Coancestry in the analysis of complex traits

Elizabeth Thompson
University of Washington

For: Simons Institute Workshop
Berkeley, California
18-21 February, 2014

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

Genetic variation, Association, and Descent

• For genetic analysis, 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 $\mathbf{Y}$.

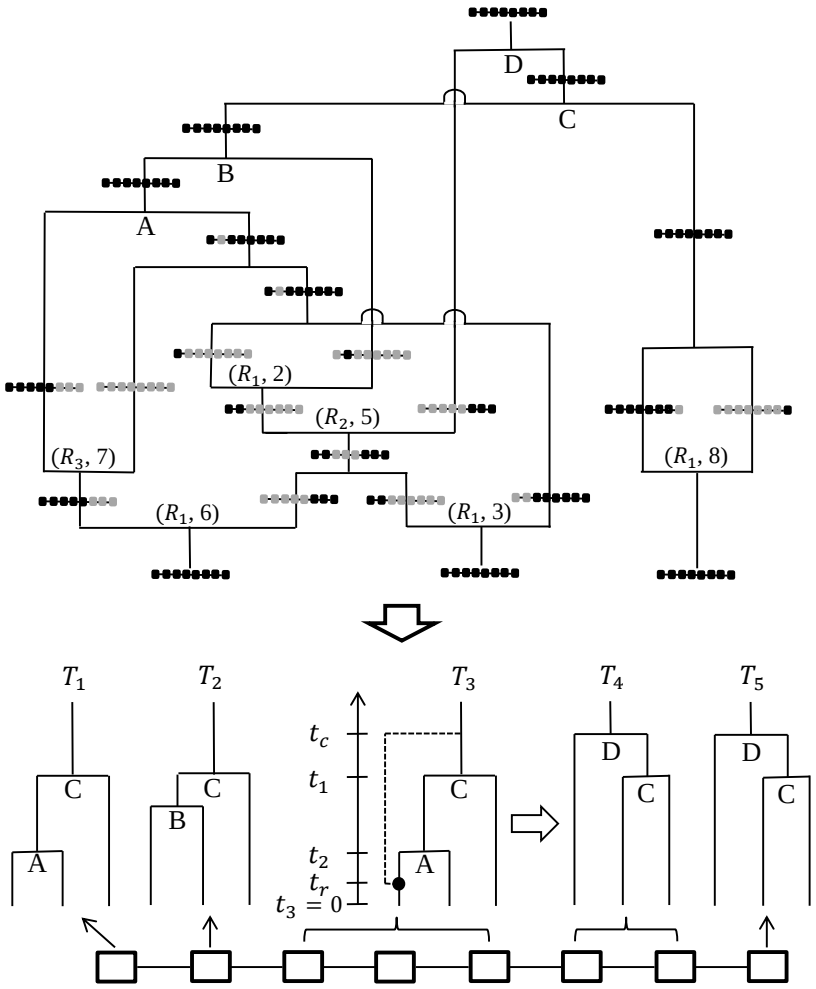
• Direct testing for an association between $\mathbf{Y}$ and allelic type $\mathbf{X}$ at each SNP location ignores the fact that DNA descends in blocks.

• 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*) relative to some ancestral population, if it is a copy of the same DNA in that population.

• 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, $\mathbf{Y}$.

An *ibd* model too complex to use



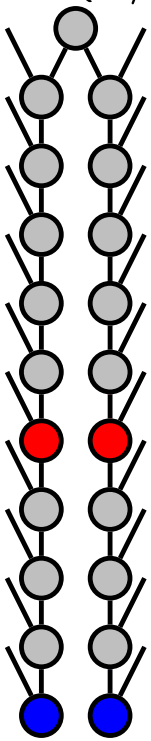
- Full specification of ancestry is the *ancestral recombination graph* or ARG: Figure due to Chaozhi Zheng.
- MCMC sampling of the ARG (Kuhner et al.) or of its sequential Markov approximations, (Zheng et al.) is hard (even for 500 kbp).
- Main problem: Our interest is in long lengths (> 1 Mbp) and short time depths < 50 generations. Most of the ARG is irrelevant.

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

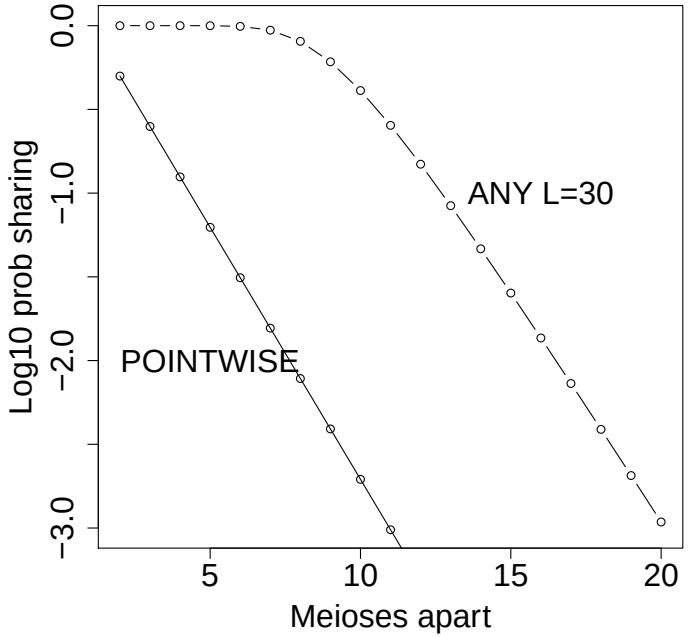
Relatives separated by m meioses.

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

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



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



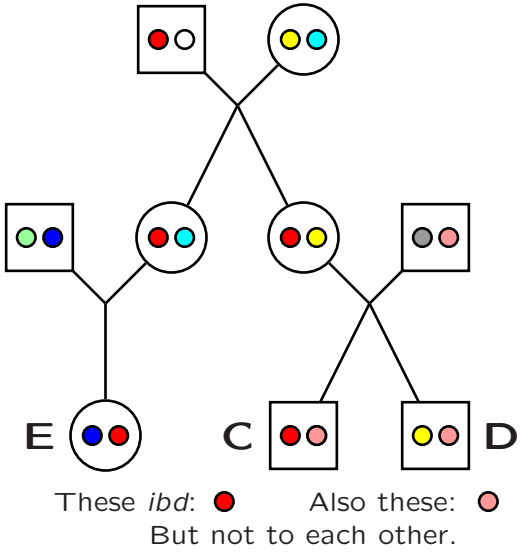
Length of *ibd* segment $\sim m^{-1} \times 10^8$ bp.

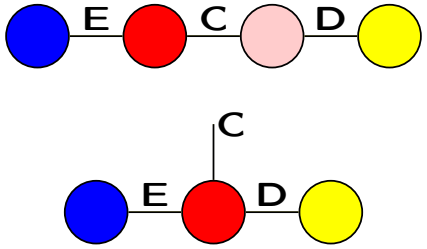
	$m = 12$	$m = 20$
ibd at point	0.0005	2×10^{-6}
any *ibd* ($L = 30$)	0.148	0.001
length *ibd* segment	8.5 Mbp	5 Mbp

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

Identity by descent is sufficient for analysis

- Given *ibd*, the pedigree is no longer relevant.
  The *ibd* may come from a pedigree or population inference.



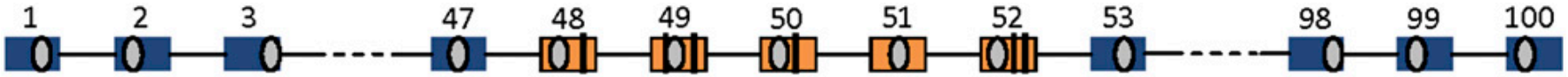


●	●	●	○	Prob	Pr(● ≡ ●)
b	a	a	b	$q_a^2 q_b^2$	h
b	a	–	b	$q_a q_b^2$	1-h

- For example: $\Pr(E = ab,\ C = aa,\ D = ab)$.
- Or $\Pr(Y_E, Y_C, Y_D) = \sum_{●} \sum_{●} (\Pr(Y_E|●,●) q(●) q(●)$
  $\sum_{●} (\Pr(Y_C|●,●) q(●) \sum_{○} (\Pr(Y_D|●,○) q(○)))$
- In a population (e.g. ● and ●),
  a population probability model is needed to provide h.
- In a pedigree/population: marker (SNP) data and pedigree/population prior give probabilities and realizations of *ibd*.

Case-Control Simulation Study of *ibd*

- Browning and Thompson, Genetics, 2012: Is there enough power?
- 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.
- Relative to $G = 25$ the location-specific *ibd*, $\mathbf{Z}$, is assumed known.



- 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.
- Total number of variants in the population in the 5 central 1kb blocks ranged from 7-10 (strongest selection) to 11-16 (weakest selection).
- Individuals with ≥ 1 of these causal variant alleles are cases with probability 0.1.

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 test SNP locations across the 200 kb region.
- 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.
- 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.)
- Power of tests in large population: $N_e = 10^5$ for $G = 25$.

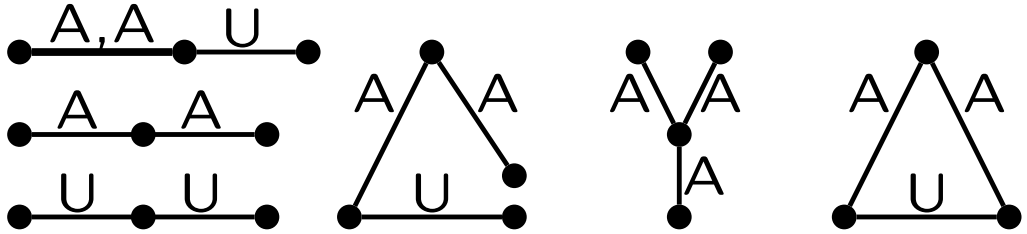
selec -tion	tot.freq variants	assoc. max R^2	# cases= # contr.	power assoc.	power *ibd*	association vs. *ibd*
0.0005	0.045-0.13	0.91-1.00	500	0.87	0.57	assoc.
0.001	0.019-0.05	0.28-1.00	500	0.65	0.53	Not-Sig
0.002	0.010-0.03	0.06-0.52	1000	0.53	0.87	*ibd*
0.005	0.004-0.01	0.03-0.16	3000	0.47	0.90	*ibd*

Joint trait-related *ibd* in population samples

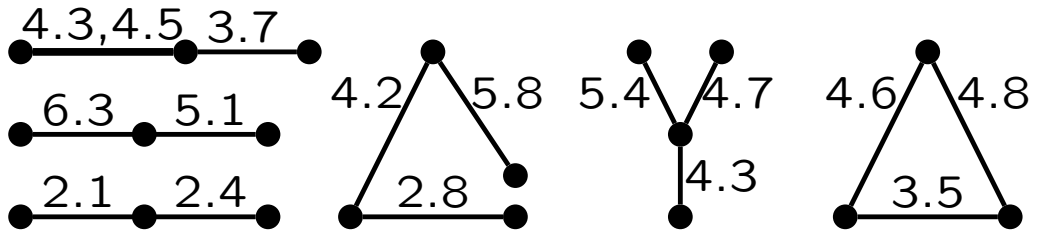
• In a population, trait-related *ibd* can indicate causal locations, but we gain by considering *ibd* among multiple individuals.

• 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, and can assess significance by permutation of trait values.

• A trait model – even ranked quantitative values – increases power.

First, detect the *ibd* among individuals

• Model-based inference of *ibd* $\mathbf{Z}$ from SNP data $\mathbf{X}$:
provides measures of uncertainty, not a point estimate,
allows realizations from the probability distribution given the data,
i.e. from the joint distribution across the genome segment.

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

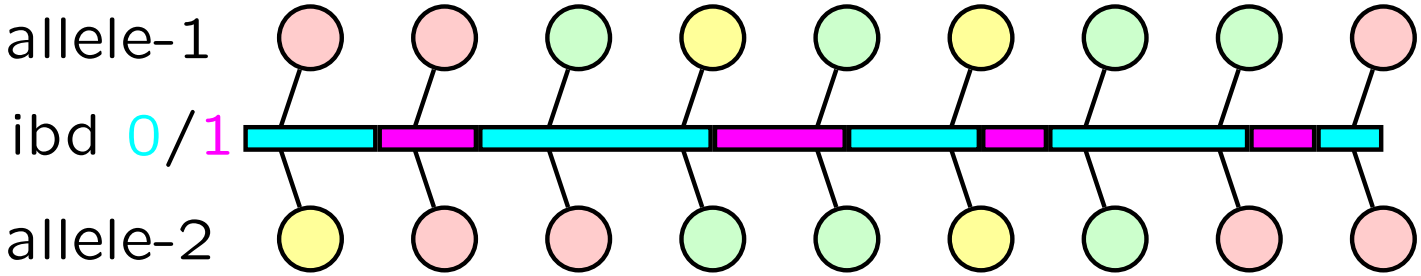
• DNA chunks that are *ibd* from a recent common ancestor are the same allelic type for the SNPs in the chunk (with high probability).
DNA that is not *ibd* will be of "independent" allelic type—
basically, there will be differences at many SNPs.

• Need a model for the process of *ibd* $\mathbf{Z}$ along the chromosome,
Need a model for the SNP data $\mathbf{X}$ given $\mathbf{Z}$.

• 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-gamete model (Leutenegger et al. 2003)



- Two-parameter Markov model: marginal prob β, rate change α.

In reality, *ibd* is not Markovian 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 gametes 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

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

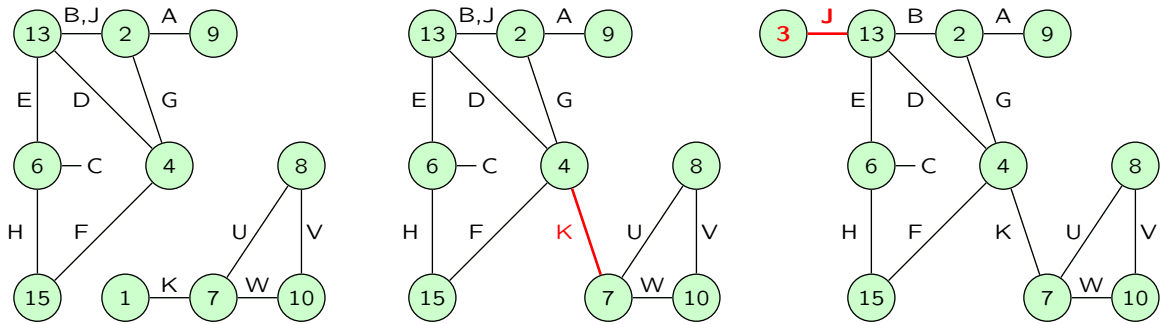
with $k = \sum_j a_j$, $n = \sum_j j a_j$.

- 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.

Changing *ibd* partitions across the chromosome



- Partition: $(\{A1\}, \{A0, B0, J0, G1\}, \{G0, D0, F0\}, \{C1, C0, E1, H1\}, \{B1, J1, D1, E0\}, \{H0, F1\}, \{K1\}, \{K0, U1, W1\}, \{U0, V1\}, \{W0, V0\})$.

- Becomes:$(\{A1\}, \{A0, B0, J0, G1\}, \{G0, D0, F0, K1\}, \{C1, C0, E1, H1\}, \{B1, J1, D1, E0\}, \{H0, F1\}, \{K0, U1, W1\}, \{U0, V1\}, \{W0, V0\})$.

- Becomes:$(\{A1\}, \{A0, B0, G1\}, \{G0, D0, F0, K1\}, \{C1, C0, E1, H1\}, \{B1, J1, D1, E0\}, \{J0\}, \{H0, F1\}, \{K0, U1, W1\}, \{U0, V1\}, \{W0, V0\})$.

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

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.)

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.

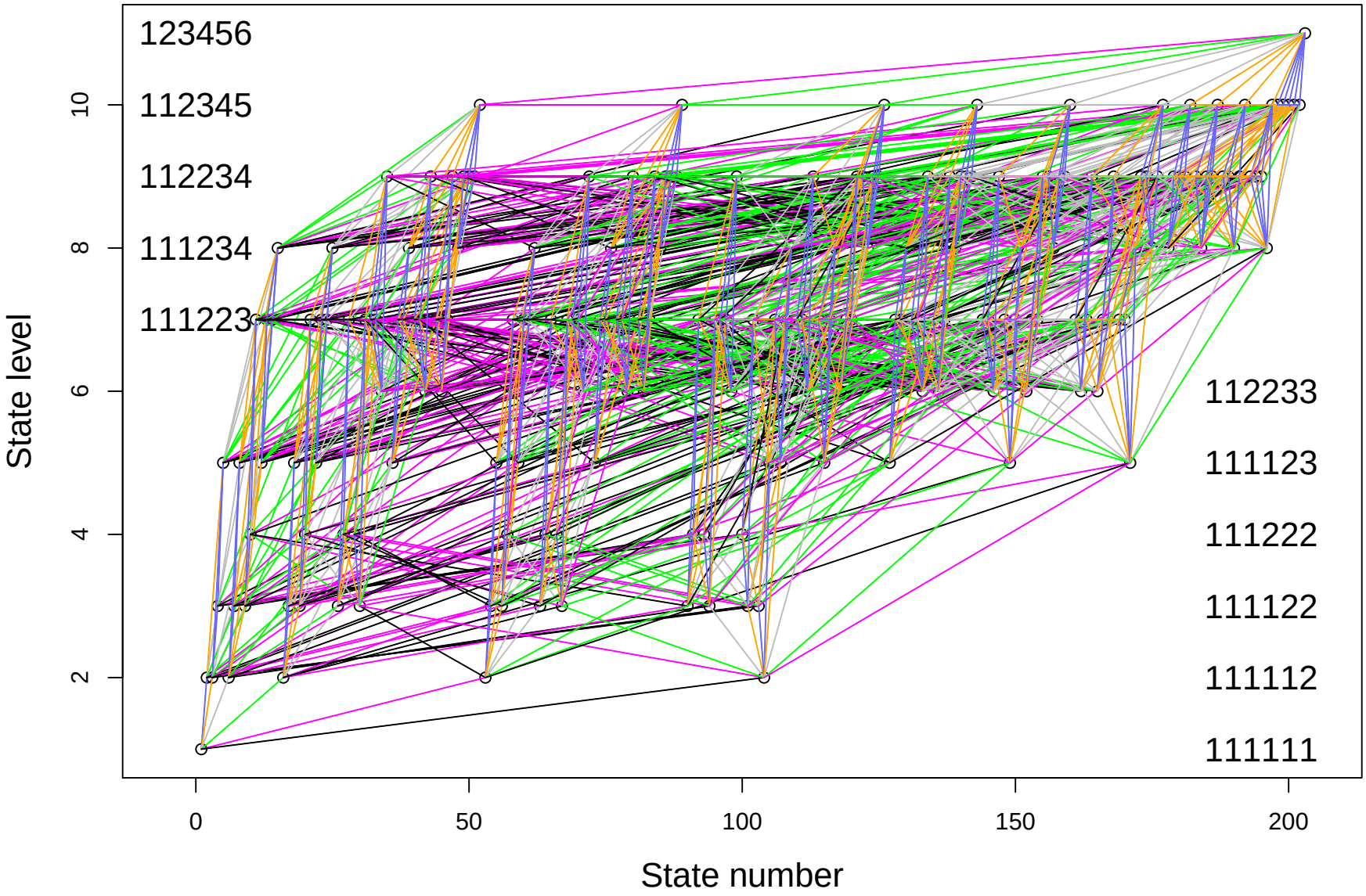
- At a potential change point:
— First, an *extra* gamete, $*$, is proposed as a singleton with prob. $\propto\ (1-\beta)$, and to join each group of size j with prob. $\propto\ j\beta$.
— Next, one of the $n+1$ gametes is selected for deletion, and, if not deleted, $*$ is given the identity of the deleted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

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

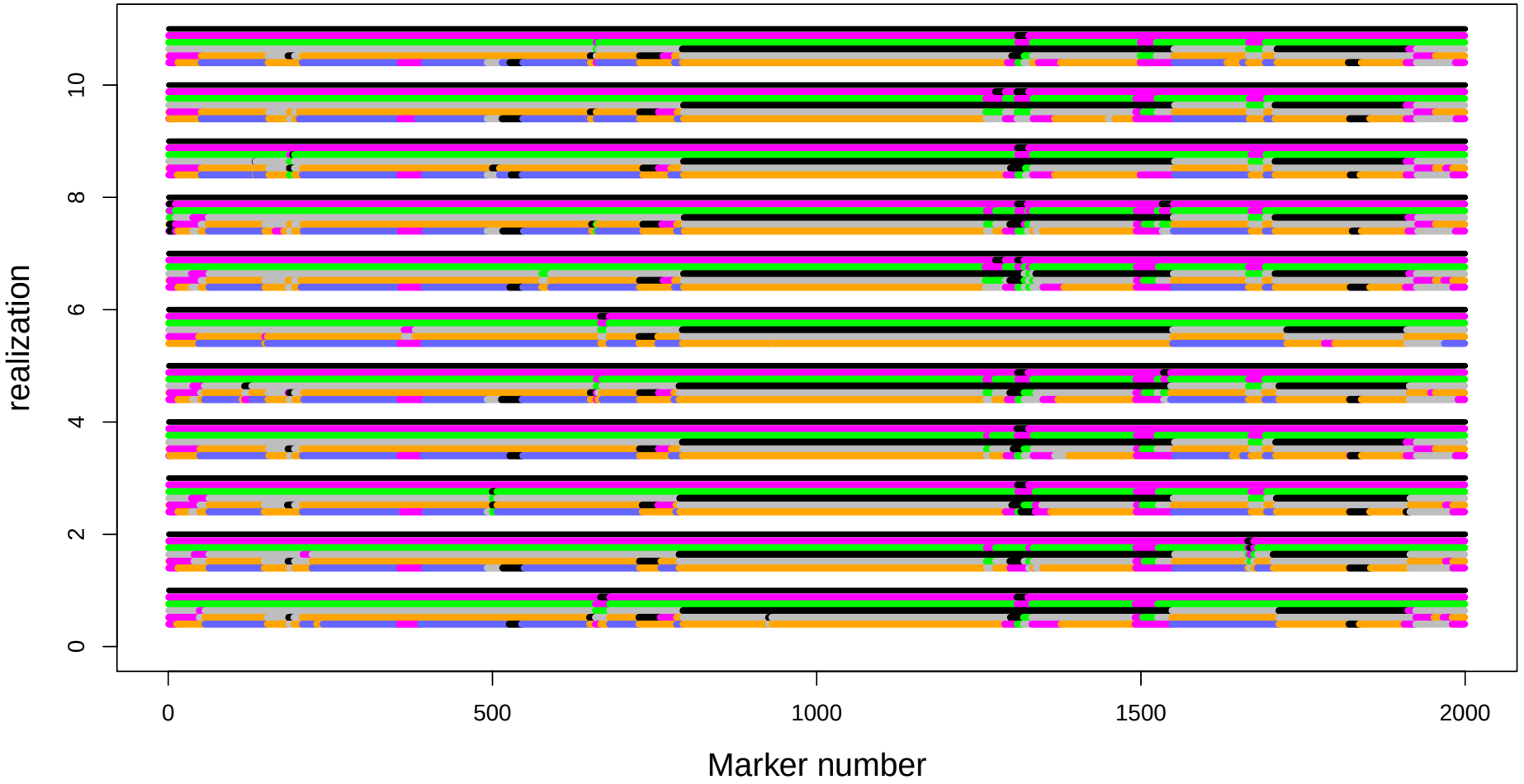
$*$ joins	probability	interim state	dies	new Z^*
(a,c,f)	$3\beta/(1+5\beta)$	$(a,c,f,*),(b,e),(d)$	d	$(a,c,d,f),(b,e)$
(b,e)	$2\beta/(1+5\beta)$	$(a,c,f),(b,e,*),(d)$	b	$(a,c,f),(b,e),(d)$
(d)	$\beta/(1+5\beta)$	$(a,c,f),(b,e),(d,*)$	e	$(a,c,f),(b),(d,e)$
$(\cdot)$	$(1-\beta)/(1+5\beta)$	$(a,c,f),(b,e),(d),(*)$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

- Now if $Z\sim\pi_6(\cdot)$, then $Z^*\sim\pi_6(\cdot)$.

Transitions in the state space for 6 gametes

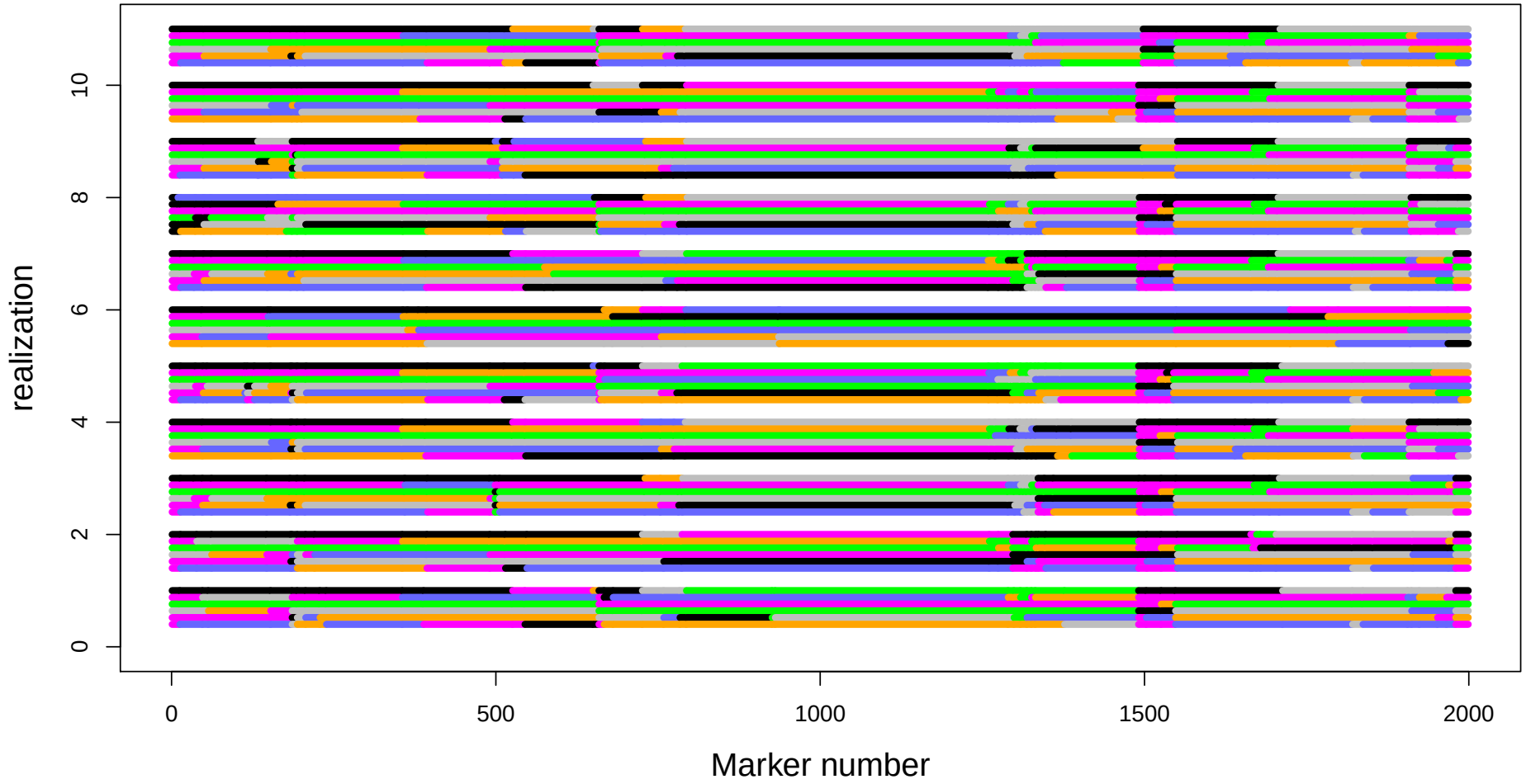


HMM realizations for six gametes



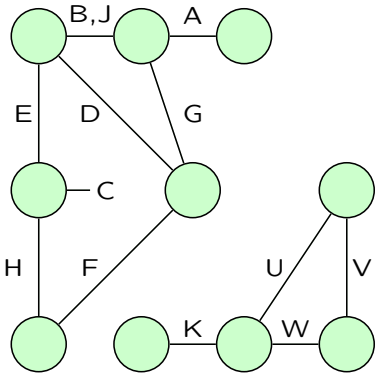
- 2000 SNPs in 51.4 Mbp from simulated 200-generation population
- One truth, and 10 independent realizations given SNP data
- Gamete *ibd* states are labelled in canonical ordering

Label switching problems of representation



- Now gamete changes color only if involved in an *ibd* transition.
- However, colors lose identity across the chromosome.
- Weight realizations by using relative local likelihoods under LD.

ibd graph equivalences across genomes



- The (unlabeled) nodes of an *ibd* graph have identity only through the (labeled) edges that connect them.
- *ibd* graphs are slowly changing across the genome (on bp scale) — in realizations only changes are recorded.
- Any feature of the graph (e.g. set of edges at a given node) has a marker or bp-range over which it exists.
- The IBDgraph software incorporates these features, identifying graph equivalences. (Koepke and Thompson, JCB, 2013).
- IBDgraph allows for efficient insertion, querying, equality testing, and set operations on ibd-graph collections, at or over markers.
- The IBDgraph software takes only a few seconds to run, and can reduce trait likelihood computations by two orders of magnitude.
- Allows trait models based on joint *ibd* at more than one locus.

Realizing *ibd* partitions among multiple gametes

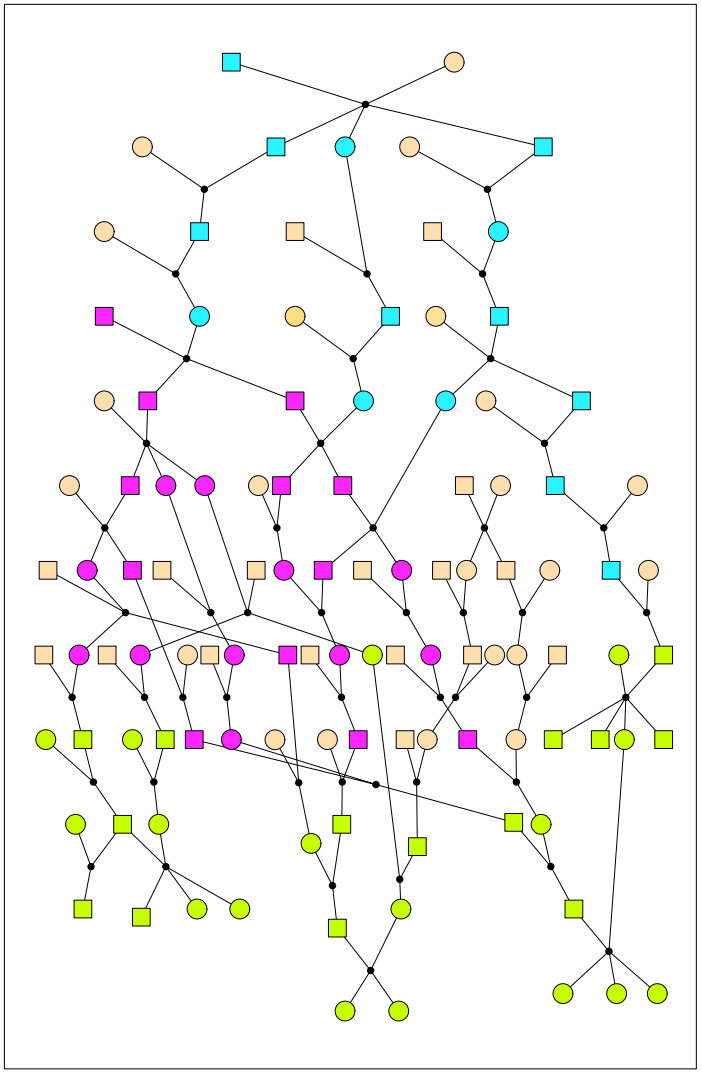
• We want joint inference, but for more than 6 gametes, the HMM is impractical – the number of partitions (*ibd* states) gets huge.

• Two possible MCMC approaches (for haploid gametes) :
—Chaozhi Zheng – full Bayesian MCMC of parameters, transition points and *ibd* transitions, given haplotype data (in press; JCB).
—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.

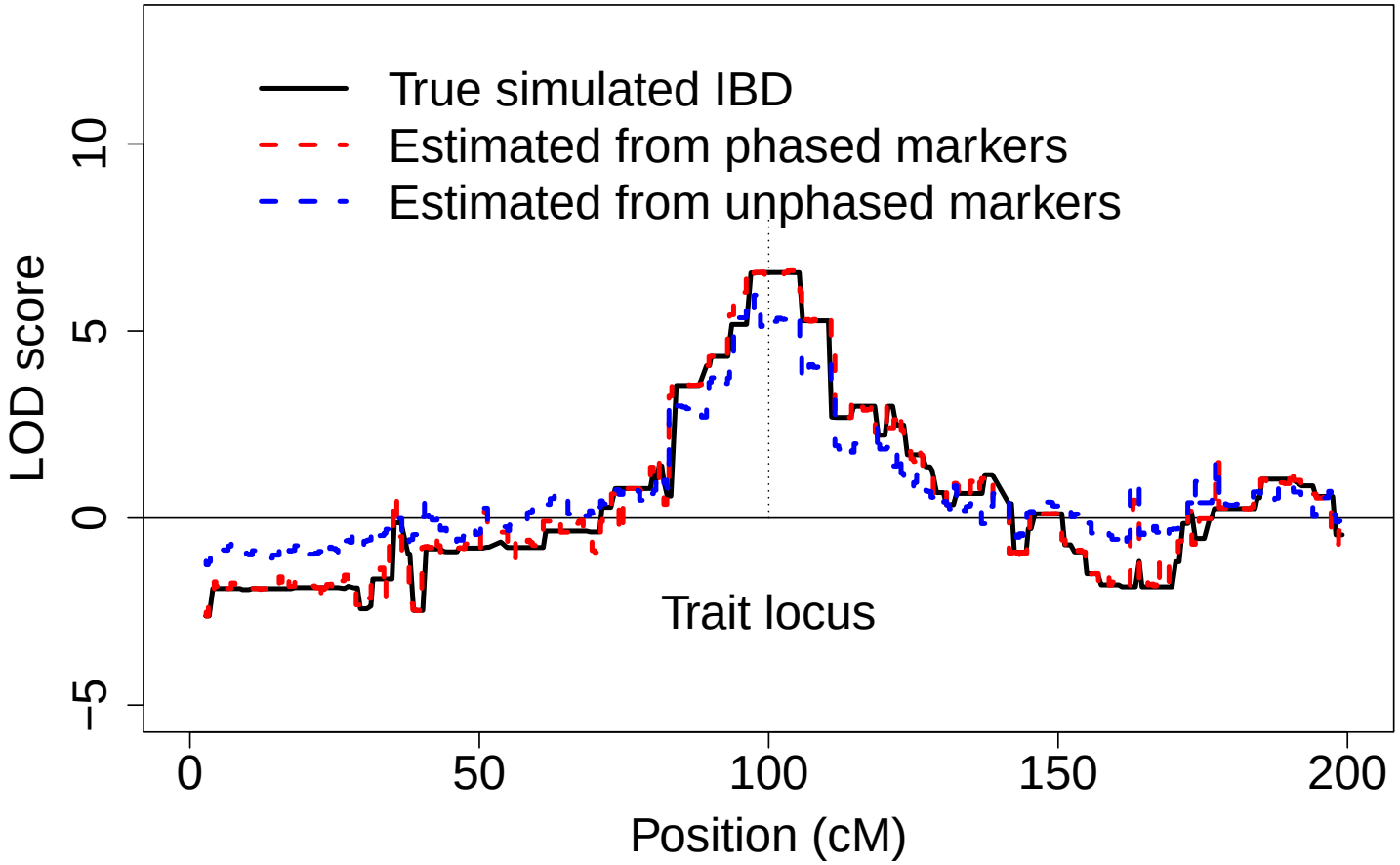
An example of related individuals in a population



- A simulation:
- 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* lod-score curve.

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

- In populations, modern SNP data enable realizations of *ibd*.
- The pedigree/population source of the *ibd* inference is irrelevant to analysis — lod scores and test statistics are funtions of *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:

— assuming a pedigree structure is not flexible.
— assuming no error in marker data is not flexible.
— assuming only transitions of a single gametes is not flexible.

References

• Brown, M. D., Glazner, C. G., Zheng, C., and Thompson, E. A. (2012) Inferring coancestry in population samples in the presence of linkage disequilibrium. Genetics, 190: 1447–1460.

• Browning, S. G. and Thompson, E. A. (2012) Detecting rare variant associations by identity by descent mapping in case-control studies. Genetics, 190: 1521-1531.

• Koepke, H. A., and Thompson, E. A. (2013) Efficient identification of equivalences in dynamic graphs and pedigree structures. Journal of Computational Biology 20: 551–570.

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

• Zheng, C., Kuhner, M. K., and Thompson, E. A. (2014) Joint inference of identity by descent along multiple chromosomes from population samples. Journal of Computational Biology: in press.