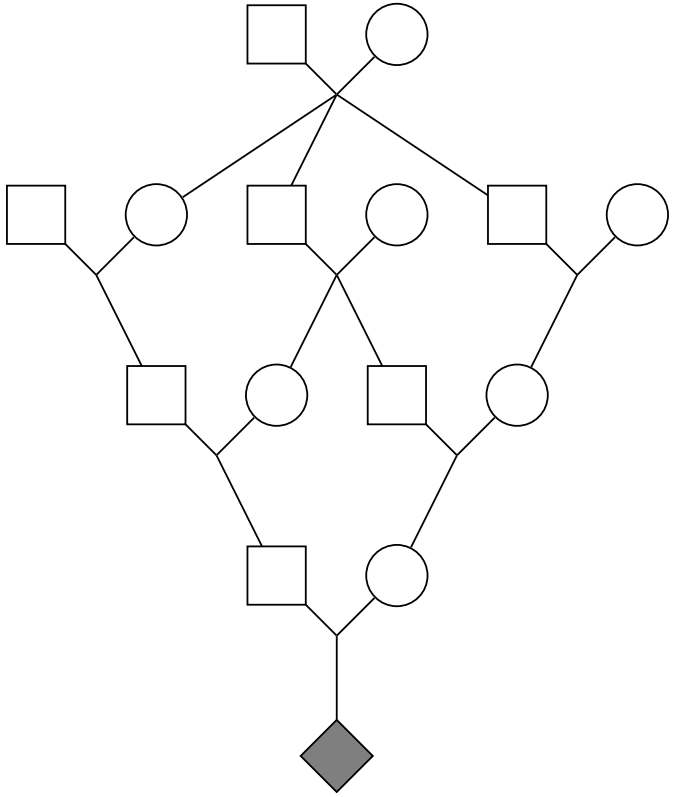
MCMC for the analysis of genetic data on pedigrees: Tutorial Session 1

Elizabeth Thompson
University of Washington

- Inheritance and the descent of genes in pedigrees.
- Computation of probabilities on pedigrees.
- MCMC sampling of inheritance patterns in pedigrees.

A PEDIGREE



Individuals have unique identifiers.

To specify pedigree: specify parent identifiers of each individual.

Founders have unspecified parents: others are non-founders.

Notation: male, female, affected or other data.

GENE IDENTITY BY DESCENT (*ibd*)

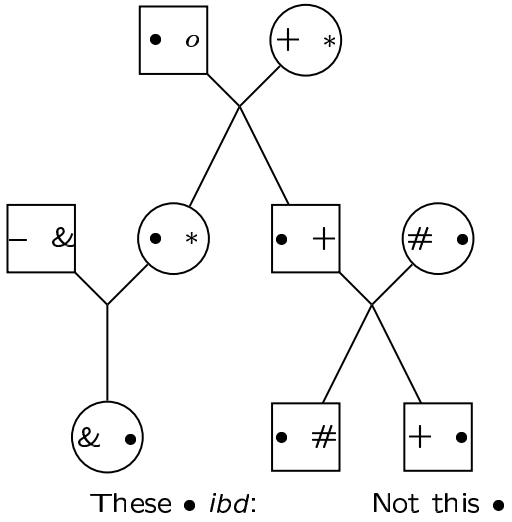
• Human individuals are diploid: every cell contains 2 copies of the human genome: one maternal, the other paternal.

•MENDEL's FIRST LAW (1866)
Each parent individual segregates a randomly chosen one of its two genes to each offspring, independently to each offspring. (All meioses are independent.)

• Genes that are copies of the same gene in a recent* common ancestor are said to be identical by descent (*ibd*).
*: *ibd* is defined relative to given pedigree or time point.

• **Simple model:** *ibd* genes are of the same allelic type, non-*ibd* genes are of independent types.

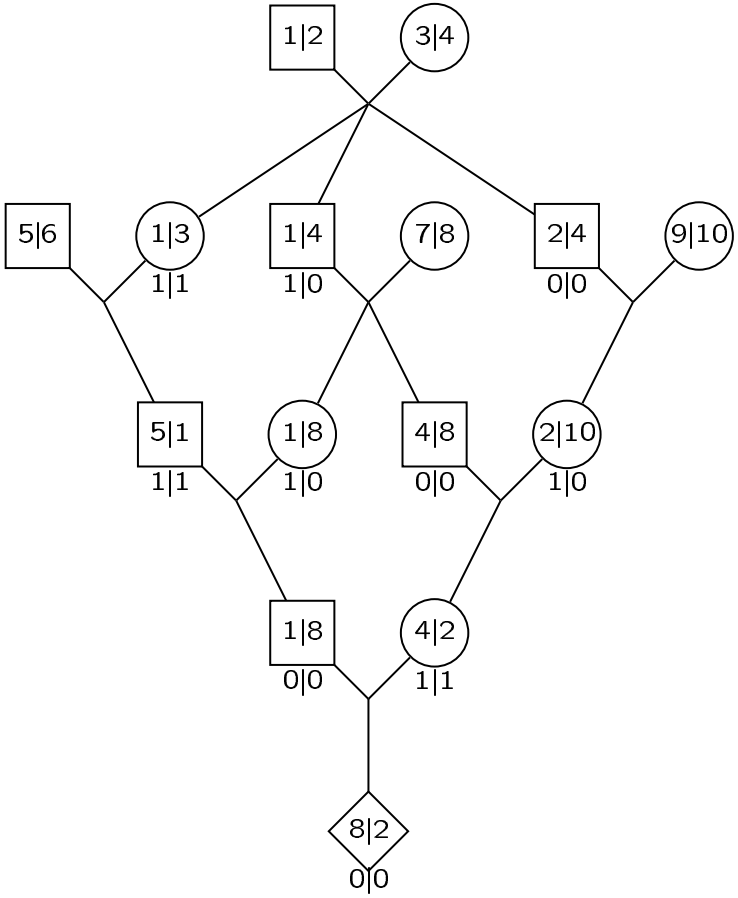
GENE IDENTITY BY DESCENT (*ibd*)



Given I have blood type O, the probability my cousins have blood type O is increased, because with some probability they share genes *ibd* with me at this locus.

- RELATIVES ARE SIMILAR because they have *ibd* genes.
- MENDELIAN GENETICS APPLIES TO MARKERS.

THE INHERITANCE OF GENOME



Label the two haploid genomes of every founder: Founder genome labels (FGL).

Inheritance of FGL:

$$S_{i,j} = 0 \text{ or } 1$$

as in meiosis i at locus j the maternal or paternal gene (respectively) of the parent is transmitted to the offspring.

INHERITANCE AT A SET OF LOCI

. . . is fully specified by *meiosis indicators*:
For loci j, $j = 1, \ldots, L$,

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

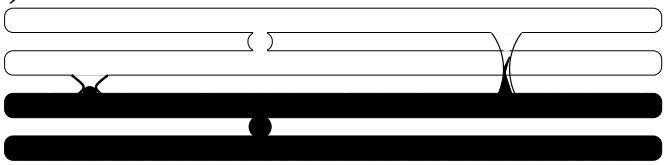
Notation:

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

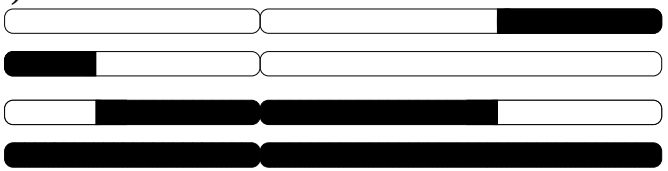
where m is the number of meioses in the pedigree, and L the number of loci along the chromosome.

CHROMOSOMES AND MEIOSIS

(a)



(b)



$\rho(d) = \Pr(\text{recomb.}|d) \nearrow d$

$\rho(0) = 0,\ \rho(\infty) = \frac{1}{2}$

Chromosomes duplicate align and exchange material.

Offspring chromosome consists of segments of two parental chromosomes.

Between two points: recombination $\exists$ if DNA is from the 2 different parental chromosomes.

A MODEL FOR LATENT INHERITANCE $\mathbf{S} = \{S_{i,j}\}$

- Meioses i are independent: $S_{i,\bullet}$ are independent, a priori.
- Mendel's First Law: $\Pr(S_{i,j} = 0) = \Pr(S_{i,j} = 1) = 1/2$
- Recombination: $\Pr(S_{i,j-1} \neq S_{i,j}) = \rho_{j-1}$ same for all i?– no

$$\Pr(S_{\bullet,j} \mid S_{\bullet,j-1}) = \rho_{j-1}^{R_{j-1}}(1-\rho_{j-1})^{m-R_{j-1}}$$

where $R_{j-1} = (\#i : S_{i,j} \neq S_{i,j-1})$

- No genetic interference: $\Pr(S_{i,j}|\mathbf{S}_{-(i,j)}) = \Pr(S_{i,j}|S_{i,j-1}, S_{i,j+1})$

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

A MODEL FOR THE ALLELIC TYPES OF FGL

Allelic types of the FGL are nuisance variables (in most contexts) which we need to marginalize over.

Model: 1. Loci j are independent– good model for $\rho > 0.005$
2. At locus j, each FGL g has type k independently with prob $q_{j,k}$.

$$\Pr(\mathcal{A}_j) \;=\; \prod_g q_j(g) \;\;=\;\; \prod_g q_{j,k}^{n_j(k)}$$

where $n_j(k)$ is number of FGL g with type k at locus j.
(For (2), better models are possible.)

STRUCTURE OF A GENETIC MODEL

• Population model: parameters $\mathbf{q}$, provide probabilities for latent $\mathcal{A}$– allelic types of FGL at each j

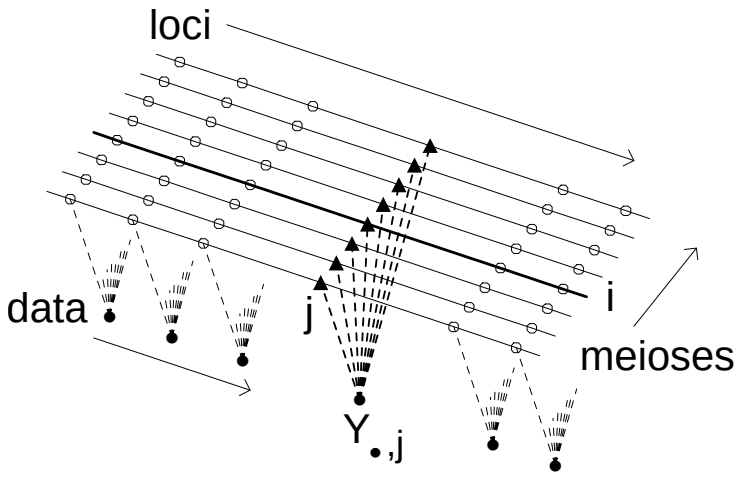
• Inheritance model: parameters ρ, provide probabilities for latent $\mathbf{S}$– inheritance of FGL at j, jointly over j.

• Individual genotypes $\mathbf{G}$ is deterministic function of $(\mathbf{S}, \mathcal{A})$

• penetrance model, parameters β relates $\mathbf{G}$ (and perhaps observable covariates) to observable data $\mathbf{Y}$

• $\xi = (\mathbf{q}, \theta, \beta)$.

STATISTICAL VIEW OF A PEDIGREE



Pedigree structure is implicit in the labelling of the meioses.

$Y_{\bullet,j}$: trait or marker data at locus j — trait specific to locus
$\Pr(Y_{\bullet,j} \mid S_{\bullet,j})$ can be computed – next we talk about this.

PEELING COMPUTATIONS: THREE CASES

(1) Peeling along a chromosome (HMM): $O(4^m)$

$$\Pr(\mathbf{Y}^{(\mathbf{j})}, \mathbf{S}_{\bullet,\mathbf{j}}) = \sum_{S_{\bullet,j-1}} \Pr(\mathbf{Y}^{(\mathbf{j-1})}, \mathbf{S}_{\bullet,\mathbf{j-1}}) \begin{matrix} \Pr(S_{\bullet,j}|S_{\bullet,j-1}) \\ \Pr(Y_{\bullet,j}|S_{\bullet,j}) \end{matrix}$$

Requires sequential summation along the chromosome.

(2) Pedigree peeling at single locus: $O(K^{3L})$, K large.

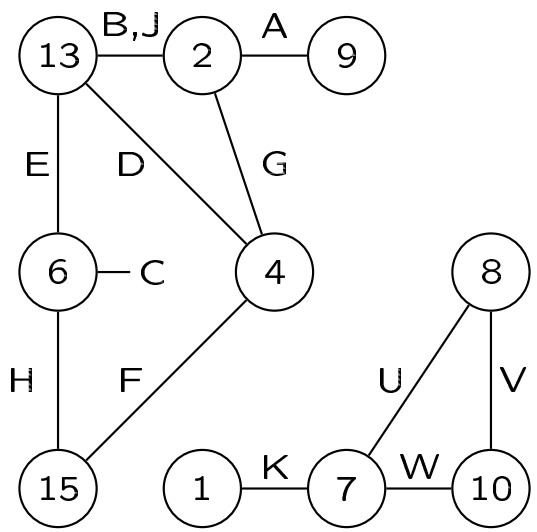
$$\Pr(Y_{\bullet,j}) = \sum_{S_{\bullet,j},\mathcal{A}_j} \Pr(Y_{\bullet,j}|S_{\bullet,j},\mathcal{A}_j)\Pr(S_{\bullet,j})\Pr(\mathcal{A}_j)$$

Involves sequential summation over the pedigree structure.

(3) For both we need

$$\Pr(Y_{\bullet,j}|S_{\bullet,j}) = \sum_{\mathcal{A}_j} \Pr(Y_{\bullet,j}|S_{\bullet,j},\mathcal{A}_j)\Pr(\mathcal{A}_j)$$

THE FGL-GRAPH STRUCTURE: FOR $\Pr(Y_{\bullet,j}|S_{\bullet,j})$



Only genes in "observed" matter– e.g. FGL 3, 11 ... are not present.

Only "connected" genes are dependent– lines connect the FGL of observed individuals.

$$\begin{aligned}
&\Pr(Y_{\bullet,j}|S_{\bullet,j}) \\
= \;& \textstyle\sum_{\mathcal{A}_j} \Pr(Y_{\bullet,j}|\mathbf{G}(\mathbf{S_{\bullet,j}}, \mathcal{A}_\mathbf{j}))\Pr(\mathcal{A}_\mathbf{j}) \\
= \;& \textstyle\sum_{\mathcal{A}_j} \left(\prod_n \Pr(Y_{n,j}|G_n(S_{\bullet,j}, \mathcal{A}_j))\right) \\
& \left(\textstyle\prod_g q_j(g)\right)
\end{aligned}$$

Nodes are FGL

Edges are observed individuals.

PEELING A COMPONENT OF THE FGL-GRAPH

$$\Pr(Y_{\bullet,j}|S_{\bullet,j}) = \sum_{\mathbf{g}} \left(\prod_n \Pr(Y_n|g_{n,1}, g_{n,2}) \right) \begin{array}{c} q(g_2)q(g_6)q(g_4) \\ q(g_9)q(g_{13})q(g_{15}) \end{array}$$

$$= \sum_{g_6} q(g_6)\Pr(Y_C|g_6) \quad \left(\sum_{g_{15}} q(g_{15})\, \Pr(Y_H|g_{15}, g_6) \right.$$

$$\left(\sum_{g_4} q(g_4)\Pr(Y_F|g_{15}, g_4) \right. \quad \left(\sum_{g_{13}} q(g_{13})\Pr(Y_E|g_6, g_{13})\Pr(Y_D|g_4, g_{13}) \right.$$

$$\left(\sum_{g_2} q(g_2)\Pr(Y_B|g_2, g_{13}) \right. \quad \Pr(Y_J|g_2, g_{13})\Pr(Y_G|g_2, g_4)$$

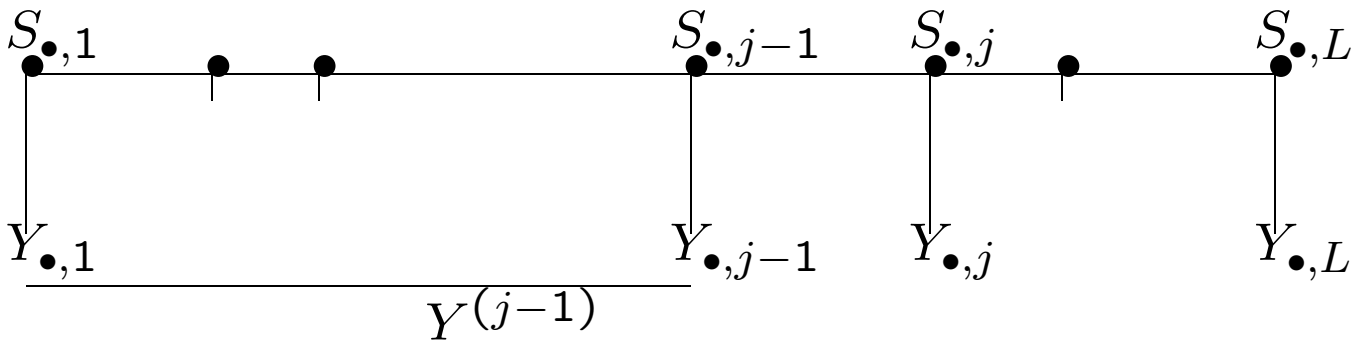
$$\left. \left. \left(\sum_{g_9} q(g_9)\Pr(Y_A|g_2, g_9) \right) \right) \right) \right) \right)$$

BAUM ALGORITHM FOR HMM: Lander-Green

For data observations $\mathbf{Y} = (\mathbf{Y_{\bullet,j}, j = 1, \ldots, L})$, we want to compute $\Pr(\mathbf{Y})$.

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

$$= \sum_{\mathbf{S}} \left(\Pr(S_{\bullet,1}) \prod_{j=2}^{L} \Pr(S_{\bullet,j} \mid S_{\bullet,j-1}) \prod_{j=1}^{L} \Pr(Y_{\bullet,j} \mid S_{\bullet,j}) \right)$$



BAUM ALGORITHM DETAILS

Forwards Baum: Define $R_1^*(s) = \mathsf{Pr}(S_{\bullet,1} = s)$ and $Y^{(j)} = (Y_{\bullet,1}, \ldots, Y_{\bullet,j})$, the data up to locus j, so

$$\begin{aligned} R_j^*(s) &= \mathsf{Pr}(Y_{\bullet,k}, k = 1, \ldots, j-1,\ S_{\bullet,j} = s) \\ &= \mathsf{Pr}(Y^{(j-1)},\ S_{\bullet,j} = s) \end{aligned}$$

Now

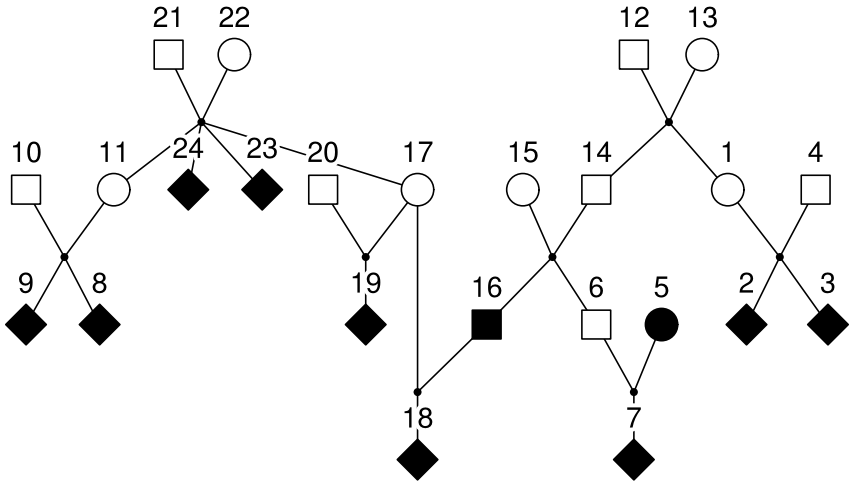
$$R_{j+1}^*(s) = \sum_{s^*} \Big[\mathsf{Pr}(S_{\bullet,j+1} = s \mid S_{\bullet,j} = s^*) \\ \mathsf{Pr}(Y_{\bullet,j} \mid S_{\bullet,j} = s^*)\ R_j^*(s^*)\Big]$$

for $j = 1, 2, \ldots, L-1$, with

$$L = \mathsf{Pr}(\mathbf{Y}) = \sum_{s^*} \mathsf{Pr}(Y_{\bullet,L} \mid S_{\bullet,L} = s^*)\ R_L^*(s^*)$$

$S_{\bullet,j}$ can take 2^m values, where m is number of meioses. Computation is limited to small pedigrees.

PEDIGREE PEELING: Elston-Stewart

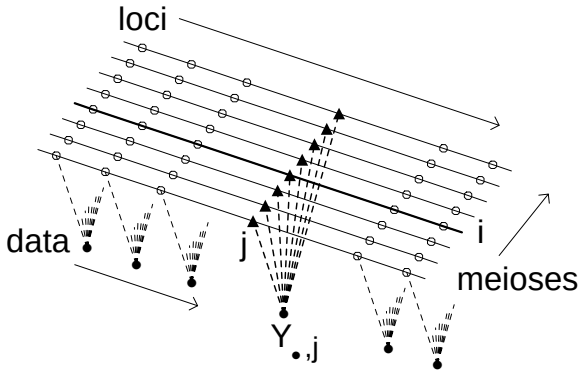


Conditional on genotypes of parents, grandparent couples and all offspring are all mutually independent.
Accumulate probabilities over pedigree, using genotypes of (cut-set of) individual(s) as latent state space.
Linear in pedigree size. Exponential in number of loci.
(Do with $\mathbf{S}$ not genotypes; do by "ordered genotype" not genotype.)

MCMC FOR LARGE PEDIGREES WITH MULTUPLE LOCI

• If pedigree (m) and number of loci (L) are large, we cannot chromosome-peel or pedigree-peel.

• We can still compute Pr($Y_{\bullet,j}|S_{\bullet,j}$) by FGL-peeling, quickly and easily (relatively).



The HMM structure lends itself to several block-Gibbs schemes, each updating a subset S_u of $\{S_{i,j}\}$ conditional on rest (S_f) and $\mathbf{Y}$.

BLOCK GIBBS SAMPLERS FOR $\mathbf{S} \sim \Pr(\cdot|\mathbf{Y})$

L-sampler: requires peeling over the pedigree to resample $S_{\bullet,j}$.

$$\Pr(Y_{\bullet,j} \mid S_{\bullet,j-1}, S_{\bullet,j+1}) = \sum_{S_{\bullet,j}} \Pr(Y_{\bullet,j} \mid S_{\bullet,j}) \Pr(S_{\bullet,j} \mid S_{\bullet,j-1}, S_{\bullet,j+1})$$

M-sampler: requires peeling along the chromosome (Baum algorithm) using $\Pr(S_{\bullet,j}|Y_{\bullet,1}, ..., Y_{\bullet,j})$.

$$\Pr(\{S_{i,\bullet}; i \in I^*\} \mid \mathbf{Y}, \{\mathbf{S}_{\mathbf{i}',\bullet}; \mathbf{i}' \notin \mathbf{I}^*\})$$

L-sampler mixes poorly for tight linkage
M-sampler mixes poorly on extended pedigrees.
L-sampler is irreducible (theoretically).
Together (LM-sampler) they can do better.