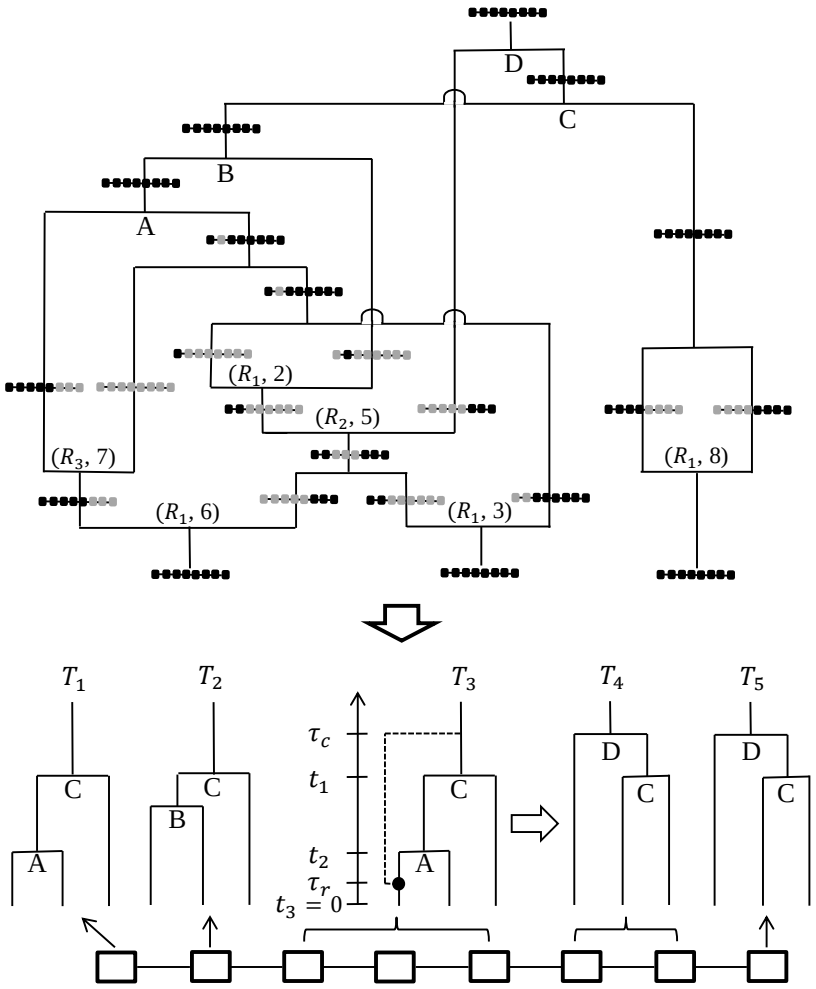
Models for the use and inference of identity-by-descent in populations

Elizabeth Thompson

For: SMEEG Conference, Angers, France
9-12 December, 2013

With acknowledgement to
Sharon Browning, Chaozhi Zheng, and Chris Glazner.

A model too complex to use



- Full specification of ancestry is the *ancestral recombination graph* or ARG: Figure due to Chaozhi Zheng.
- MCMC sampling of the ARG (Kuhner et al.) or of its sequential Markov approximations, (Zheng et al.) is hard (even for 100 kbp).
- Main problem: Our interest is in long lengths (> 1 Mbp) and short time depths < 50 generations. Most of the ARG is irrelevant.

Figure 1

Genetic variation, Association, and Descent

• There is a large amount of variation in our genomes: at about 1 in 1000 bp, there will be two different possible alleles (a and b). These are SNPs; *single nucleotide polymorphisms*.

• The data are genetic marker (SNP) data $\mathbf{X}$ at known locations in the genome, and trait data $\mathbf{Y}$ (qualitative or quantitative).

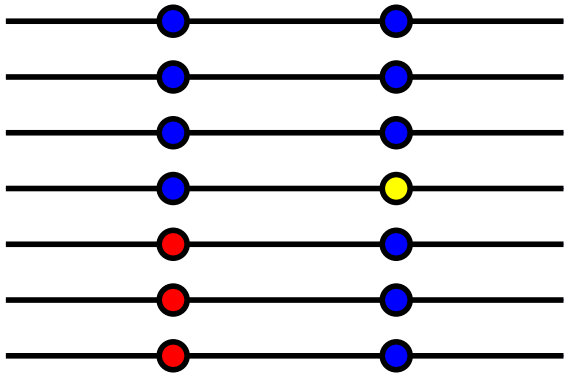
• The goal is to find where in the genome are there DNA variants that affect the trait values $\mathbf{Y}$.

• Direct testing for an association between $\mathbf{Y}$ and allelic type $\mathbf{X}$ at each SNP location ignores the fact that DNA descends in blocks.

• Also ignores the fact that functional genes are blocks of DNA and is confounded by allelic heterogeneity: many ways to mess up a local block of DNA that is a functional gene.

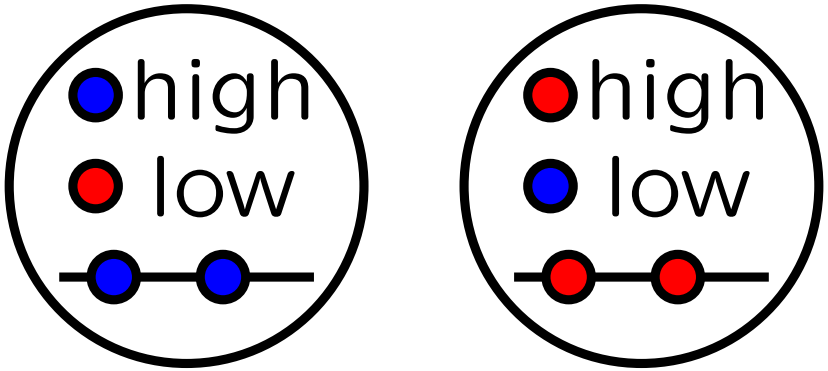
• Instead consider association in descent of $\mathbf{X}$ and $\mathbf{Y}$:
DNA is identical by descent (*ibd*) if it is a copy of the same DNA in a common ancestor.

Relatedness is the source of allelic association



- A new causal mutation, ●, arises.
- Associations of interest come from descent of small chromosome segments over many generations.
- The association is maintained by genetic linkage.

- Associations also arise from demographic history and random genetic drift, resulting in differing allele frequencies among population subdivisions.



- Both are forms of relatedness; the first can signal a causal location.
- Idea of *ibd*-based mapping is to detect excess location-specific relatedness (identity by descent, *ibd*) $\mathbf{Z}$ at test locations, among individuals of similar phenotype.

Case-control study: Excess relatedness among cases

• In association tests, we compare frequency of an allele in N_1 cases vs N_2 controls, at dense test SNP locations across the genome:

$$\left(\frac{1}{2N_1} \sum_{\text{cases}} X_i \; - \; \frac{1}{2N_2} \sum_{\text{cont.}} X_i \right)$$

where $X_i = 0, 1, 2$ is number of alleles of specified type in i.

• In *ibd* test, we compare the frequency of *ibd* between M_1 case-case pairs and M_2 case-(non-case) or (non-case)-(non-case) pairs:

$$\left(\frac{1}{M_1} \sum_{\text{case-case}} Z_i \; - \; \frac{1}{M_2} \sum_{\text{other}} Z_i \right)$$

where $Z_i = 1$ or 0 as pair does/does not share genome by descent at test location.

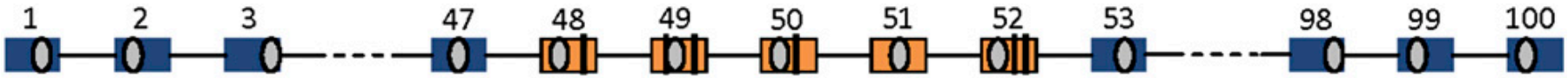
• To adjust for population heterogeneity or structure, adjust for the genome-wide average in each group.

• Assess significance by permutation of case-control labels. (No distributional assumptions.)

Simulation Study: is there enough power?

- Study by Sharon Browning.

- Long population evolutionary simulation at $N_e = 10^4$ with mutation, selection and recombination. Then run forward at larger population ($N_e = 10^5$) for $G = 25$ generations.



- Each simulation is a 200kb region, with central 10kb containing also causal SNPs arising in the population simulation.

- Retain 100 common SNPs; best in alternating 1kb blocks. These are used for association mapping.

- Individuals with ≥ 1 causal variant alleles in the 5 central 1kb blocks are cases with prob 0.1

- Note the *ibd* (location-specific relatedness), $\mathbf{Z}$, is assumed known.

Results of the simulation study

- Results from Browning and Thompson, Genetics, April 2012.

- Properties of simulated causal variants.

selec -tion	# var.	var.freq.	total freq. of var-hap.	max assoc R^2 w/marker SNP
0.0005	11-16	0.00015-0.0060	0.045-0.13	0.91-1.00
0.001	9-14	0.00010-0.0031	0.019-0.050	0.28-1.00
0.002	8-13	0.00010-0.0020	0.0097-0.031	0.06-0.52
0.005	7-10	0.000088-0.001	0.0045-0.011	0.03-0.16

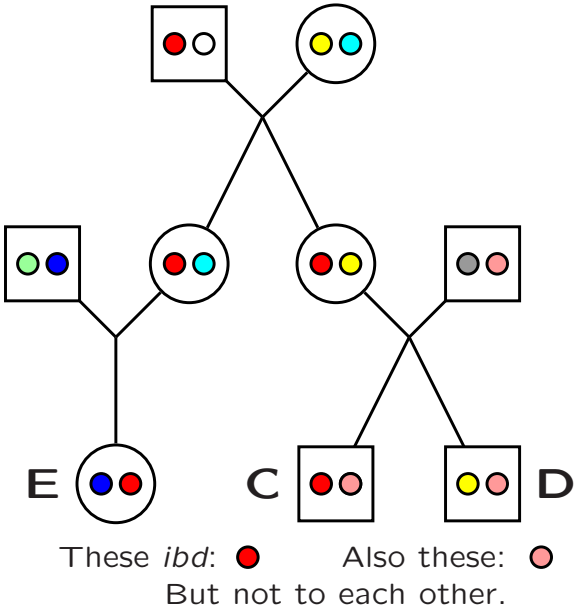
- Power of tests in large population: $N_e = 10^5$ for 25 generations.

selec -tion	# cases= # controls	power assoc.	power *ibd*	association vs. *ibd*
0.0005	500	0.87	0.57	assoc.
0.001	500	0.65	0.53	Not-Sig
0.002	1000	0.53	0.87	*ibd*
0.005	3000	0.47	0.90	*ibd*

Mendelian segregation: Identity by descent

• Mendel's first law (1866); Each individual has two genome copies; one maternal, one paternal. At every location, to each offspring independently, a parent copies a random one of his/her two copies.

• DNA is identical by descent (*ibd*) if it is a copy of the same DNA in a common ancestor.



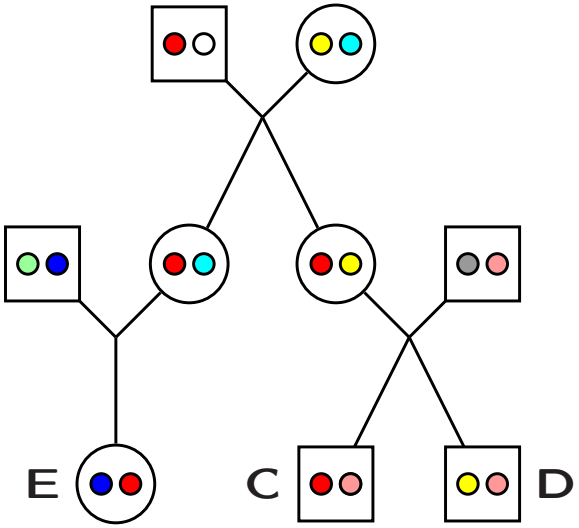
• DNA that is *ibd* is (with high probability) the same allelic type, whereas non-*ibd* DNA is of independent allelic type.

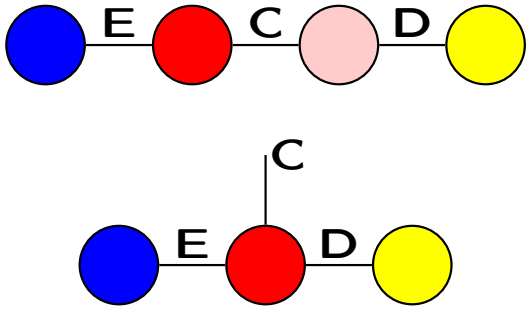
• Whether or not pedigree relationships are known, *ibd* underlies patterns of phenotypic similarity among relatives.

• In a pedigree: *ibd* relative to the founders may be inferred given marker data $\mathbf{X}$ and pedigree prior.

In a population: *ibd* at a locus may be inferred from local marker data haplotypes $\mathbf{X}$ (e.g. • and •).

Given *ibd*, we know all there is to know





●	●	●	●	Prob	
b	a	a	b	$q_a^2 q_b^2$	h
b	a	–	b	$q_a q_b^2$	1-h

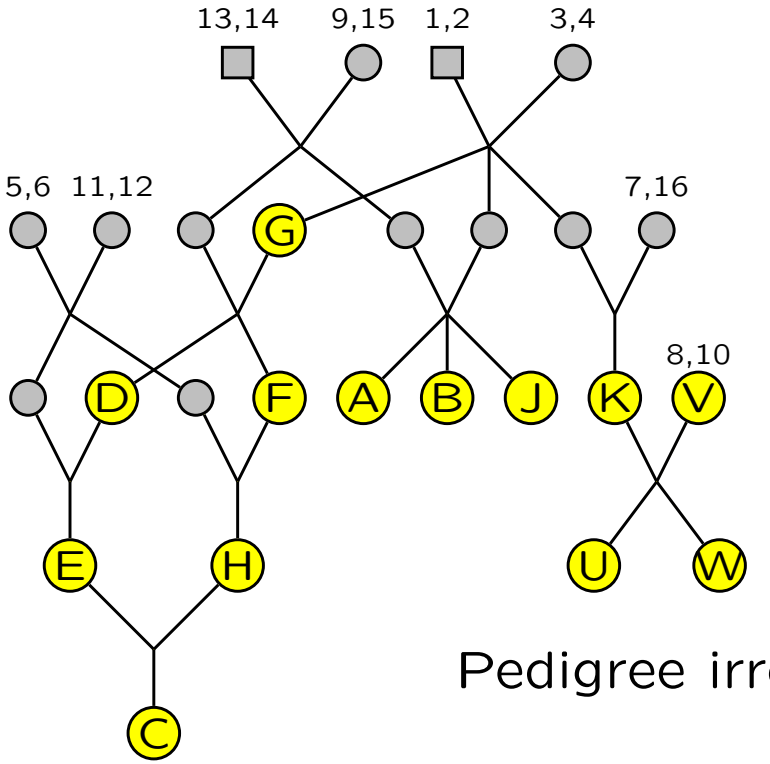
- For example:

  $\Pr(E = ab,\ C = aa,\ D = ab)$

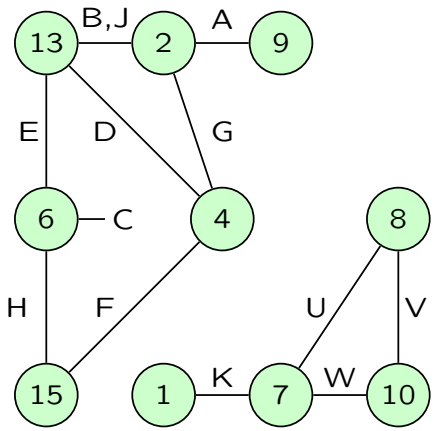
- Or $\Pr(Y_E, Y_C, Y_D) = \sum_{●} \sum_{●} (\Pr(Y_E|●,●) q(●) q(●)$
  $\sum_{●} (\Pr(Y_C|●,●) q(●) \sum_{●} (\Pr(Y_D|●,●) q(●)))$

- Given *ibd*, the pedigree is no longer relevant.
  The *ibd* may come from a pedigree or population inference.
  A population probability model is needed to provide h.

Multiple *ibd* among closely related individuals



FGL = founder genome label.



Pedigree irrelevant once *ibd* is known.

• Edges are observed individuals; nodes represent *ibd* genome.
For example: G, D, and F carry FGL "4": B, J, E and D carry "13".

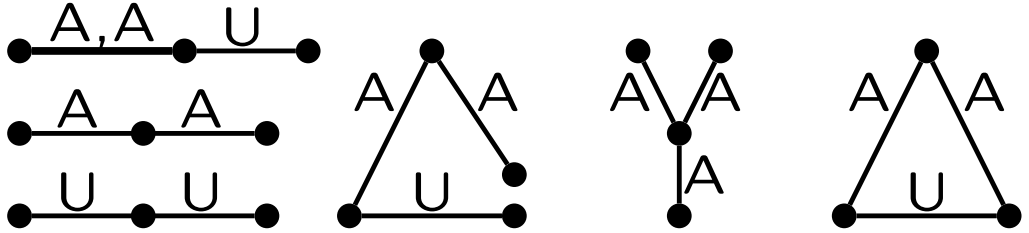
• The *ibd* state at a locus is a partition of the gametes of observed individuals: $(\{Ap\}, \{Am, Bm, Jm, Gp\}, \{Gm, Dm, Fm\}, \{Cp, Cm, Ep, Hp\}, \{Bp, Jp, Dp, Em\}, \{Hm, Fp\}, \{Kp\}, \{Km, Up, Wp\}, \{Um, Vp\}, \{Wm, Vm\})$.

Trait-related *ibd* in population samples

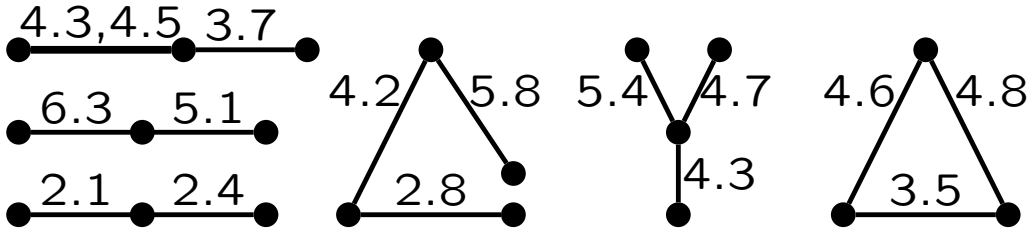
• In a population, *ibd* levels may be lower, and partitions simpler, but trait-related *ibd* can still indicate causal locations.

• Edges are individuals observed for a trait. Two edges sharing a node indicate *ibd* of those individuals at that locus.

(a)



(b)



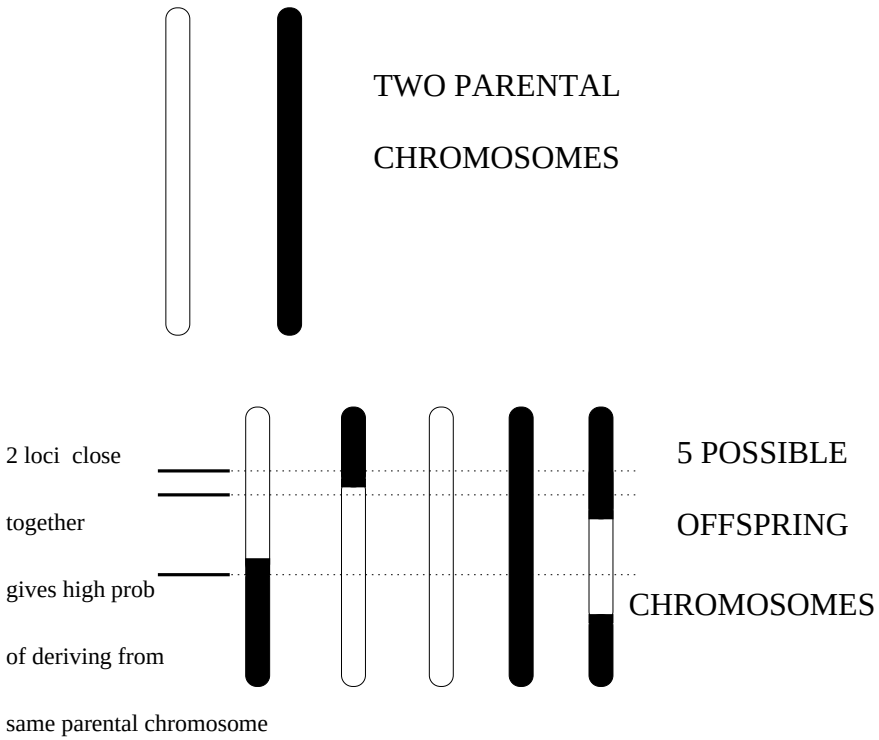
• Trait data may be (a) qualitative, or (b) quantitative.

• Individuals not showing any *ibd* are omitted.

• In regions of the genome with causal DNA, we should detect a clustering of *ibd* associated with trait similarity.

• Assess significance by permutation of trait values.

Inheritance of chromosome segments



- Each mat/pat genome of 3×10^9 bp ($\sim 3{,}000$ Mbp) is packaged into 22 chromosomes sized from 51 to 245 Mbp.

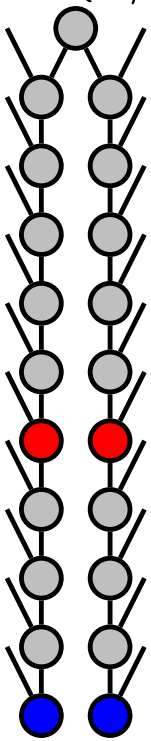
- Chromosomes are inherited in large chunks, $\sim$ 10^8bp or 100 Mbp.
- In any meiosis, crossovers occur as a Poisson process along the chromosome, rate 1 per 10^8 bp.
- Over m meioses, collectively crossovers occur as a Poisson process, rate m per 10^8 bp.
- The distance to the next crossover is exponential with mean $10^8/m$ bp.
- Exponential distributions have standard deviation equal to the mean.

ibd in remote relatives; (K. P. Donnelly, 1983)

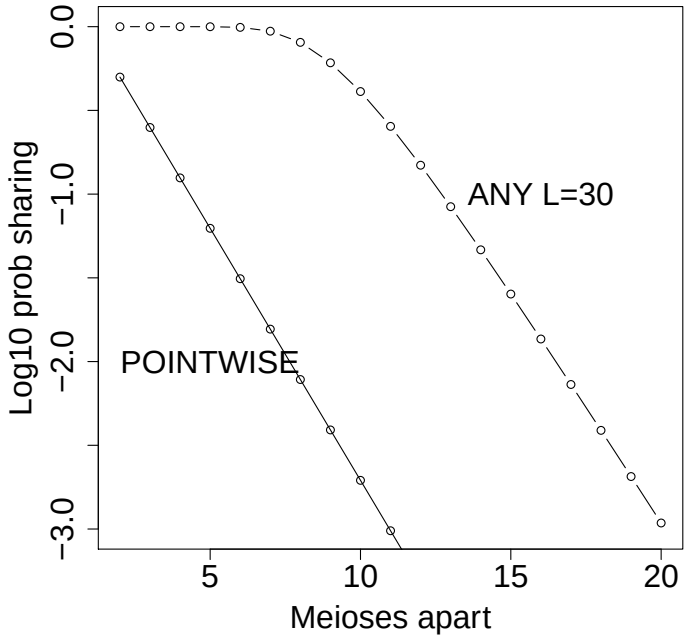
Relatives separated by m meioses.

Pr(2 kids get same) $= 1/2$

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



Pr(share any genome length L (10^8bp)) $= 1 - \exp(-(m-1)L/2^{m-1})$



Length of *ibd* segment $\sim m^{-1} \times 10^8$ bp.

	$m = 12$	$m = 20$
ibd at point	0.0005	2×10^{-6}
any *ibd* ($L = 30$)	0.148	0.001
length *ibd* segment	8.5 Mbp	5 Mbp

- *ibd* segments are rare but not short. The human genome is short.

Detecting *ibd* among individuals in populations

• For model-based inference of *ibd* $\mathbf{Z}$ from SNP data $\mathbf{X}$:
—need a model for the process of *ibd* $\mathbf{Z}$ along the chromosome,
—need a model for the SNP data $\mathbf{X}$ given $\mathbf{Z}$.

• Each SNP alone gives almost no information, but *ibd* comes in segments, with more and larger segments in closer relatives.

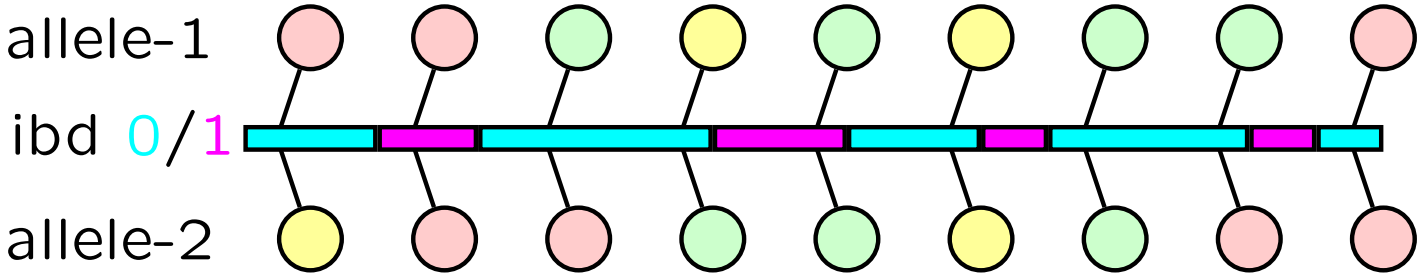
• DNA chunks that are *ibd* from a recent common ancestor are the same allelic type for the SNPs in the chunk (with high probability).

DNA that is not *ibd* will be of "independent" allelic type—basically, there will be differences at many SNPs.

• For model-based inference of *ibd*, use common variation!
Models require allele and/or haplotype frequencies;
Only for common SNPs can we have good estimates of the relevant population allele and local haplotype frequencies.

Realizing *ibd* segments from $\mathbf{X}$ in populations

- Two-haplotype model (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.

- *ibd* $\Rightarrow$ same allele; non-*ibd* $\Rightarrow$ independent alleles.
  Allow error so different alleles can still be *ibd*.

- Given a model, a standard HMM forward-backward algorithm gives realizations of *ibd* $\{\mathbf{Z}(j); j = 1, ..., \ell\}$ given $\mathbf{X}$, jointly over j, where $\mathbf{X}$ are allele types on the chromosomes over all loci.

Model for pointwise *ibd* among multiple gametes

- Ewens' sampling formula (ESF; Ewens, 1971) was originally deveoped to model allelic variation, but provides a one-parameter model for the partition of any n exchangeable objects.
- Each partition $\mathbf{Z}$ of n gametes into $k = |\mathbf{Z}|$ *ibd* groups v

$$\pi_n(\mathbf{Z}) \;=\; \frac{\Gamma(\theta)\;\theta^{|\mathbf{Z}|}}{\Gamma(n+\theta)} \prod_{v \in \mathbf{Z}} (|v|-1)!$$

- If $|\mathbf{Z}| = k$ and $\mathbf{Z}$ has a_j groups of size j

$$\pi_n(\mathbf{Z}) \;=\; \frac{\Gamma(\theta)\;\theta^{k}}{\Gamma(n+\theta)} \prod_{j} ((j-1)!)^{a_j}$$

with $k = \sum_j a_j$, $n = \sum_j j a_j$.

- Note for two gametes b and c, the probability of 1 group size 2 is

$$\pi_2(\mathbf{Z} = \{b,c\}) \;=\; \frac{\theta}{\theta(1+\theta)}((2-1)!)^1 \;=\; \frac{1}{(1+\theta)} \;\equiv\; \beta$$

is the probability of *ibd* between two gametes.

The Chinese restaurant process for building the ESF

- Tavaré and Ewens, 1997.

- Given a state with n people, at k tables, with a_j tables at which there are j people.
— New person sits at an empty table with probability $\propto \ (1-\beta)$, and to join each group of size j with prob. $\propto \ j\beta$.

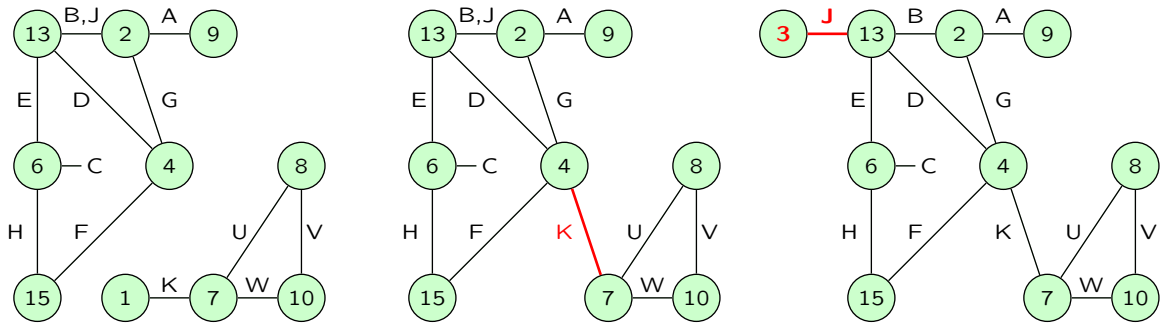
- $k = \sum_j a_j$, $n = \sum_j j a_j$.

- Example: New gamete g added to
$Z = (a,c,f),(b,e),(d) \ \sim \ \pi_6(\cdot)$ which has $k = 3, \ a_3 = a_2 = a_1 = 1$:

g joins	probability	new state Z^*	state character
(a,c,f)	$3\beta/(1+5\beta)$	$(a,c,f,g),(b,e),(d)$	$k=3, a_4 = a_2 = a_1 = 1$
(b,e)	$2\beta/(1+5\beta)$	$(a,c,f),(b,e,g),(d)$	$k=3, \ a_3 = 2, a_1 = 1$
(d)	$\beta/(1+5\beta)$	$(a,c,f),(b,e),(d,g)$	$k=3, \ a_3 = 1, a_2 = 2$
$(\cdot)$	$(1-\beta)/(1+5\beta)$	$(a,c,f),(b,e),(d),(g)$	$k=4, a_3 = a_2 = 1, a_1 = 2$

If $Z \sim \pi_6(\cdot)$, then $Z^* \sim \pi_7(\cdot)$. (n changes from 6 to 7.)

Changing *ibd* partitions across the chromosome



- Partition: $(\{Ap\}, \{Am, Bm, Jm, Gp\}, \{Gm, Dm, Fm\}, \{Cp, Cm, Ep, Hp\}, \{Bp, Jp, Dp, Em\}, \{Hm, Fp\}, \{\underline{Kp}\}, \{Km, Up, Wp\}, \{Um, Vp\}, \{Wm, Vm\})$.

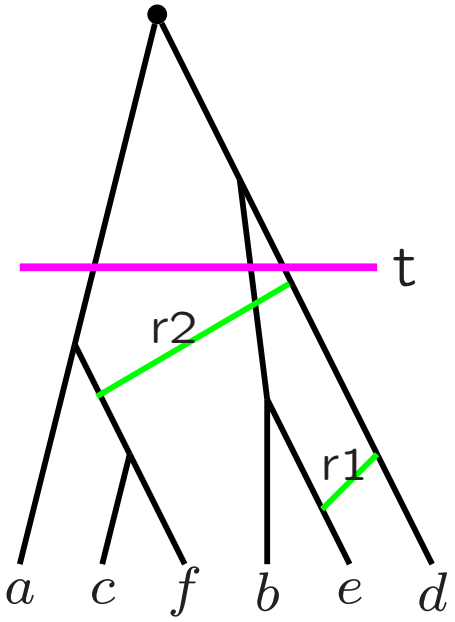
- Becomes:$(\{Ap\}, \{Am, Bm, \underline{Jm}, Gp\}, \{Gm, Dm, Fm, Kp\}, \{Cp, Cm, Ep, Hp\}, \{Bp, Jp, Dp, Em\}, \{Hm, Fp\}, \{Km, Up, Wp\}, \{Um, Vp\}, \{Wm, Vm\})$.

- Becomes:$(\{Ap\}, \{Am, Bm, Gp\}, \{Gm, Dm, Fm, Kp\}, \{Cp, Cm, Ep, Hp\}, \{Bp, Jp, Dp, Em\}, \{Jm\}, \{Hm, Fp\}, \{Km, Up, Wp\}, \{Um, Vp\}, \{Wm, Vm\})$.

- Recombination events in the ancestry of the gametes will move them among elements of the partition – we need a model for this process.

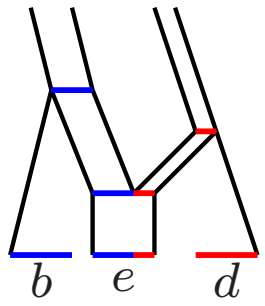
ibd partitions at a locus: the coalescent ARG

- The coalescent traces co-ancestry of chromosomes at a particular locus, back to the most recent common ancestor (MRCA).
- *ibd* is always relative. Relative to time t generations ago; $Z = ((a, c, f)(b, e)(d))$. Changing t, changes Z.
Pairwise *ibd* probability β is surrogate for t.
- Along a chromosome, the coalescent changes due to recombination events, and we have the *ancestral recombination graph* (ARG).



For example:

r1: $((b, e), (d)) \leftrightarrow ((b), (e, d))$



- If chromosomes share a recombination breakpoint, changes may involve > 1 chrom.

For example:

r2: $((a, c, f), (d)) \leftrightarrow ((a), (c, f, d))$

- But ARG model is too complex for genome-wide use.

Model for changing *ibd* among multiple gametes

• Modified CRP due to Chaozhi Zheng, allows any 1 gamete to move from one *ibd* subset to another, and has ESF as equil. dsn.

• Potential changes in *ibd* occur at some rate α per Mbp along the chromosome, a normalized recombination rate ρ.

• At a potential change point:
— First, an *extra* gamete, $*$, is proposed as a singleton with prob. $\propto\ (1-\beta)$, and to join each group of size j with prob. $\propto\ j\beta$.
— Next, one of the $n+1$ gametes is selected for deletion, and, if not deleted, $*$ is given the identity of the deleted gamete.

• Examples only, (each "dies" prob 1/7):

$*$ joins	probability	interim state	dies	new Z^*
(a,c,f)	$3\beta/(1+5\beta)$	$(a,c,f,*),(b,e),(d)$	d	$(a,c,d,f),(b,e)$
(b,e)	$2\beta/(1+5\beta)$	$(a,c,f),(b,e,*),(d)$	b	$(a,c,f),(b,e),(d)$
(d)	$\beta/(1+5\beta)$	$(a,c,f),(b,e),(d,*)$	e	$(a,c,f),(b),(d,e)$
$(\cdot)$	$(1-\beta)/(1+5\beta)$	$(a,c,f),(b,e),(d),(*)$	$*$	$(a,c,f),(b,e),(d)$

A note about models

• In pedigrees and in populations, Mendelian segregation and the crossover processes along a chromosome are real.

	Pedigrees	Populations
Model	Mendelian segregation and crossover process	Ewens sampling formula or coalescent
Prior for inferring *ibd* from $\mathbf{X}$	Yes (if correct)	Yes
Null distribution for $\mathbf{Z}$	Yes (if correct)	NO

• In pedigrees, we can base both *ibd* realizations and null distribution directly on this highly informative prior.

• In populations, the models based on ESF provides a good prior for realizations of *ibd* given $\mathbf{X}$ – because the data dominate.

• The model is only a (flexible) prior; can be made more flexible e.g. by including a component allowing a transition to a realization from $\pi_n(\mathbf{Z})$ independent of current state with small probability δ.

Realizing *ibd* partitions among multiple gametes

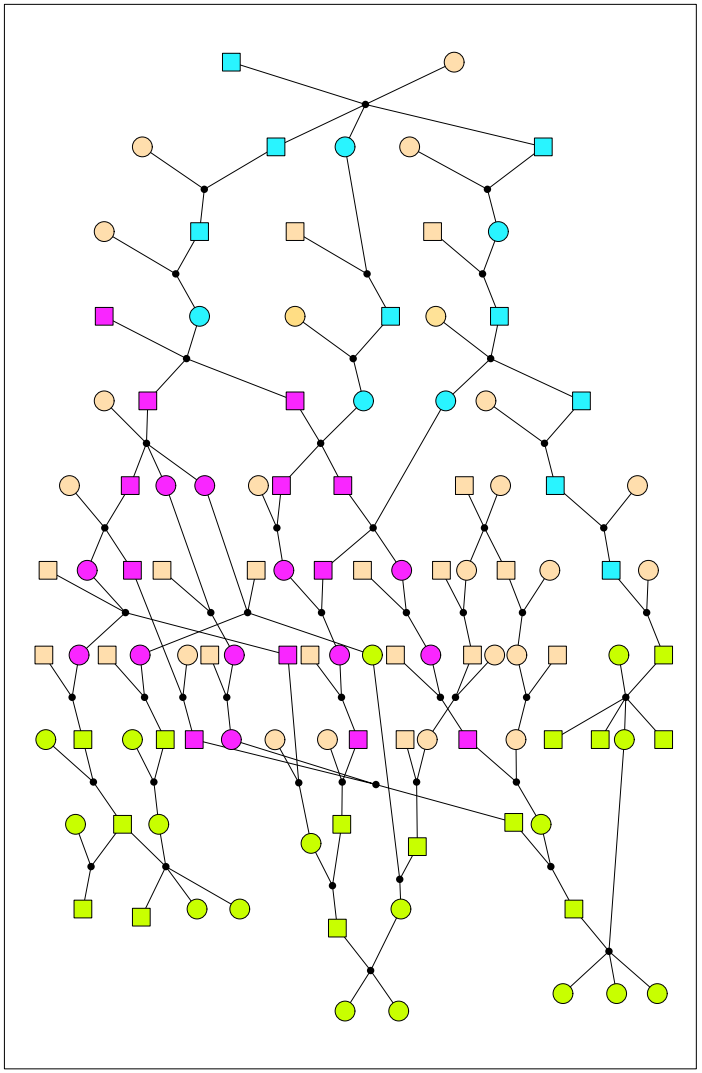
• We want joint inference, but for more than 6 gametes, the HMM is impractical – the number of partitions (*ibd* states) gets huge.

• Two possible MCMC approaches (for haploid gametes) :
—Chaozhi Zheng – full Bayesian MCMC of parameters, transition points and *ibd* transitions, given haplotype data.
—Chris Glazner – particle filter Monte Carlo approach.

• Another approach (due to Chris Glazner); (Results below).
Building the *ibd* state across a chromosome by adding diploid individuals successively to the *ibd* state, sampling from approximate conditionals, constrained by current state:
Sample *ibd* among A, B, C: first sample $(\mathbf{Z}(A,B)|X_A, X_B)$, then $(\mathbf{Z}(B,C)|\mathbf{Z}(A,B), X_B, X_C)$, then $(\mathbf{Z}(A,C)|\mathbf{Z}(B,C), \mathbf{Z}(A,B), X_A, X_C)$.
Likelihood is "*Product of approximate conditionals*"

• Using Markov models for latent *ibd*, with marker data $\mathbf{X}$ dependent on the latent *ibd* state, we can realize *ibd* $\mathbf{Z}$ among gametes of individuals not known to be related.

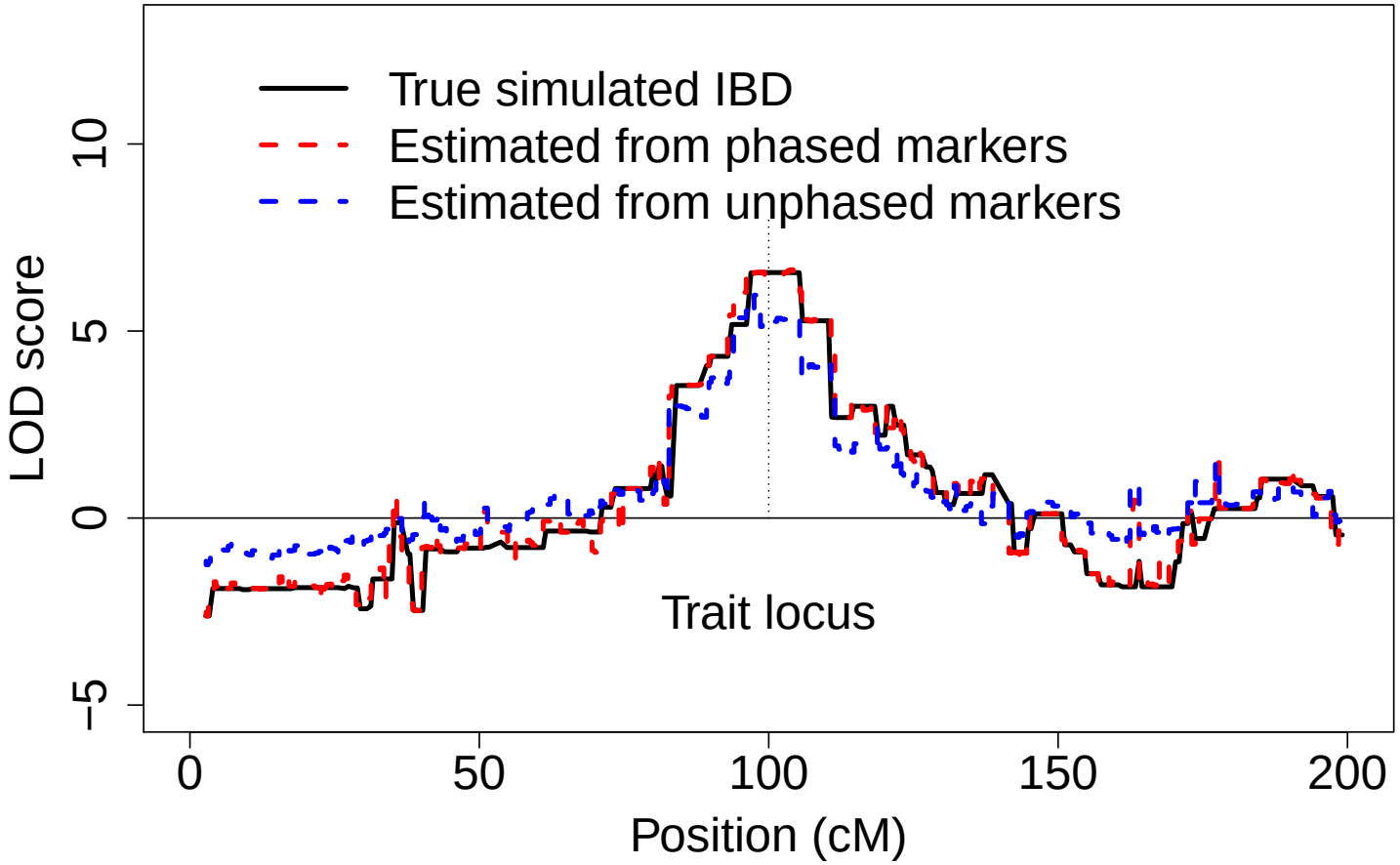
An example of related individuals in a population



- Causal DNA descends from magenta founder to the three green families.
- Quantitative trait is simulated on green families, given genotypes at the causal locus.
- Descent across the chromosome is simulated given descent at the causal locus.
- SNP marker data are simulated on the three green families, given each SNP marker location descent.

Lod scores based on inferred *ibd*; No pedigree info!

- Results due to Chris Glazner.



- Results assessed by ability to recover linkage lod score.
- Information comes from between family *ibd*

- If data can be phased (i.e. we can identify the haplotypes that make up the genotypes of the observed individuals) we can almost perfectly recover the true-*ibd* curve.

Summary:
Genetic analyses can be based on inferred *ibd*

- Modeling descent is important: *ibd* measures relevant location-specific relatedness, whether in pedigrees or in populations

- Modeling genomes is important: our genomes are not 3 million exchangeable SNPs. In terms of *ibd* segments, human genomes are short.

- Models are important: Models do not mimic reality. Models provide a map to assess inferences and information.

- Models should be flexible:
— an unvalidated pedigree prior is not flexible.
— assuming no error in marker data is not flexible.

- In pedigrees and populations, modern SNP data, $\mathbf{X}$ enable realizations of *ibd* given $\mathbf{X}$, but the source of the *ibd* inference is almost irrelevant to analysis. (Pedigrees, if correct, provide a "true null".)

- Genetic analyses can be based on inferred *ibd*.