Pedigree information in contrast to population-inferred descent

Elizabeth Thompson, Department of Statistics,
University of Washington, Seattle, WA, USA

For: BIRS Workshop,
New Methods for Family-based studies.
Aug 7, 2018

No sequence data! Inferring regions of interest
using common SNP variation,
before we consider sequence-variant associations.

Finding genes in the SNP era

- Goal: to find where in the genome are DNA variants that affect the values Y of a trait of interest.
- For genetic analysis, the data are:
  — genetic marker (SNP) data X; the allelic DNA types at known locations in the genome, and
  — and trait data Y (qualitative or quantitative).
- Association mapping considers directly the association between marker types X and trait values Y:
- But associations arise from descent of genome:

$$X \leftrightarrow Y$$

  --- genomes descend in large segments,
  --- functional genes are segments and
  --- there is variant heterogeneity in any functional gene.
- So consider association of X and Y through descent Z.

$$X \leftarrow Z \rightarrow Y$$

IBD-based gene mapping

- Similarity of phenotype Y increases probability of shared descent Z in causal regions, relative to

  --- that expected given pedigree (?) relationships

  --- similarly related (?) control (?) individuals

  --- same individuals in non-causal regions (assume exist)

- Idea: detect location-specific shared descent, Z, at locations of common SNP markers, X, among individuals of similar trait values, Y.

$$X \rightarrow Z \rightarrow Y$$

  – causal variants need not be pre-identified, hypothesized, or even typed.

  – ibd-based test integrates across (rare) variants, somewhat (?) addressing allelic heterogeneity.

Computing P(Y|X) using $X \to Z \to Y$

- Assume that, given Z, X and Y are independent.
- Given model Θ_X for X, Θ_Y for Y, and causal DNA at locations λ, compute $L_Y(\Theta) = P(Y \mid X ; \Theta_X, \Theta_Y, \lambda)$

$$P(Y \mid X;\Theta_X,\Theta_Y,\lambda) = \sum_{Z_\lambda} P(Y \mid Z_\lambda;\Theta_Y,\lambda) P(Z_\lambda \mid X,\Theta_X)$$

- But in general the number of possible Z_λ is huge.
- So we use a Monte Carlo estimate:

$$\hat{P}(Y \mid X;\Theta_X,\Theta_Y,\lambda) = \frac{1}{N}\sum_{k=1}^{N} P(Y \mid Z_\lambda^{(k)};\Theta_Y,\lambda)$$

Z➔ Y

where, for k=1,…,N,

$$Z_\lambda^{(k)} \sim P(\bullet \mid X;\Theta_X)$$

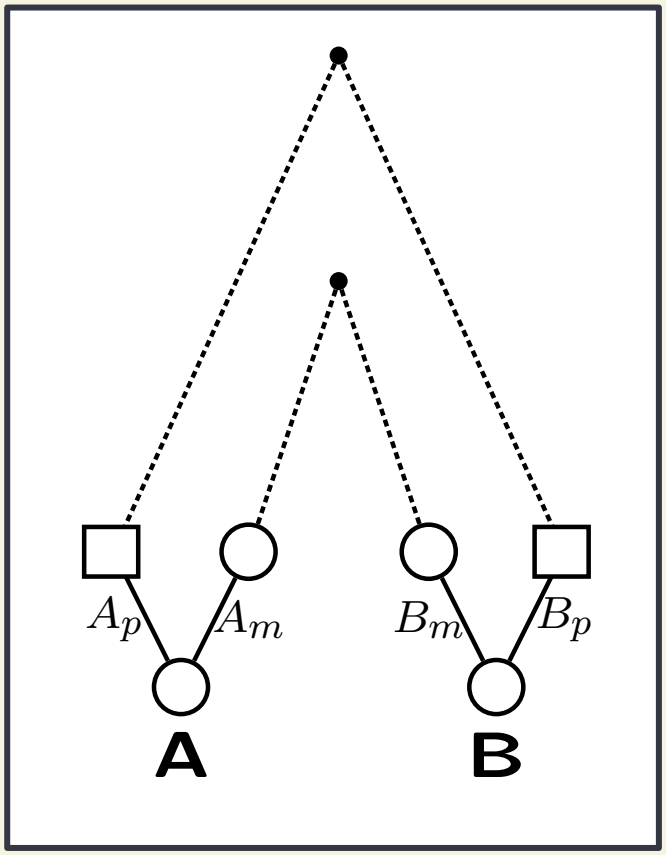
X ➔Z

Defining Z: IBD states at a locus

- At a locus, an ibd state on *n* haploid genomes is a partition of *n* labelled objects.
- The number of states (partitions) increases very rapidly with *n*. For n = 4, 6, 12, we have 15, 203, > 4×10^6.
- For the 15 states in pairs of individuals (*n*=4), each gives a kinship value 0, 1/4, 1/2, or 1.

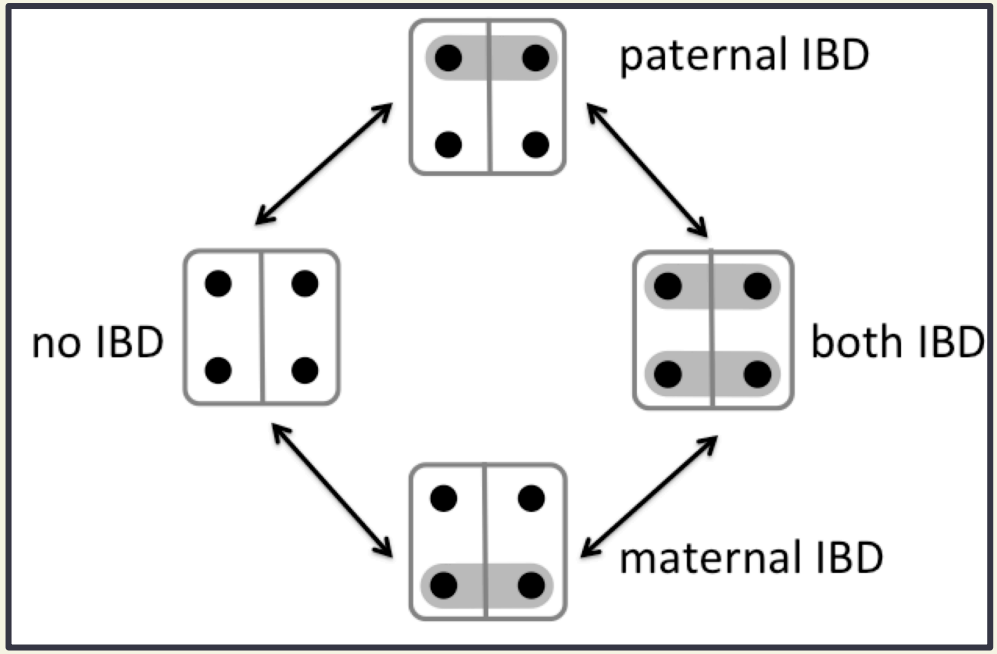
1		$\{A_p, A_m, B_p, B_m\}$	1
2		$\{A_p, A_m\}, \{B_p, B_m\}$	0
3		$\{A_p, A_m, B_p\}, \{B_m\}$	1/2
4		$\{A_p, A_m, B_m\}, \{B_p\}$	1/2
5		$\{A_p, A_m\}, \{B_p\}, \{B_m\}$	0
6		$\{A_p, B_p, B_m\}, \{A_m\}$	1/2
7		$\{A_p\}, \{A_m, B_p, B_m\}$	1/2
8		$\{A_p\}, \{A_m\}, \{B_p, B_m\}$	0
9		$\{A_p, B_p\}, \{A_m, B_m\}$	1/2
10		$\{A_p, B_m\}, \{A_m, B_p\}$	1/2
11		$\{A_p, B_p\}, \{A_m\}, \{B_m\}$	1/4
12		$\{A_p, B_m\}, \{A_m\}, \{B_p\}$	1/4
13		$\{A_p\}, \{A_m, B_p\}, \{B_m\}$	1/4
14		$\{A_p\}, \{A_m, B_m\}, \{B_p\}$	1/4
15		$\{A_p\}, \{A_m\}, \{B_p\}, \{B_m\}$	0

IBD at a locus and over loci



- At a locus, DNA may descend from common ancestors, resulting in *ibd.*

$$P(ibd) \propto (1/2)^m$$



- Over loci*: ibd* changes due to recombination in ancestral lineage

$$length(ibd) \propto (1/m)$$

IBD segments rare but not short

Probabilities of	*m*=12	*m*=20
ibd at locus	0.0005	0.000002
any *ibd* (human)**	0.148	0.001
Length *ibd* segment	8.5 Mbp	5 Mbp

- In remote relatives, there is no ibd with high probability.
- If there is ibd it comes in long segments

** K. Donnelly (1983) – my first PhD student.

Realizing Z from X

$$X \rightarrow Z$$

- All models are false but some are useful (George Box). We need tractable, flexible, models for Z that allow the SNP data to "speak".

$$P(Z \mid X) \propto P(X \mid Z)P(Z)$$

- For the descent, Z or ibd,
  - Use the fact that segments of ibd are large
  - we may need to estimate joint descent among n haploid genomes (n/2) individuals.
  - Even for two individuals, there are 4 genomes.
- For SNP markers, X,
  - There are many SNPs and good models.
  - Each SNP is quite uninformative
  - Need to combine information over SNPs

(1) Population Model for IBD: P(Z)

- We need a Markov false but useful "flexible prior":
- For any pair of haploid genomes: (Leutenegger et al, 2003)
  - a level of ibd: β (measures relatedness/kinship)
  - a change rate of ibd: α
    (controls lengths of ibd and non-ibd segments)
- Among multiple (or 4) haploid genomes:
  - Ewens' sampling formula (ESF) is a population genetics model for ibd partition with single parameter β which is the pairwise probability of ibd
  - Potential changes at rate α, with a model for consistent combination of changes of ibd state Z that maintains ESF marginally at all points (Chaozhi Zheng).

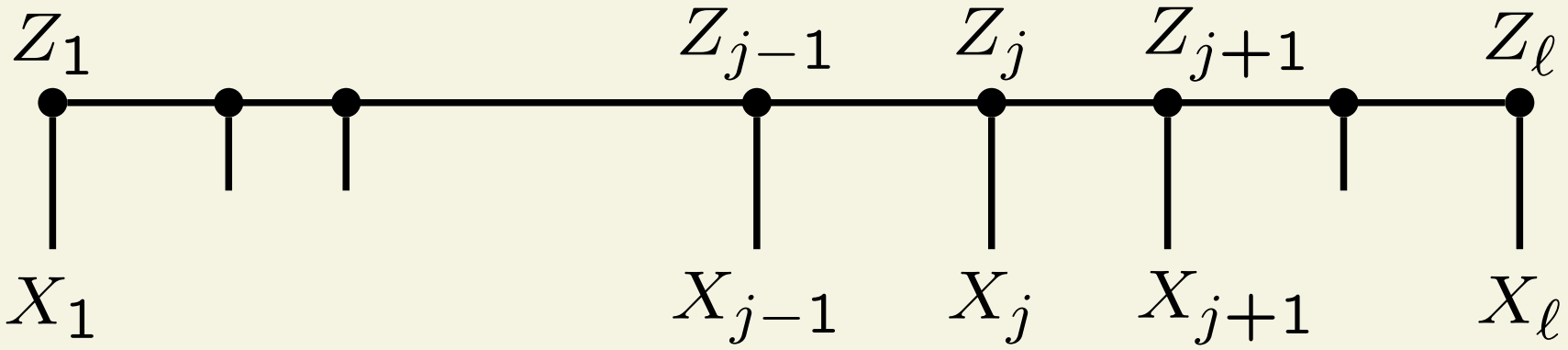
(2) Model for $P(X_j | Z_j)$

- DNA in current individuals that descends from a recent single ancestral DNA (ibd) is very likely the same allelic type.

A simple model is:

- ibd genes are of the same allelic type:
  - ignores mutation etc.
- Non-ibd genes are of independent types:
  - ignores population structure etc.
- Allow a small probability of error for flexibility.
- All models are false, but some are useful (George Box): This one is VERY useful.

Realizing Z given X: the HMM



- 1. A Markov model for
  a) Pointwise ibd among haploid genomes
  (15 states for n=4; pairs of individuals)
  b) Changing ibd across a chromosome
- 2. A model for marker data X given ibd (Z)

Realize ibd states Z across all chromosomes, given X:

$$P(Z \mid X) \propto P(X \mid Z) P(Z)$$

Estimate realized ibd states (hence kinship):

location-specific and genome-wide

IBD-based Likelihoods for a VC model

- Variance component (Random effects) model:

  At each location j, the vector of quantitative observations Y on the individuals is modeled as

  $Y = \mu 1 + \tau_j w_j + \sigma_a g + \sigma_e e$

  $Z \rightarrow Y$

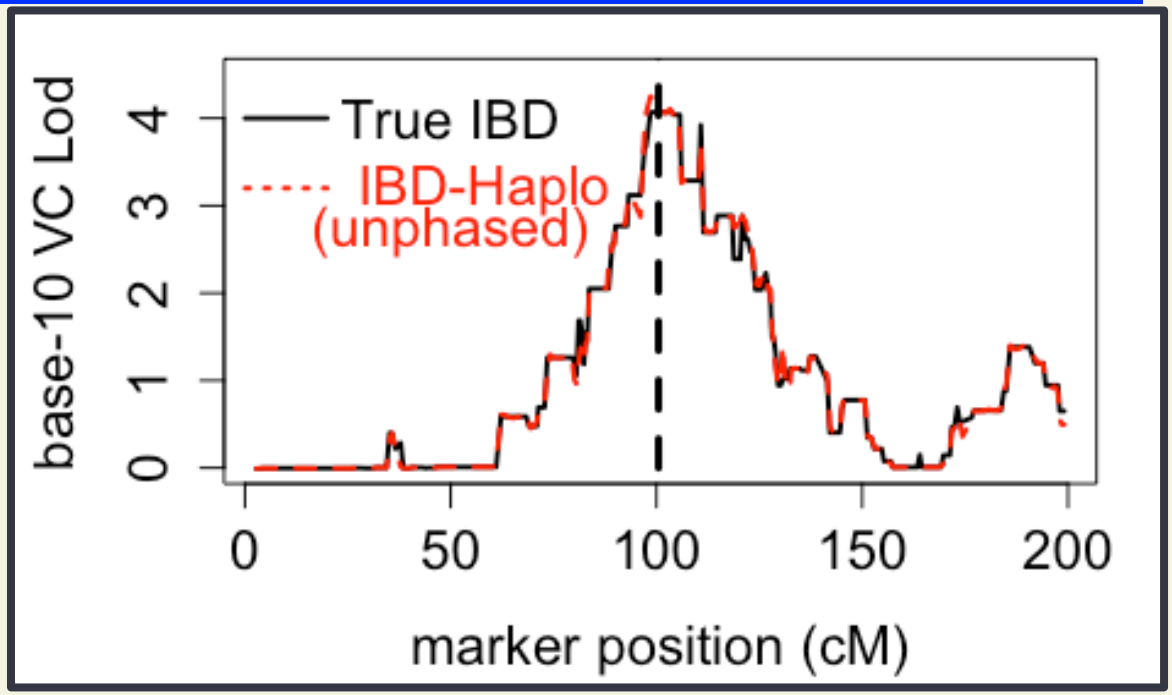
- w_j, g, e mean 0, $var(w_j) = 2\Phi_j$, var(g) = 2 Ψ, var(e) = I.
  where Φ_j is the pairwise ibd (kinship) matrix at location *j*, and Ψ is the genome-wide realized ibd.
- Compare the model in which the is a causal effect at *j*, with the model of no effect $\tau_j^2=0$.

$$l_j = \log_{10}\left(\frac{\max_{\mu,\sigma_a^2,\tau_j^2\sigma_e^2} L_Y(\mu,\sigma_a^2,\tau_j^2,\sigma_e^2;\Phi_j,\Psi)}{\max_{\mu,\sigma_a^2,\sigma_e^2} L_Y(\mu,\sigma_a^2,\tau_j^2=0,\sigma_e^2;\Psi)}\right)$$

Lod scores without pedigrees: it works!

Simulated data:
31 individuals in 3 connected families.

1. Z_0 at 100.5 cM
2. Z_0 ➔ **Z**,
3. **Z** ➔ X (~10K SNPs)
4. Z_0 ➔ Y



- Compute a "lod score" (a base-10 log-likelihood ratio) at sparse/few locations *j* across the chromosome.
  - ibd is slowly varying (relative to 10,188 SNPs)
  - maximization over variance parameters required at each location.
- In this example, we recover almost perfect ibd information, without use of any pedigree information.
- Earlier methods (ours and others') did NOT do well.

Pedigree vs Population prior

- Population model for P(Z) provides a prior:
  - Works because SNP data are highly informative
  - Does not provide a null model for testing
- Pedigree meiosis model also provides a P(Z):
  - Pedigree constraints give poor MCMC mixing
  - Pedigree does provide a null model for testing
- But does it ??
  - Ascertainment distorts ibd in causal regions
  - Selection (viability) distorts ibd in causal regions.
- Can we combine pedigree and population models to
  - Assess ascertainment biases
  - infer genome regions subject to selection?

Lod score ascertainment biases

- We imposed strong "ascertainment effect" by forcing segregation of causal DNA (Z_0) to three families.
  - results in high ibd mid-chromosome.
  - higher ibd gives higher likelihoods
- Kullback-Leibler information provides the expected lod score (over potential data Y) as a function of

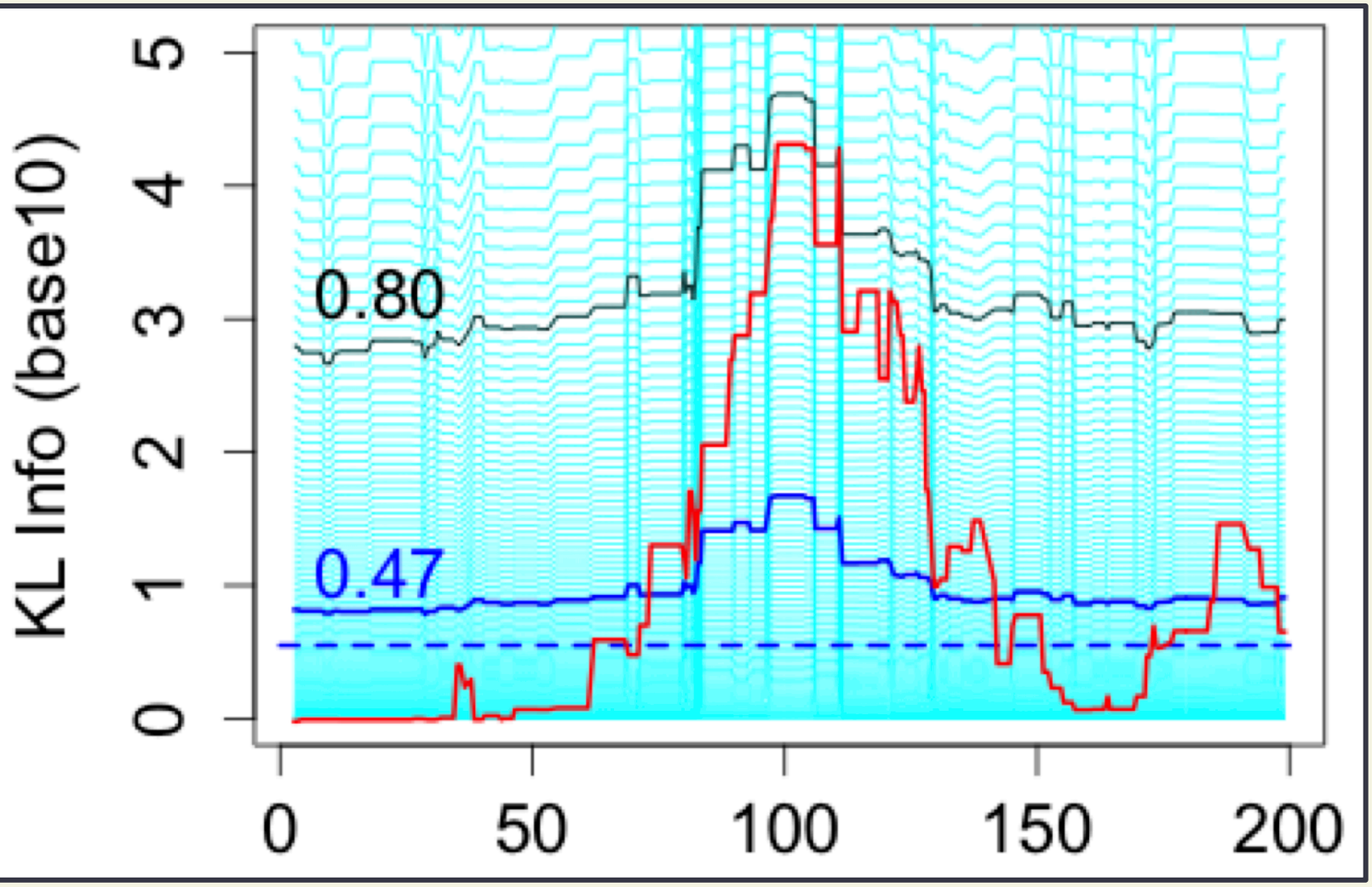
$$V = \mathrm{var}(\mathbf{Y}) = 2\,\Psi\,\sigma_a^2 + 2\,\Phi\,\tau^2 + \sigma_e^2\,\mathbf{I}$$

- Between two trait variances V^* and V:

$$0 \le E_{V^*}(\log L_Y(V^*) - \log L_Y(V)) =$$

$$\frac{1}{2}(\log(|V|/|V^*|)) + tr(V^* V^{-1}) - \dim(V))$$

Back to the simulation example



Lod score

$h^2 = 0.80$

Model truth
$h^2 = 0.47$

Solid: using Φ_j

Dashed: using Ψ

- KL given location-specific ibd, h^2 =0.01 to 0.99

Detecting inbreeding depression

- St Kilda – 4 islands in extreme NW Scotland

  (110 miles from mainland)
- Soay Sheep -- Primitive domestic breed
- 4000 years on Soay -- came with first human settlers
- >1000 yrs Vikings “Soay”
- 107 sheep moved to main Island, Hirta in 1932
- Hirta population studied since 1985
- Population fluctuates

  600 to 2100
- Eff pop size = 194

Levenish Dun Hirta Soay Boreray

Soay ewe

The Soay Sheep

- Data due to Josephine Pemberton, Sue Johnstone, and Jisca Huisman, Univ. Edinburgh.
  - Genotypes at 32,000+ SNPs across 26 autosome pairs
- Highly inbred, highly interrelated, but classic GRM and ROH methods did not give useful results.
- The data set: 596 M-F-O trios (in connected partial pedigree) – total 1101 animals.
- Can we estimate parental relatedness (population model)?
- From surviving offspring:
  - Can we detect inbreeding depression?
  - Can we detect recessive lethals?

Infer location-specific parental kinship and offspring autozygosity, and compare at locations across genome.

Combining population and pedigree

- Need to analyze the three members of each trio jointly (pairwise analysis does not work well).
- For unphased genetic marker data we can reduce from 203 states to 66, but only some of the 66 are permitted for a M-F-O trio
- We do not have 6 “exchangeable” genomes: Transitions in state are different between M-F

  from the one-generation step M-O and F-O.
- Need to combine population and pedigree ibd models
- That is: population model for the M-F ibd.

  then add segregation from parents to offspring.

The M-F-O-trio genotypic HMM

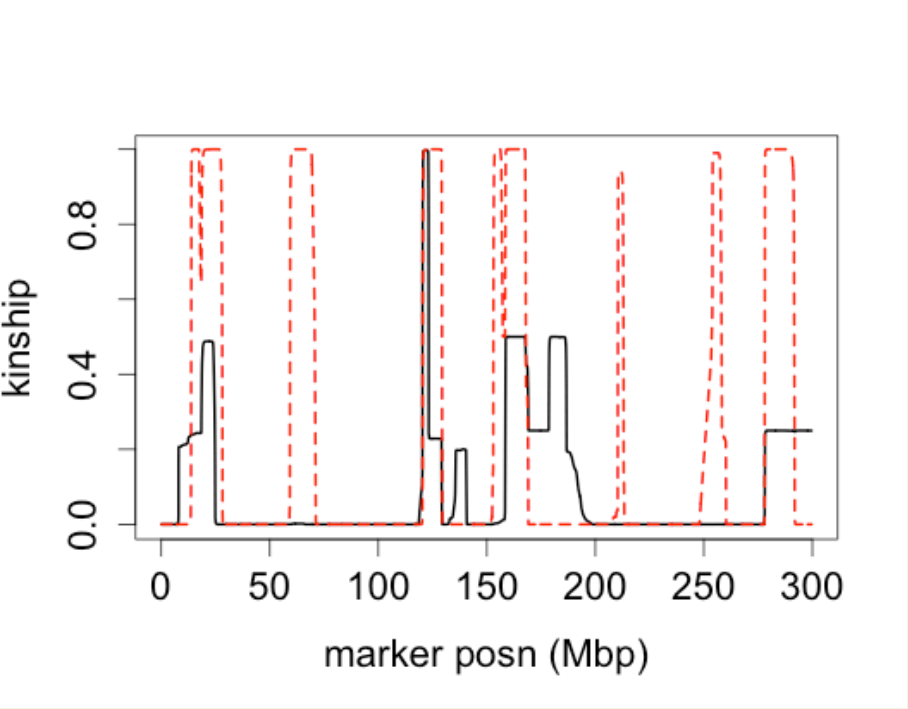
M	F	O	ϕ(F,M)	f(O)
11	11	11	1	1
11	12	11	1/2	1
11	21	12	1/2	0
11	22	12	0	0
11	23	12	0	0
12	11	11	1/2	1
12	12	11	1/2	1
12	13	11	1/4	1
12	21	12	1/2	0
12	22	12	1/2	0
12	23	12	1/4	0
12	31	13	1/4	0
12	32	13	1/4	0
12	33	13	0	0
12	34	13	0	0

- M-F states modeled according to population model
- Offspring receives first mat/pat parental DNA
- Recombinants parents to kid, become switches in parental chromosomes.
- No information on parental phase (no LD)
- New P(X|Z) for unphased trio genotypes given ibd state.

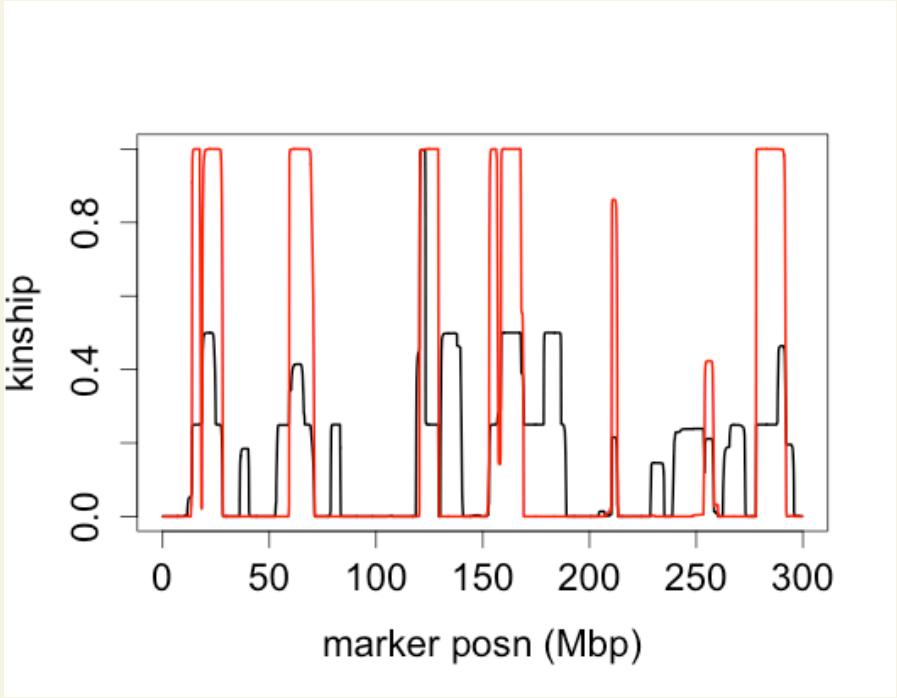
M-F kinship vs O-autozygosity

- Example: 1 trio, chromosome-1 inferences; 3610 SNPs.
- Black line—parental kinship; red –offspring inbreeding

Joint; no constraints:



Joint; assuming M-F-O trio:

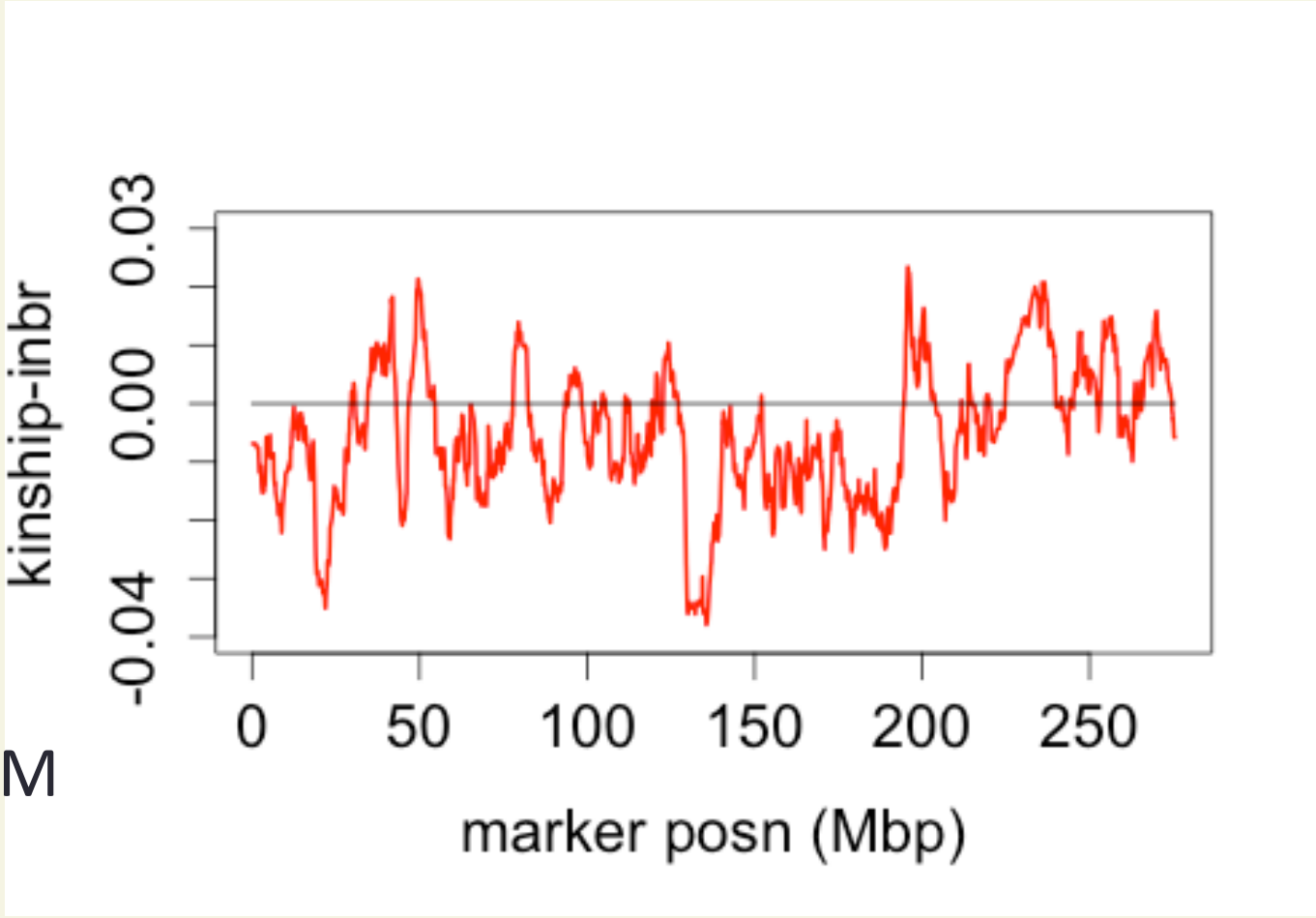


- Clearly using the parental constraint improves result.
  - Offspring information modifies parental IBD -- 60 Mbp
  - Parental constraint modifies Offspring IBD -- 255 Mbp

ParentKinship - inbreeding: (ϕ – f)

- Compare location-specific kinship, ϕ, of parents with autozgosity, f, of offspring.

- 150 secs CPU on laptop.
- Over 596 trios: mean f > mean ϕ. NS
- At 140Mbp is centromere.
- At 230 Mbp ϕ > f over 20cM
- Significant??



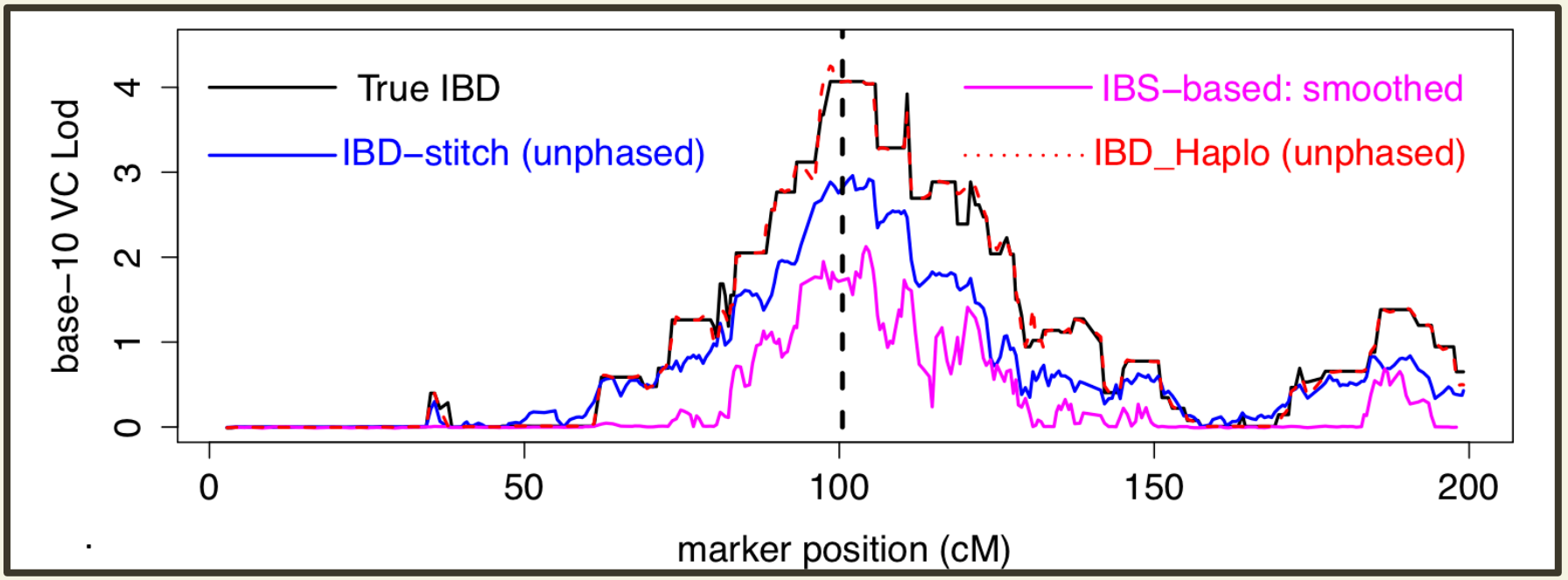
Summary

- Old: To infer ibd from marker data need a model using segmental properties of ibd to combine information from multiple SNPs.
- Old: ibd underlies genetic associations, and can be used in genetic mapping. Trait likelihoods can be based on realizations of ibd inferred from common SNP variants using a population model using no pedigree information.
- However: There are biases in realized ibd:
  - 1. Due to ascertainment
  - 2. Due to selection (e.g. inbreeding depression)
- Without a constraining pedigree, the level of inferred ibd varies widely: the lod score may reflect only ibd level.
- New: The KL information may be computed and provides a normalization for the lod score that adjusts for ibd.
- New: Combining pedigree and population models may enable location-specific selection to be detected.

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. (ibd haplo – HMM for pairs)
- Donnelly, K. P. (1983) The probability that related individuals share some section of genome identical by descent. Theor. Pop.Biol. 23: 34—63.
- Johnston S.E., Bérénos C., Slate J. and Pemberton, J.M. (2016). Conserved genetic architecture underlying individual recombination rate variation in a wild population of Soay sheep (*Ovis aries*). Genetics 203: 573—581.
- Leutenegger, A., Prum, B., Genin, E., Verny, C., 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.
- Raffa J. D. and Thompson E. A. (2016) Power and effective study size in heritability studies, Statistics in Biosciences, 8: 264-283.

Example: it works!



Simulated data: Z_0 → **Z**, then **Z** → X, and Z_0 → Y;
a simple example, provides proof of principle:

- Black = "true" (lod score if we knew the true ibd)
- Magenta: first pairwise method (due to Day-Williams et al.)
- Blue: Chris Glazner's multi-individual ibd-based PAC method
- Red dashed – current HMM pairwise method

– simple models are sometimes best!