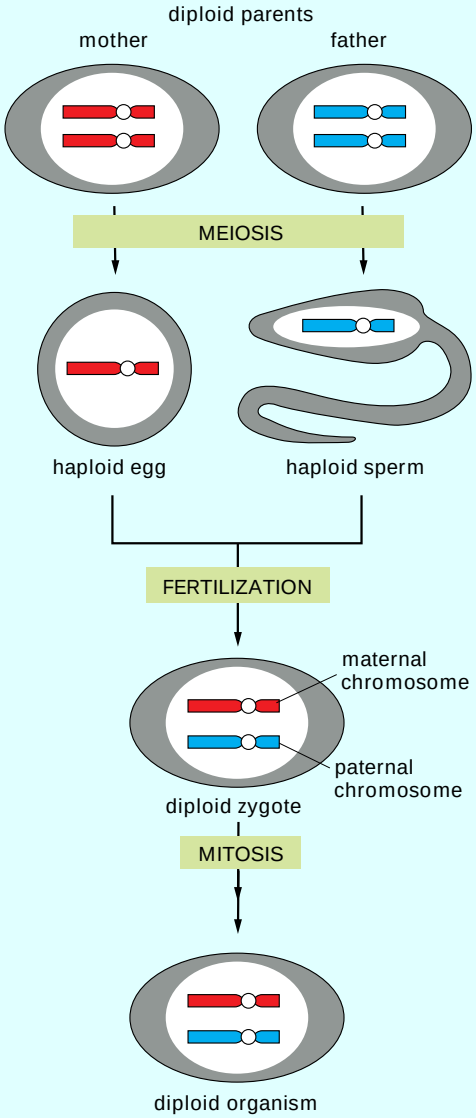
Untangling sexual reproduction – a problem in ecological genetics

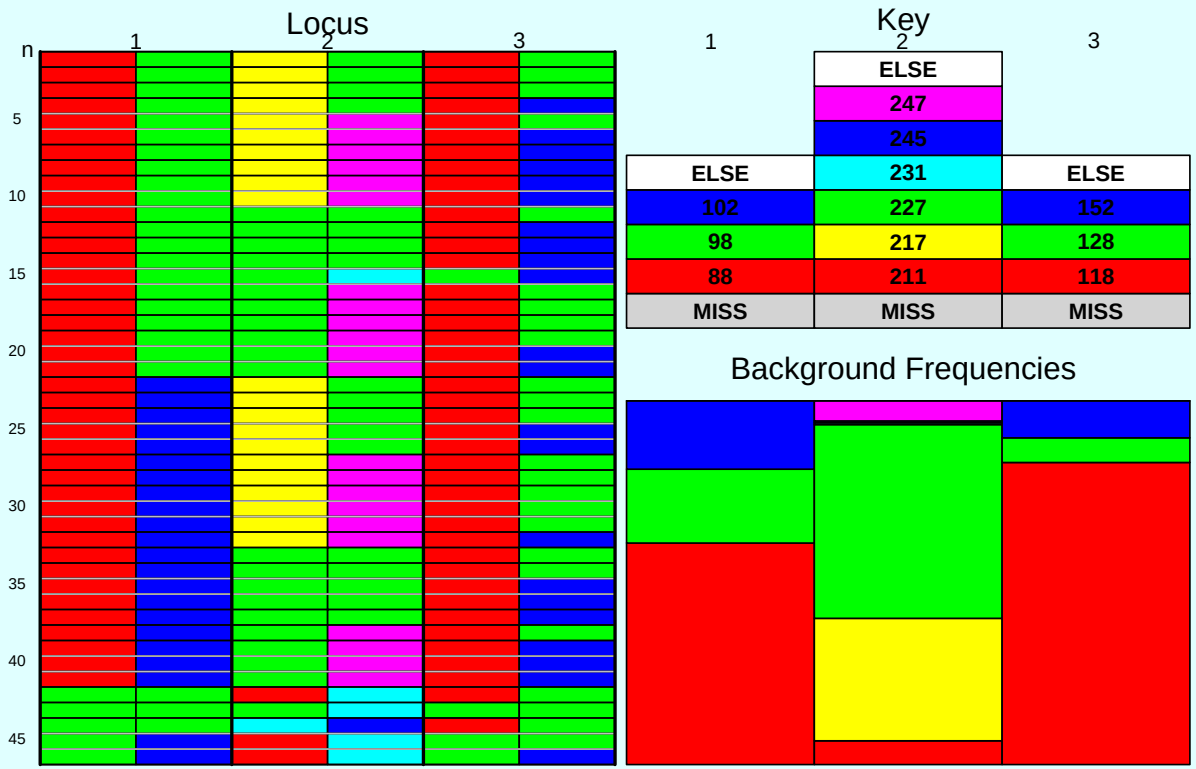
Ian Wilson
University of Aberdeen

9 July 2004

- **Motivation & problem description**
- Modelling sexual reproduction as a mixture problem
- Restoring sibling relationships
- Problems and extensions


diploid parents
mother
father
MEIOSIS
haploid egg
haploid sperm
FERTILIZATION
maternal chromosome
paternal chromosome
diploid zygote
MITOSIS
diploid organism

But sexual reproduction is seldom straightforward . . .


Locus
n
1
2
3
5
10
15
20
25
30
35
40
45
Key
1
2
3
ELSE
247
245
ELSE
231
ELSE
102
227
152
98
217
128
88
211
118
MISS
MISS
MISS
Background Frequencies

Why do we want to know relationships?

- Estimation of heritability from natural populations
- Genotypes as markers in ecological studies
- Dispersal for extinction/recolonisation models
- Direct measurement of fecundity

Two examples:

- *Lologi Forbesi* – squid lays eggs in deep water off the West Coast. Want to estimate heritability in growth experiments (work with Aiden Emery and Les Noble).

- Blue Gill Sunfish – male nest guarding (data supplied by Bryan Neff).

Other Methods

- Parsimony: fit "by eye", reconstructing maternal genotype, and then then fit fathers.
- Simple clustering. Measure genetic distance between all pairs of individuals and cluster.
- Likelihood, but extremely difficult for all but the most simple problems - combinatorial explosion.

We have no idea about the uncertainty of our estimate of the relationships for any of the above.

Note that Kevin Dawson and Thomas & Hill have methods that are similar.

The generalised problem is to reconstruct relationships between individuals based on the individual data when we may have:

- Genotypes of (some or all) potential parents
- "Background" allele frequencies from a random sample of the population
- None, some or all relationships known
- "Mutation"

- Motivation & problem description
- **Modelling sexual reproduction as a mixture problem**
- Restoring sibling relationships
- Problems and extensions

Take a Bayesian "Kitchen Sink" Approach, augmenting variables to include the maternal and paternal genotypes, and index variables for all relationships.

An index for the mother and father is associated with each individual.

Each parent has a genotype.

Likelihood equations:

e.g Individal with genotype AB at a locus

Parent 1	Parent 2	Probability
AA	BB	1
AC	BB	$\frac{1}{2}$
AB	AB	
AC	BC	$\frac{1}{4}$
CC	CD	0

"Mutation" can be included trivially. Full likelihood calculated by multiplying over all loci for all individuals

Modelling Issues:

- How do we model the background allele frequency data – when we have alleles from the egg string not seen in the background?

- How do we model the relative reproductive success of different fathers/mothers?

Models for Number of Parents and Offspring Share

One possible prior for is a Poisson number of parents.

We treat the mother and father of any hatchling as independent – this will not be generally true.

To model this we need the joint density of the number of fathers and the patterns of shared paternity.

One possibility would be equal probability for any father, conditional on k fathers:

$$\mathsf{P}(\mathbf{a_m}, n_m) = n_m! \frac{\mathsf{P}(n_m)}{n_m} \tag{1}$$

A better model may be an multinomial number of offspring with an exchangeable Dirichlet prior on the share of paternity:

$$\mathsf{P}(\mathbf{a_m}, n_m) = n_m! \prod_{i=1}^{n_m} \Gamma(n_i + \alpha) \qquad (2)$$

Taking the limit as n_m gets large with αn_m constant gives the ESF. Also used in ecology for species assemblages (Lambshead).

We use an exchangeable dirichlet prior for the allele frequencies in the background. Prior probability of relative frequencies x

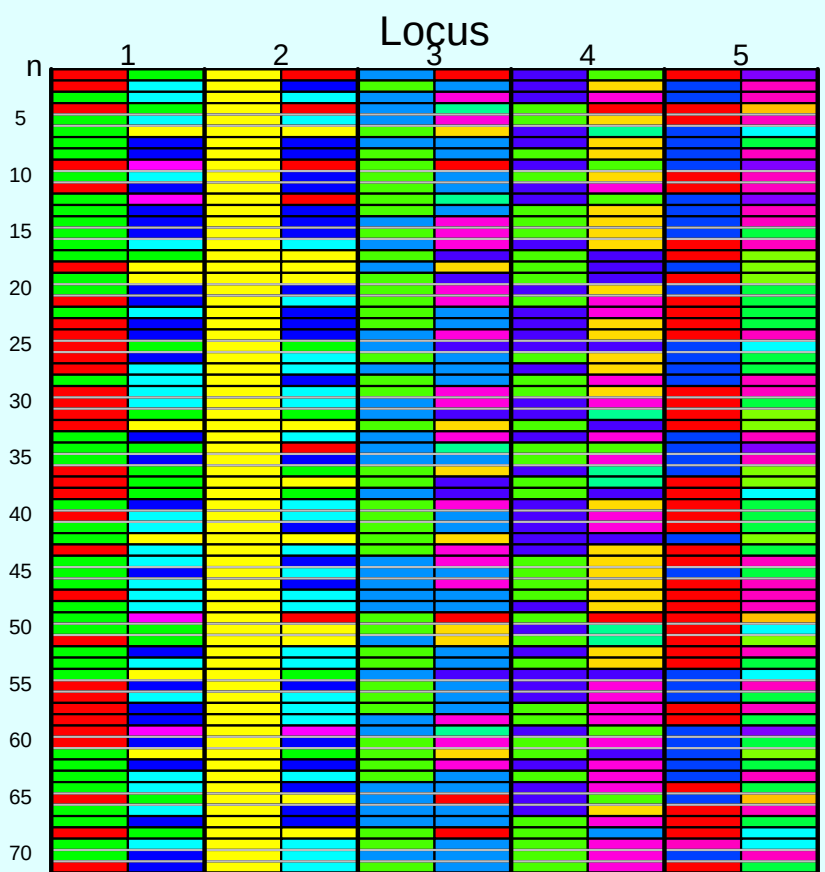
$$f(\mathbf{x}) \propto \prod_{i=1}^{k} x_i^{\lambda-1} \tag{3}$$

Then the posterior distribution of allele frequencies after observing the background data is Dirichlet distributed with parameters $b_i + \lambda$.

MCMCMC: A random scan was taken from:

- Gibbs sampling of allocation of females to offspring;
- Gibbs sampling of allocation of males to offspring;
- Gibbs sampling of parentage for both mothers and fathers;
- Gibbs sampling of maternal genotypes;
- Gibbs sampling of paternal genotypes; and
- Metropolis Hastings sampling of mutation rate, μ,
- Metropolis Hastings updates to the number of males or females by splitting/joining and resampling parental genotypes;
- Metropolis-Hastings updates of α and β the Ewens' sampling formula parameters and a
- Step to swap between chains – coldest chain an importance sampler.

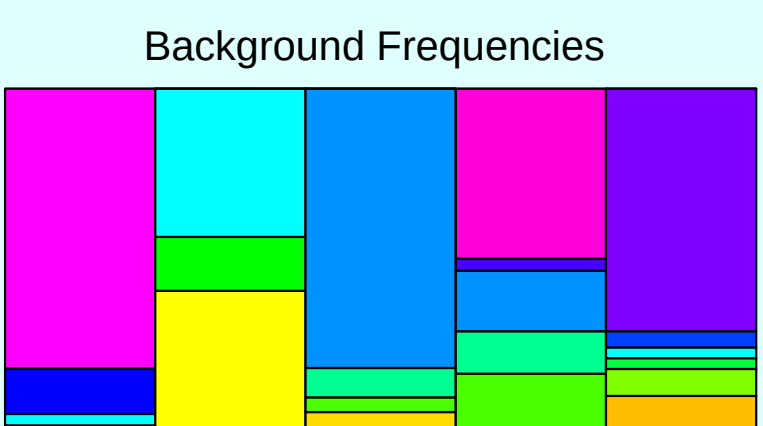
- Motivation & problem description
- Modelling sexual reproduction as a mixture problem
- **Restoring sibling relationships**
- Problems and extensions



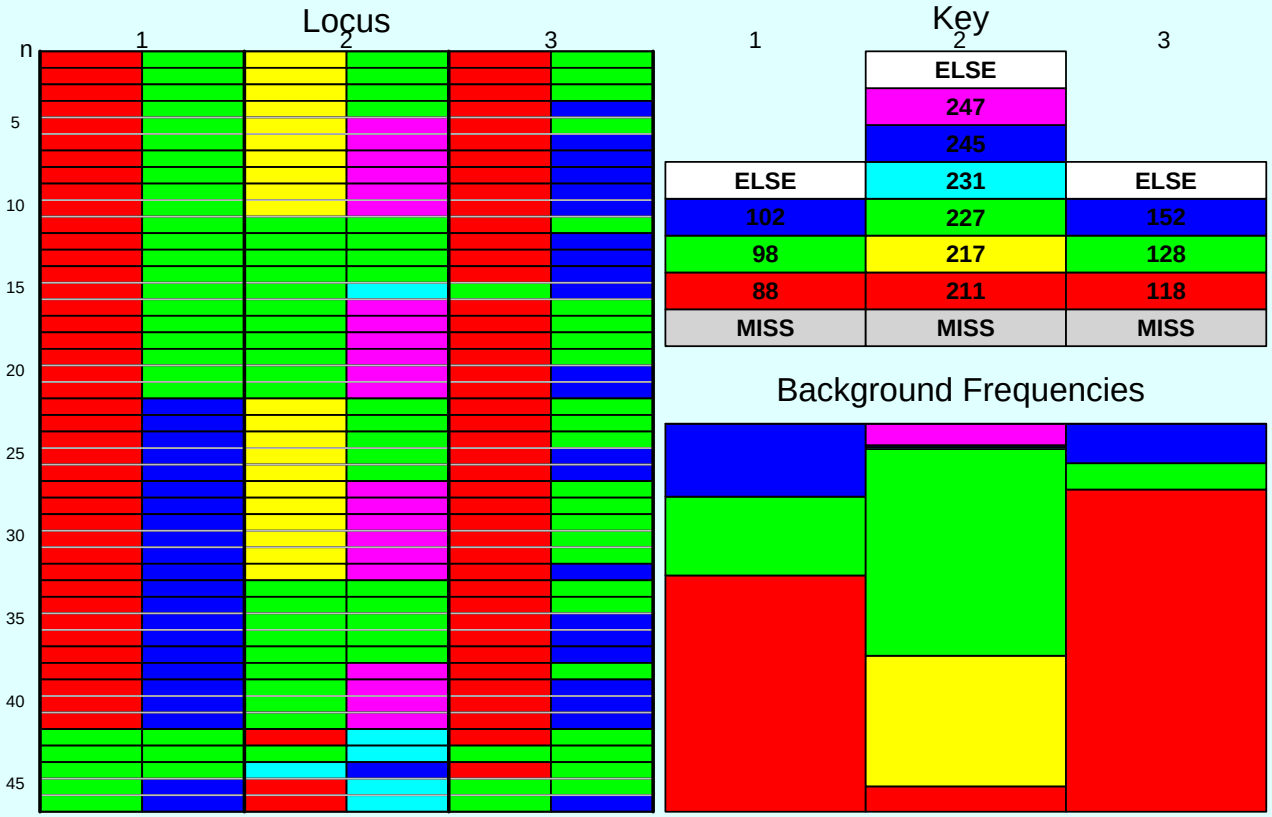
Key

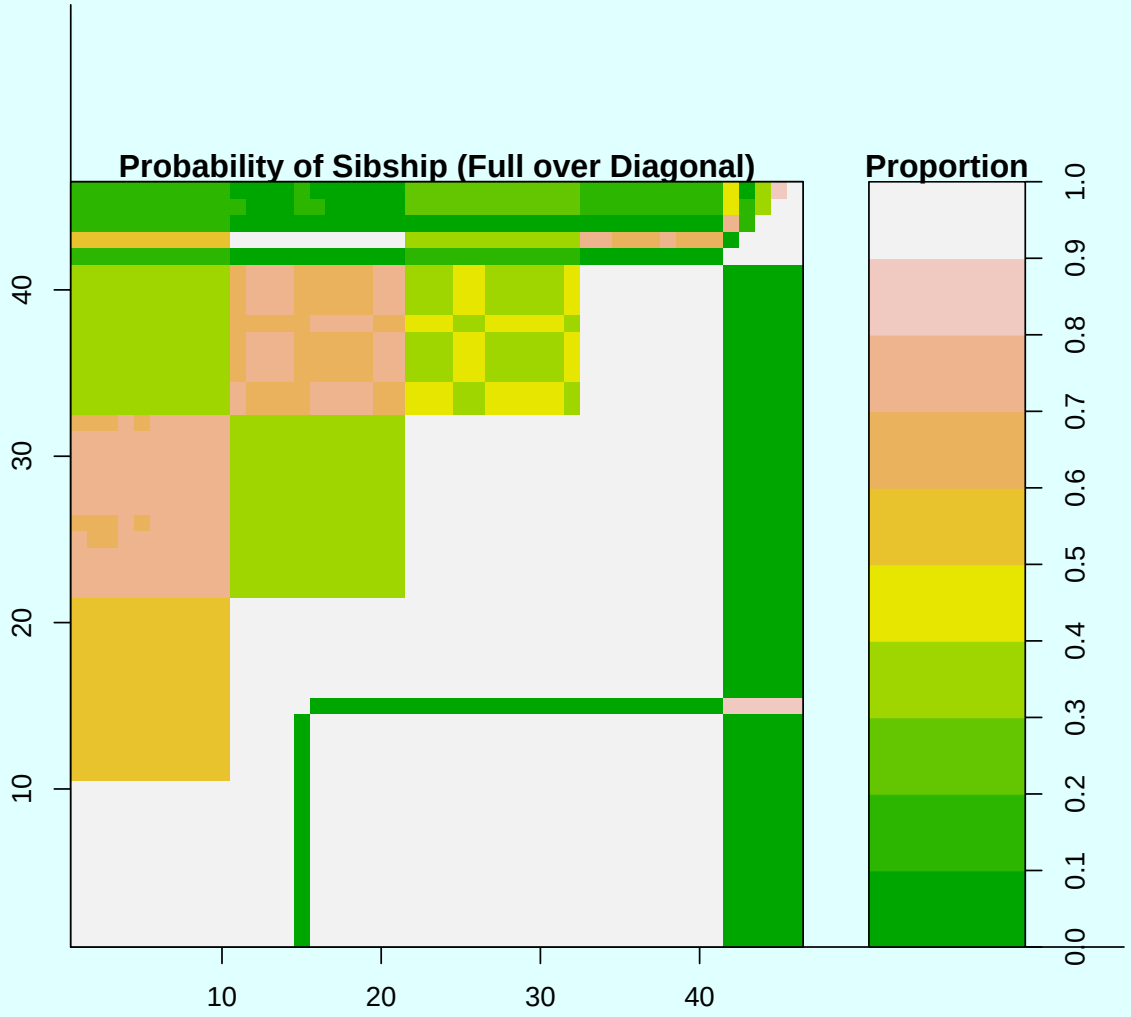
1	2	3	4	5
				ELSE
		ELSE	ELSE	294
ELSE	ELSE	131	251	292
234	126	122	247	278
231	111	118	243	276
228	109	112	241	274
222	107	110	237	272
212	93	108	235	263
206	90	104	233	256
MISS	MISS	MISS	MISS	MISS

Background Frequencies



		0	1	2	3	4	≥ 5
fathers	Model	0	0	0.0004	0.8598	0.1394	0.0004
	prior	0.4562	0.2560	0.1452	0.0744	0.0354	0.0328
mothers	Model	0.9994	0.0006	0	0	0	0
	Prior	0.9762	0.0234	0.0004	0	0	0
muts	Model	0.0004	0.7550	0.2312	0.0128	0.0006	0
	Prior	0.5032	0.2888	0.1314	0.0498	0.0166	0.0102


Locus
n
1
2
3
5
10
15
20
25
30
35
40
45
Key
1
2
3
ELSE
247
245
ELSE
231
ELSE
102
227
152
98
217
128
88
211
118
MISS
MISS
MISS
Background Frequencies


Probability of Sibship (Full over Diagonal)
Proportion
10
20
30
40
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

Simulation Results

Estimated Posterior Probabilities of Sibship for True Siblings

Loci	$\widehat{p} = 0$	$0 \leq \widehat{p} \leq 0.05$	$0.05 < \widehat{p} \leq 0.95$	$0.95 < \widehat{p} \leq 1$	$\widehat{p} = 1$
1	0	0.43	0.67	0	0
2	0	0.07	0.86	0.07	0
3	0	0	0.36	0.64	0
4	0	0	0.21	0.14	0.64
5	0	0	0.21	0.07	0.71

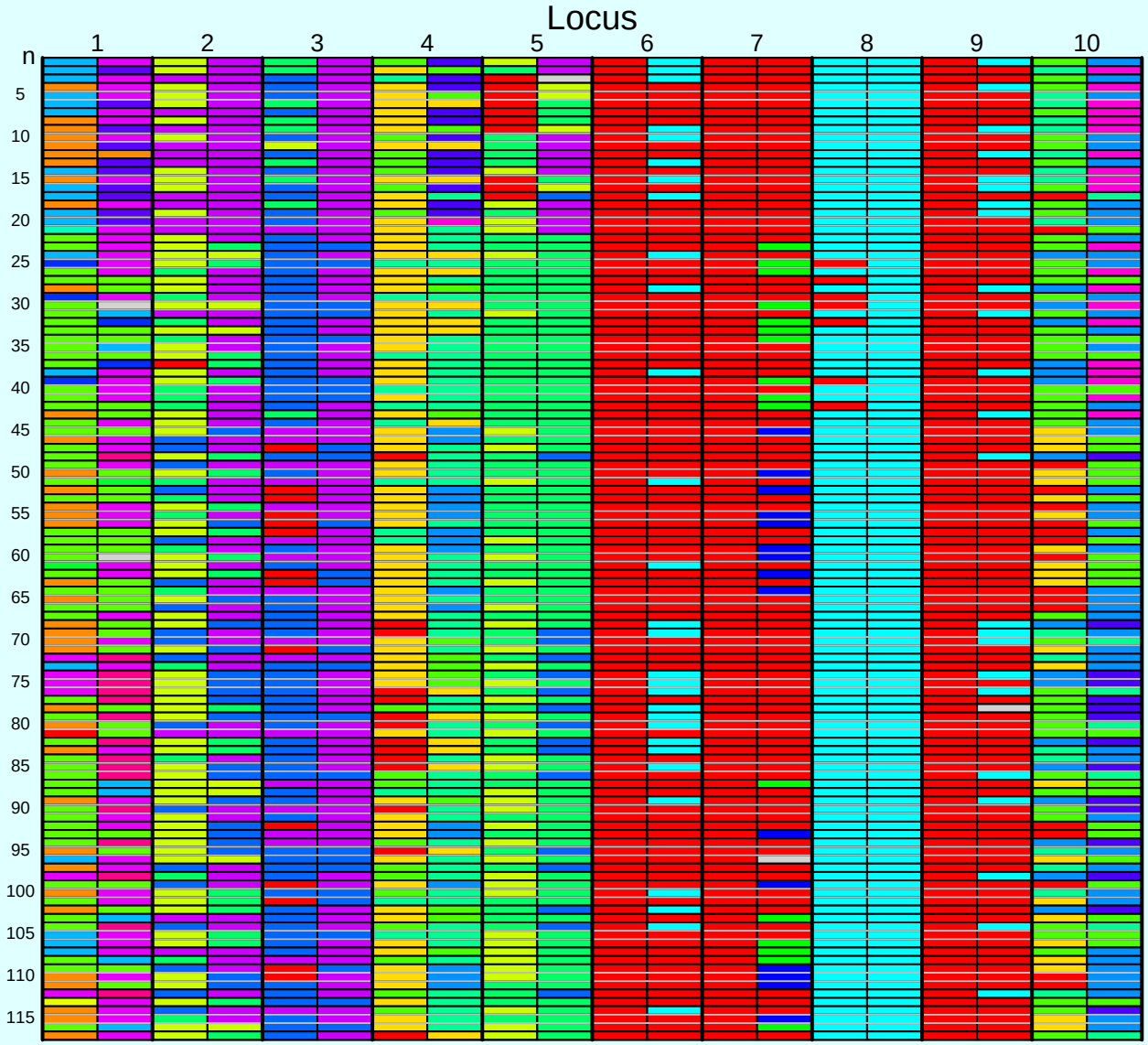
Estimated probabilities ($n = 1000$) of correct identification of full siblings for simulated data. Twenty individuals sampled with 5 mothers and 2 fathers. Each locus had 10 equally frequent alleles.

Estimated Posterior Probabilities of Sibship for Non-Siblings

Loci	$\widehat{p} = 0$	$0 \leq \widehat{p} \leq 0.05$	$0.05 < \widehat{p} \leq 0.95$	$0.95 < \widehat{p} \leq 1$	$\widehat{p} = 1$
1	0	0.48	0.52	0	0
2	0.20	0.54	0.26	0	0
3	0.84	0.13	0.03	0	0
4	0.97	0.02	0.01	0	0
5	0.98	0.01	0.01	0	0

Estimated probabilities ($n = 1000$) of incorrect identification of full siblings for simulated data. Twenty individuals sampled with 5 mothers and 2 fathers. Each locus had 10 equally frequent alleles.

- Motivation & problem description
- Modelling sexual reproduction as a mixture problem
- Restoring sibling relationships
- **Problems and extensions**


Locus
1
2
3
4
5
6
7
8
9
10
n
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
105
110
115

- The joint distribution of offspring number does not adequately capture the features of **real** data.

  In general there is an association between the identity of mother and father in most (all?) cases. I have an independent prior.

- Potential parents may be related.