairs: Analysis with and without Linkage Disequilibrium. Technical report # 519, Department of Statistics, University of Washington

Glazner C, Brown MD, Cai Z, Thompson EA (2010) Inferring coancestry in structured populations. Abstract, West North American Region of the IBS Annual Meeting

Graham J, Thompson EA (1998) Disequilibrium likelihoods for fine-scale mapping of a rare allele. American Journal of Human Genetics 63:1517–1530

Haseman JK, Elston RC (1972) The investigation of linkage between a quantitative trait and a marker locus. Behavior Genetics 2:3–19

Heath SC (1997) Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. American Journal of Human Genetics 61:748–760

Hudson R (1991) Gene genealogies and the coalescent process. In R Dawkins, M Ridley, eds., *Oxford Surveys in Evolutionary Biology*, vol. 7, 1–44. Oxford University Press: Oxford

Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES (1996) Parametric and nonparametric linkage analysis: A unified multipoint approach. American Journal of Human Genetics 58:1347–1363

Laurie CC, Doheny KF, Mirel DB, Pugh EW, Bierut L, Bhangale T, Boehm F, Caporaso N, Cornelis MC, Edenberg HJ, Gabriel SB, Harris EL, Hu FB, Jacobs K, Kraft P, Landi MT, Lumley T, Manolio TA, McHugh C, Painter I, Paschall J, Rice JP, Rice K, Zheng X, Weir B (2010) Quality control and quality assurance in genotypic data for genome-wide association studies. Genetic Epidemiology 34:591–603

Leutenegger A, Labalme A, Génin E, Toutain A, Steichen E, Clerget-Darpoux F, Edery P (2006) Using genomic inbreeding coefficient estimates for homozygosity mapping of rare recessive traits: Application to Taybi-Linder Syndrome. American Journal of Human Genetics 79:62–66

Leutenegger A, Prum B, Genin E, Verny C, Clerget-Darpoux F, Thompson EA (2003) Estimation of the inbreeding coefficient through use of genomic data. American Journal of Human Genetics 73:516–523

Li N, Stephens M (2003) Modelling linkage disequilibrium, and identifying recombination hotspots using SNP data. Genetics 165:2213–2233

McKeigue P (2005) Prospects for Admixture Mapping of Complex Traits: a review. American Journal of Human Genetics 76:1–7

McPeek MS, Strahs A (1999) Assessment of linkage diseqilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. American Journal of Human Genetics 65:858–875

McVean G, Cardin N (2005) Approximating the coalescent with recombination. Philosophical Transactions of the Royal Society of London Series B 360:1387–1393

Patterson N, Hattangadi N, Lane B, Lohmueller KE, Hafler DA, Oksenberg JR, Hauser SL, Smith MW, OBrien SJ, Altshuler D, Daly MJ, Reich D (2004) Methods for High-Density Admixture Mapping of Disease Genes. American Journal of Human Genetics 74:79–1000

Price AL, Butler J, Patterson N, Capelli C, Pascali VL, Scarnicci F, Ruiz-Linares A, Groop L, Saetta AA, Korkolopoulou P, Seligsohn U, Waliszewska A, Schirmer C, Ardlie K, Ramos A, Nemesh J, Arbeitman L, Goldstein DB, Reich D, Hirschhorn JN (2008) Discerning the ancestry of European Americans in genetic association studies. PLoS Genetics 4:e236

Pritchard JK, Donnelly P (2001) Case-control studies of association in structured or admixed populations. Theoretical Population Biology 60:227–237

Pritchard JK, Rosenberg NA (1999) Use of unlinked genetic markers to detect population stratification in association studies. American Journal of Human Genetics 65:220–228

Pritchard JK, Stephens M, Donnelly PJ (2000a) Inference of population structure using multilocus genotype data. Genetics 155:945–959

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000b) Association mapping in structured populations. American Journal of Human Genetics 67:170–181

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, et al. (2007) PLINK: a tool-set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics 81:559–575

Rabinowitz D (1997) A transmission disequilibrium test for quantitative trait loci. Human Heredity 47:342–350

Rosenberg NA, Li LM, Ward R, Pritchard JK (2003) Informativeness of genetic markers for inference of ancestry. American Journal of Human Genetics 73:1402–1422

Smith MW, Patterson N, Lautenberger JA, Truelove AL, McDonald GJ, Waliszewska A, Kessing BD, et al (2004) A high-density admixture map for disease gene discovery in African Americans. American Journal of Human Genetics 74:1001–1013

Sobel E, Lange K (1996) Descent graphs in pedigree analysis: Applications to haplotyping, location scores, and marker-sharing statistics. American Journal of Human Genetics 58:1323–1337

Spielman RS, Ewens WJ (1996) The TDT and other family-based tests for linkage disequilibrium. American Journal of Human Genetics 59:983–989

Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus. American Journal of Human Genetics 52:506–516

Storey JD (2002) A direct approach to false discovery rates. Journal of the Royal Statistical Society Series B 64:479–498

Tavare S, Ewens WJ (1997) Multivariate Ewens Distribution. In *Discrete Multivariate Distributions*, 232–246. Wiley, New York, NY

Thompson EA (2008) The IBD process along four chromosomes. Theoretical Population Biology 73:369–373

— (2009) Inferring coancestry of genome segments in populations. In *Invited Proceedings of the 57th Session of the International Statistical Institute*, IPM13: Paper 0325.pdf. Durban, South Africa

Tong L, Thompson EA (2008) Multilocus lod scores in large pedigrees: Combination of exact and approximate calculations. Human Heredity 65:142–153

Visscher PM, Andrew T, Nyholt DR (2008) Genome-wide association studies of quantitative traits with related individuals: little (power) lost but much to be gained. European Journal of Human Genetics 16:387–390

Wright S (1951) The genetical structure of populations. Annals of Eugenics 15:323–354

Zöllner S, Pritchard JK (2005) Coalescent-based association mapping and fine mapping of complex trait loci. Genetics 169:1071–1092

Available software and documentation

1. PLINK
2. BEAGLE
3. *ibd* INFERENCE WITH BEAGLE
4. MORGAN 2.9
5. MORGAN 3.0.1 with IBD_HAPLO
6. UPCOMING: IBDgraph and IBD_Merge

1. PLINK:

- PLINK Download site
http://pngu.mgh.harvard.edu/ purcell/plink/gplink.shtml

- PLINK documentation link
http://pngu.mgh.harvard.edu/ purcell/plink/dist/plink-doc-1.07.pdf

- Problem with PLINK is that source is not available: what is implemented is not always what is in published papers (e.g. for *ibd* inference using pairs of individuals).

2. BEAGLE:

- BEAGLE download site:
http://faculty.washington.edu/browning/beagle/beagle.html

- BEAGLE documentation:
http://faculty.washington.edu/browning/beagle/beagle_3.326Dec10.pdf

- Great package, well documented, but again no source.

3. *ibd* INFERENCE WITH BEAGLE:

● This information was supplied by Brian Browning.

● Example command:

```
java -Xmx1000m -jar beagle.jar unphased=ibd.region.bgl
markers=ibd.region.markers ibdpairs=ibd.region.ibd.pairs missing=N
gprobs=false out=ibd verbose=true
```

● To generate IBD probabilities,
1) You must include a markers file (specified with the with genetic distances in cM positions (markers file format is described in the Beagle docs).
2) The .bgl file must contain a sample identifier line ("I id ...").
3) You must include the ibdpairs= argument.
4) The specified ibd pairs file must contain two white-space-delimited sample identifiers per line.

4. MORGAN 2.9:

● Main software page:
http://www.stat.washington.edu/thompson/Genepi/pangaea.shtml

● MORGAN 2.9 download link
.../Genepi/MORGAN/Morgan.shtml

● MORGAN 2.9 tutorial and examples
Download tutorial /examples:
.../Genepi/MORGAN/Morgan.shtml#tut
Online tutorial with link to examples file:
.../Genepi/MORGAN/morgan-tut-html-v29/morgan-tut.html

● With MORGAN source code is freely available.

5. MORGAN 3.0.1 with IBD_HAPLO:

- MORGAN 3.0.1 download site
http://www.stat.washington.edu/thompson/Genepi/MORGAN/Morgan.shtml

- MORGAN 3 changes; documented, but no tutorial. (tlocs etc.) MORGAN-3 is MUCH better.

- MORGAN-3 includes IBD_HAPLO.
.../thompson/Genepi/MORGAN/ibd_haplo.tar.gz

- IBD_Haplo README:
.../thompson/Genepi/MORGAN/README_ibdhap

6. UPCOMING: IBDgraph and IBD_Merge:

- IBDgraph download site and README file
See main software page:
http://www.stat.washington.edu/thompson/Genepi/pangaea.shtml

- IBDgraph examples,

- IBD_Merge – not released yet, but ...