MCMC for the analysis of genetic data on pedigrees:

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

University of Washington
Elizabeth Thompson
Individuals have unique identifiers.

To specify pedigree: specify pair of identifiers of each individual.

Notation: male, female.

Founders have unspecified parents; others are non-founders.

Affected or other data.
Human individuals are diploid: every cell contains 2 copies of the same genome. One maternal, the other paternal. 

Mendel's First Law (1866) - Mendel's First Law (1866): Segregation of an individual's identical homologous pair of chromosomes (alleles) is independent of each other.

Each parent individual segregates a randomly chosen one of its two genes to each offspring, independently of each other. 

- Simple model: \( ibd \) genes are of the same allelic type, non-\( ibd \) genes are of independent types.

\( ibd \) is defined relative to a given pedigree or time point. \( ibd \) genes that are copies of the same gene in a recent common ancestor are said to be identical by descent (ibd). All meioses are independent. 

\( ibd \) is defined relative to a given pedigree or time point. 

\( ibd \) 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 a given pedigree or time point. 

\( ibd \) genes that are copies of the same gene in a recent common ancestor are said to be identical by descent (ibd).
Given I have blood type O, the probability my cousins have blood type O is increased, because they have \( \text{idb} \) genes. Relatives are similar because they have \( \text{idb} \) genes.

- Mendelian genetics applies to markers.
- Relatives are similar because they have \( \text{idb} \) genes.

Gene identity by descent (\( \text{idb} \))
Label the two haploid genomes of every founder:

- Founder 1
- Founder 2

Label the two haploid genomes (FGL).

Inheritance of FGL:
- If the parent is transmitting a maternal or paternal gene (respectively) at locus \(j\) in meiosis \(i\), then the maternal or paternal gene (respectively) of the parent is transmitted to the offspring.

\[ S_{i,j} = 0 \text{ or } 1 \]
INHERITANCE AT A SET OF LOCI

Notation:

\[ I = 1 \quad \text{if gene at meiosis } j \text{ locus } j \text{ is parent's paternal gene.} \]
\[ I = 0 \quad \text{if gene at meiosis } j \text{ locus } j \text{ is parent's maternal gene.} \]

For loci \( j \) where \( I = 1 \), \( T, \cdots, T \) is fully specified by meiosis indicators.
Different parental chromosomes.

Between two points: Recombination.

Chromosomes duplicate and exchange material.

\[
\frac{\pi}{2} = (\infty)d \quad 0 = (0)d
\]

\[p \not\rightarrow (p|\text{recomb.}) \quad p_{\text{recomb}} = (p)d\]
\[ \prod_{i=1}^{\infty} \Pr(S_i) = \Pr(S) \]

Meioses are independent: \( \Pr(S_i) = \Pr(S_i | S) = \) prior. 

A MODEL FOR LATENT INHERITANCE

\{\{S_i\}\} = \text{A MODEL FOR LATENT INHERITANCE}
Allelic types of the FGL are nuisance variables (in most contexts) which we need to marginalize over. (For (2), better models are possible.)

$$\Pr(A_j) = \prod_{\ell} P_{\ell}^{b_j} = \prod_{\ell}^{b_j} q_{nj}(k) = \prod_{\ell} q_{nj}(k)$$

where $u_j^{\ell} = \sum_{\ell} (g_j^{\ell})$ is number of FGL $g$ with type $k$ at locus $j$.

Model 1: Loci $j$ are independent—good model for $d < 0.005$

Model 2: At locus $j$, each FGL $g$ has type $k$ independently with prob $q_{nj}(k)$ with prob $q_{nj}(k)$

A Model for the Allelic Types of FGL
STRUCTURE OF A GENETIC MODEL

\( y \) observable covariates (to observable data and perhaps \( \theta \) population model, parameters \( q \), provide probabilities for latent allelic types of \( F \) at each \( j \)).

\( \bar{\theta} \) relates \( G \) is deterministic function of \( (S,A) \).

\( \mathcal{A} \) - Inheritance model: parameters \( p \) provide probabilities for latent allelic types of \( F \) at each \( j \).

\( \mathcal{Y} \) - Inheritance model: parameters \( g \) provide probabilities for latent allelic types of \( F \) at each \( j \).

\( \mathcal{Z} \) - Inheritance model: parameters \( \rho \) provide probabilities for latent allelic types of \( F \) at each \( j \).
Pedigree structure is implicit in the labelling of meioses.

The meioses.

\[ \Pr(Y_{ij} | S_{ij}) \] can be computed — next we talk about this.

Pr(Y_{ij} | S_{ij}) trait or marker data at locus \( j \) — trait specific to locus \( j \)
(1) Peeling along a chromosome (HMM):

\[
\Pr(Y^j | \lambda_{-j}, S^j_{-j}) = \sum_{i} \Pr(Y^j | \lambda_{-j}, S^j_{-j}, i_{-j}^i) \Pr(i_{-j}^i)
\]

Requires sequential summation along the chromosome.

(2) Pedigree peeling at single locus: \(O(K^3_T)\), \(K\) large.

\[
\Pr(Y^j | \lambda_{-j}, S^j_{-j}) = \sum_{i} \Pr(Y^j | \lambda_{-j}, S^j_{-j}, i_{-j}^i) \Pr(i_{-j}^i)
\]

Involves sequential summation over the pedigree structure.

(3) For both we need

\[
\Pr(Y^j_{-j} | \lambda_{-j}, S^j_{-j}) = \sum_{i} \Pr(Y^j_{-j} | \lambda_{-j}, S^j_{-j}, i_{-j}^i) \Pr(i_{-j}^i)
\]

Requires sequential summation over the pedigree structure.
Nodes are FGL

Edges are observed individuals.

FGL of observed individuals.

Dependent lines connect the "connected" genes are not present. E.g. FGL 3, 11 are not observed. 

Only genes in "observed"
Peeling a component of the FGL-graph
For data observations we want to compute \( \lambda \).

---

BAUM ALGORITHM FOR HMM: Lander-Green
Computation is limited to small pedigrees. $S^\ast \downarrow_j$ can take $2^m$ values, where $m$ is number of meioses.

\[
(\ast s)^{T^\ast \downarrow_j}(\ast s = T^\ast S \mid T^\ast \lambda)^{P\downarrow_j} \sum_s = (\lambda)^{P\downarrow_j} = T
\]

For $\ell = 1, 2, \ldots, T - 1$, with

\[
(\ast s)^{f^\ast \downarrow_j}(\ast s = f^\ast S \mid f^\ast \lambda)^{P\downarrow_j}
\]

\[
(\ast s = f^\ast S \mid s = 1 + f^\ast S)^{P\downarrow_j} \sum_s = (s)^{1 + f^\ast \downarrow_j}
\]

Now

\[
(s = f^\ast S \mid (1 - f^\ast) \lambda)^{P\downarrow_j} =
\]

\[
(s = f^\ast S \mid 1 - f^\ast \lambda)^{P\downarrow_j} = (s)^{f\downarrow_j}
\]

Forwards Baum: Define $R^\ast = \{s \mid (f^\ast \lambda, \ldots, f^\ast \lambda) = (f) \lambda$, the data up to locus $\ell$, so

$R^\ast = \{s \mid (f^\ast \lambda, \ldots, f^\ast \lambda) = (f) \lambda$, and

(s) = (s)^{P\downarrow_j}$

BAUM ALGORITHM DETAILS

NUS-I: Mar 2004
Condition on genotypes of parents, grandparents 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.
If pedigree size and number of loci are large, we cannot peel chromosome or pedigree-peel easily. We can still compute Pr(Y | S_i, f) by FGL-peeling, quickly and easily (relatively). If pedigree size (m) and number of loci (I) are large, we cannot.

MULTIPLE LOCI MCMC FOR LARGE PEDIGREES WITH...
Together (LM-sampler) they can do better.

L-sampler is irreducible (theoretically).

M-sampler mixes poorly on extended pedigrees.

L-sampler mixes poorly for tight linkage.

\[ \Pr(\mathbf{Y} \mid \mathbf{S}) = \sum_{\mathbf{S}} \Pr(\mathbf{S} \mid \mathbf{Y}) \Pr(\mathbf{S}) \]

M-sampler: Requires peeling along the chromosome (Baum algorithm) using \( \Pr(\mathbf{S} \mid \mathbf{Y}) \).

L-sampler: Requires peeling over the pedigree to resample \( \mathbf{S} \).

\[ \Pr(\mathbf{Y} \mid \mathbf{S}) \sim \Pr(\cdot | \mathbf{S}) \]