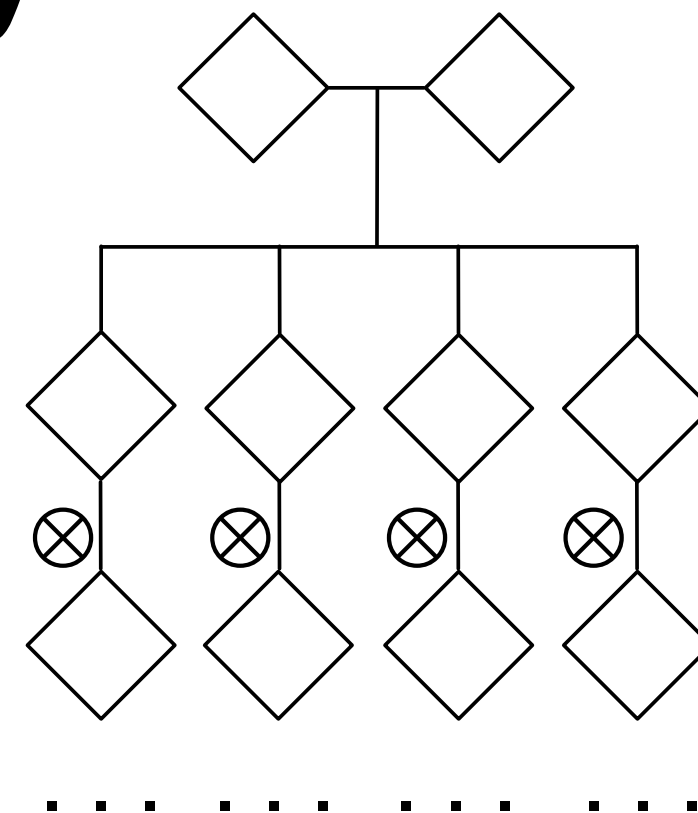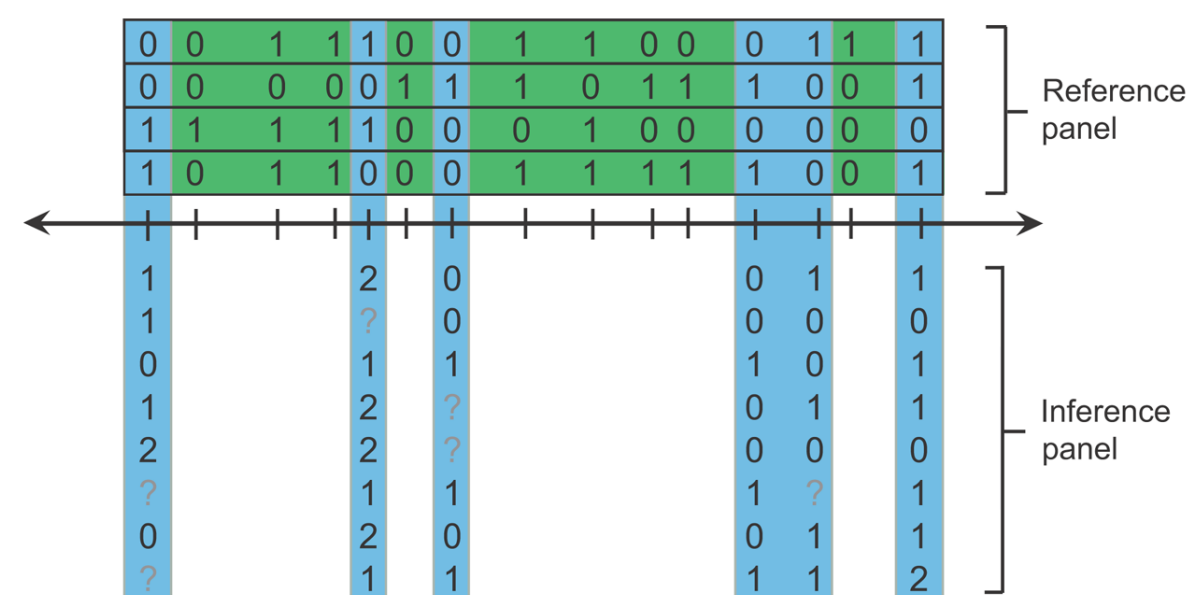# Informatics Interest Group Meeting

Tom Kono
2018-05-16

Terminology

- Pedigree: Known breeding structure of a population or family

- Imputation: Inferring unobserved genotypes in individuals

- Phasing: Assigning alleles to chromosomes (haplotypes)

0	0	1	1	1	0	0	1	1	0	0	0	1	1	1
0	0	0	0	0	1	1	1	0	1	1	1	0	0	1
1	1	1	1	1	0	0	0	1	0	0	0	0	0	0
1	0	1	1	0	0	0	1	1	1	1	1	0	0	1

Reference panel

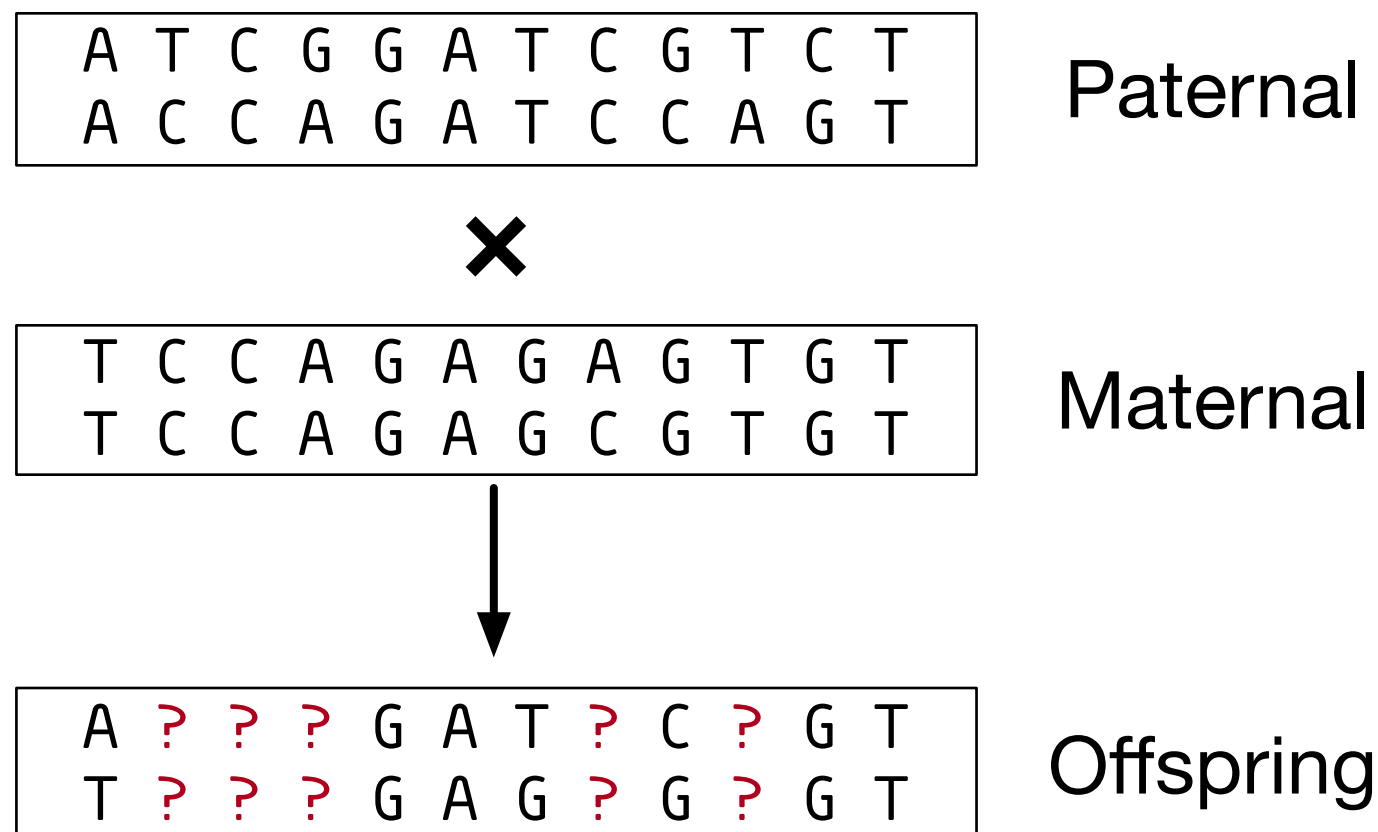Inference panel

T = SNPs typed in both panels

U = SNPs typed only in reference panel

Two Flavors of Imputation

- Pedigree-based: populations with known breeding or relationship structure. Has high power.

- Linkage disequilibrium-based: populations with mixed or unknown relationships. Requires huge amounts of data.

- AlphaPeel (will be discussed later) falls somewhat in between these two

Pedigree Imputation

- Takes advantage of *identity by descent* and the principles of transmission genetics

- Requires trios (or duos) - experimentally challenging for some systems

```
A T C G G A T C G T C T
A C C A G A T C C A G T
```
Paternal

✖

```
T C C A G A G A G T G T
T C C A G A G C G T G T
```
Maternal

↓

```
A ? ? ? G A T ? C ? G T
T ? ? ? G A G ? G ? G T
```
Offspring

Pedigree Imputation

- Takes advantage of *identity by descent* and the principles of transmission genetics

- Requires trios (or duos) - experimentally challenging for some systems
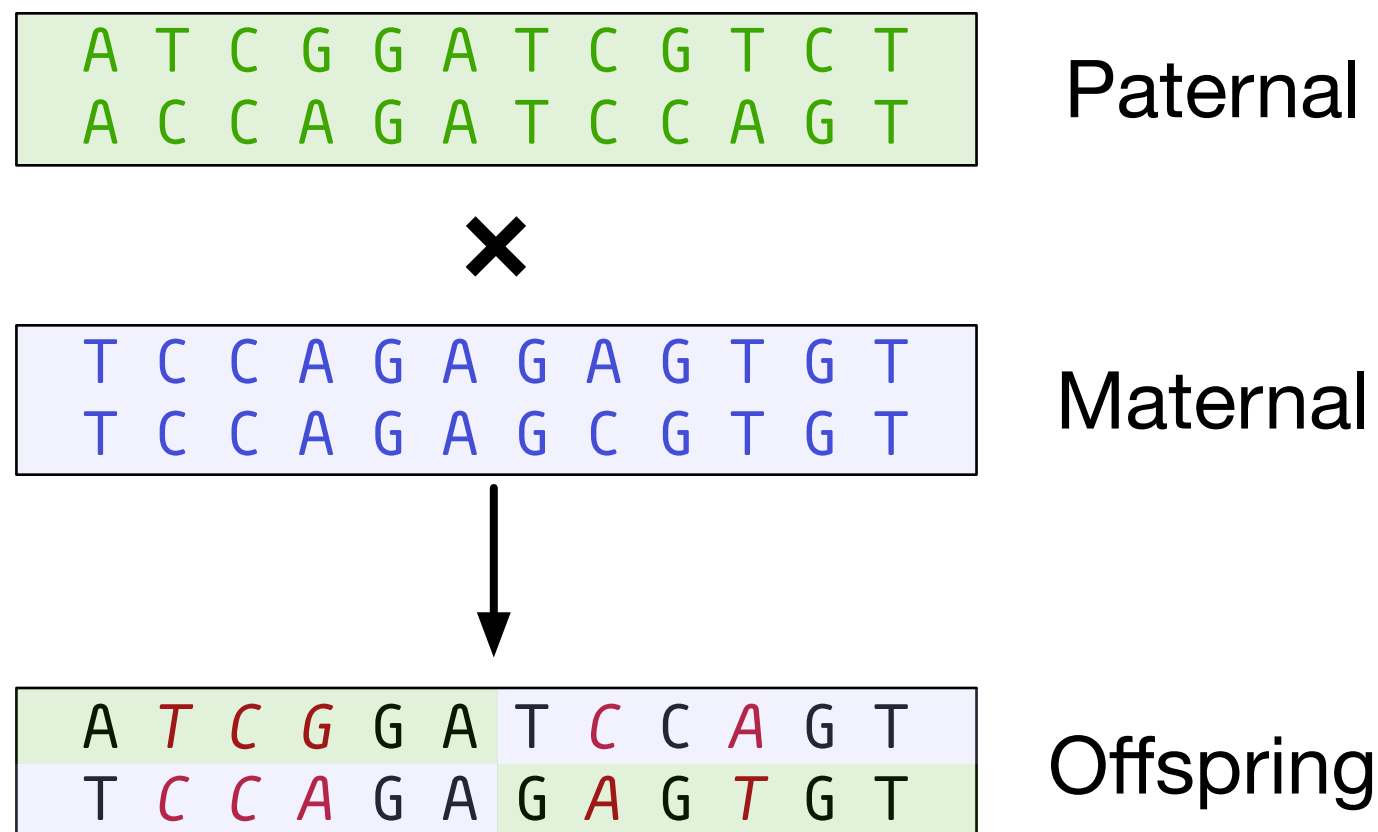
Population Imputation

- Imputation in individuals of *mixed* or *unknown* relatedness (many call this "unrelated")

- Similar principle to pedigree imputation, but uses linkage disequilibrium structure and *identity by state* rather than *descent*

- Instead of parental genotypes (or haplotypes), it uses a reference panel of haplotypes

- Requires computational phasing as an intermediate

Population Imputation

a Genotype data with missing data at untyped SNPs (grey question marks)

c Each sample is phased and the haplotypes are modelled as a mosaic of those in the haplotype reference panel

e The reference haplotypes are used to impute alleles into the samples to create imputed genotypes (orange)

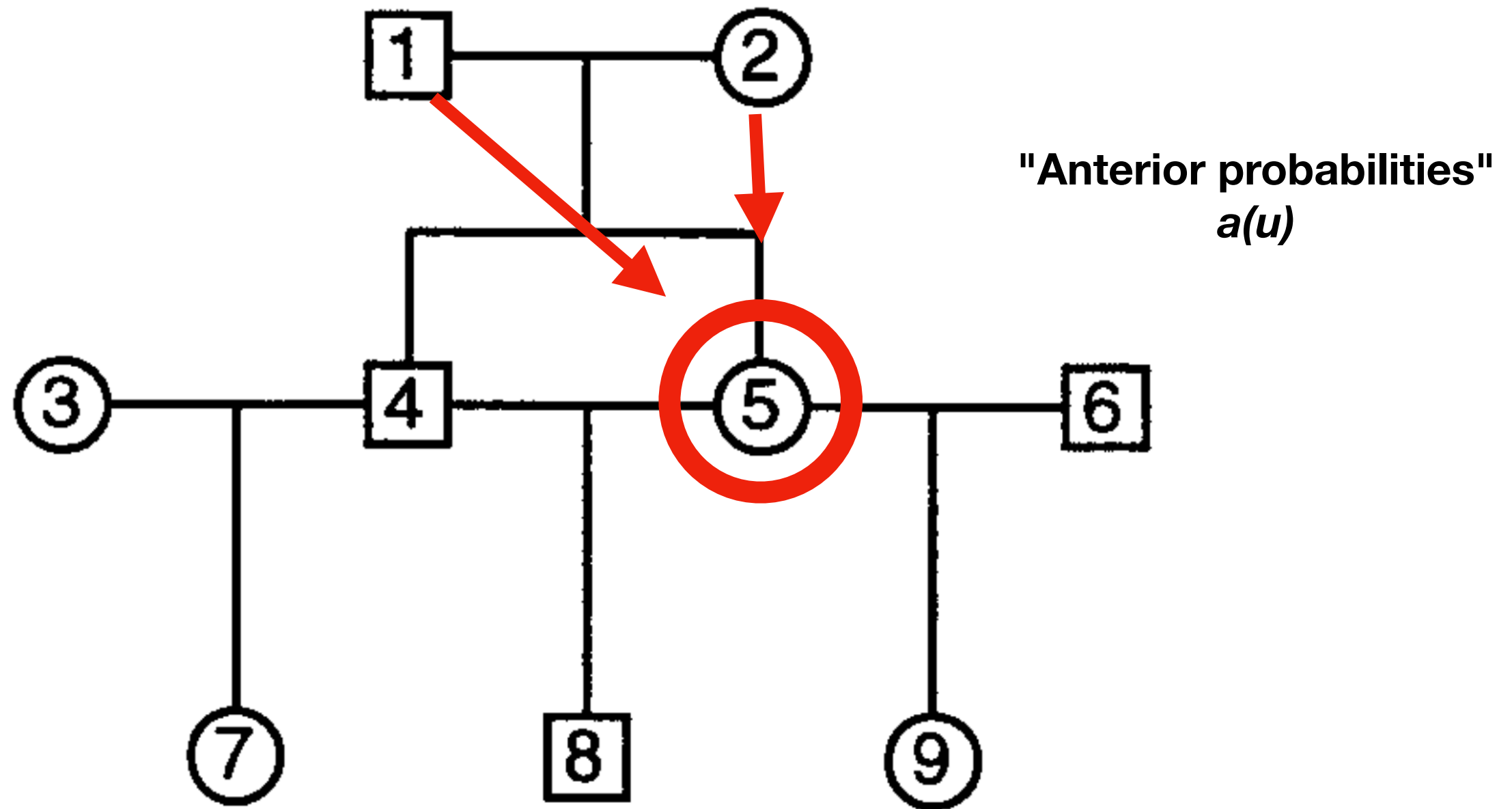Marchini and Howie, 2010, *Nature Rev Genet*

So, Peeling?

- Iterating or recursing over pedigree relationships

  - "peeling down" - moving from ancestors to descendants, calculating probabilities of transmitting alleles

  - "peeling up" - moving from descendants to ancestors, calculating probabilities of inheritance

- Neat because it can predict genotypes for totally unobserved individuals in the pedigree

- Can also predict phase
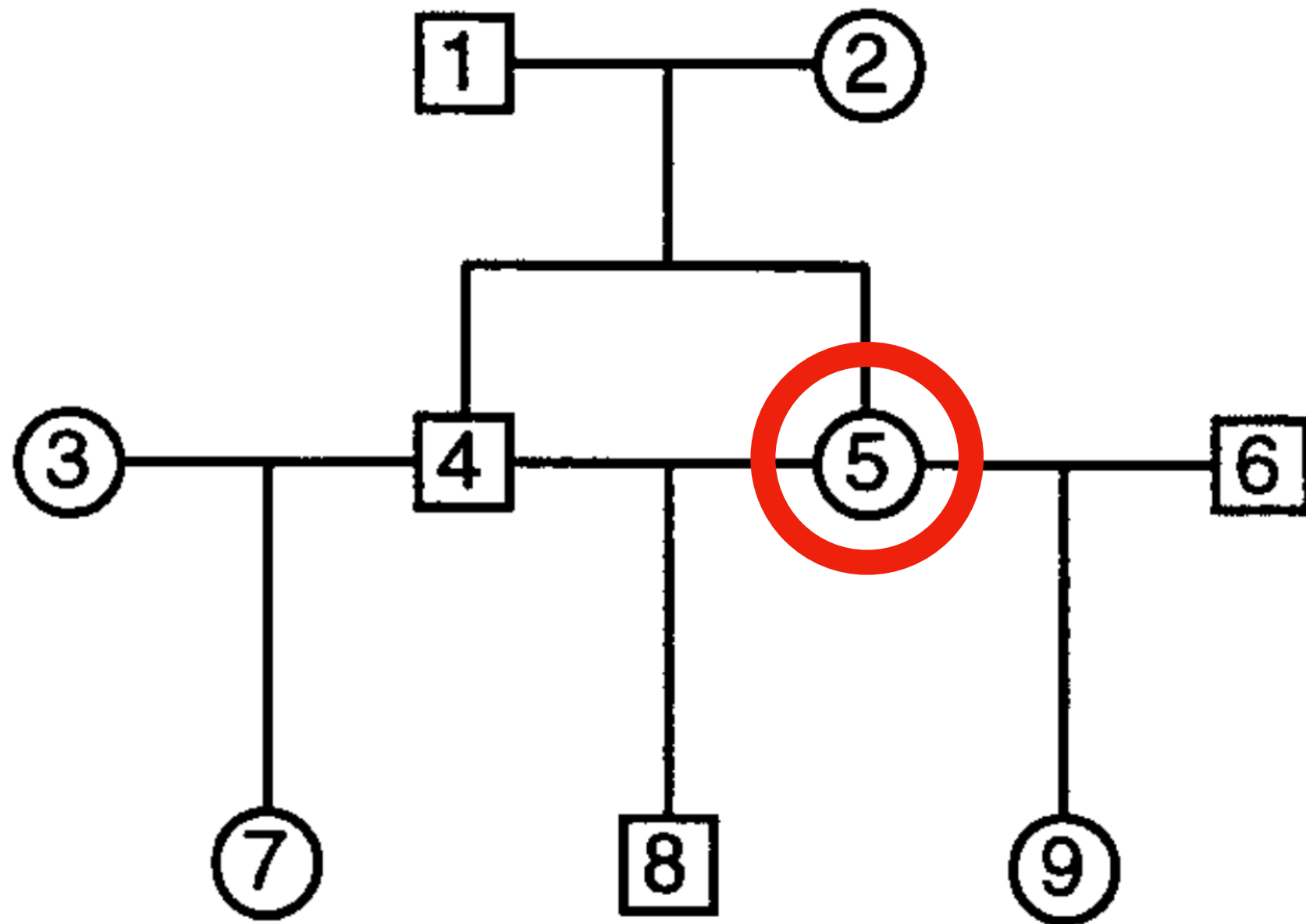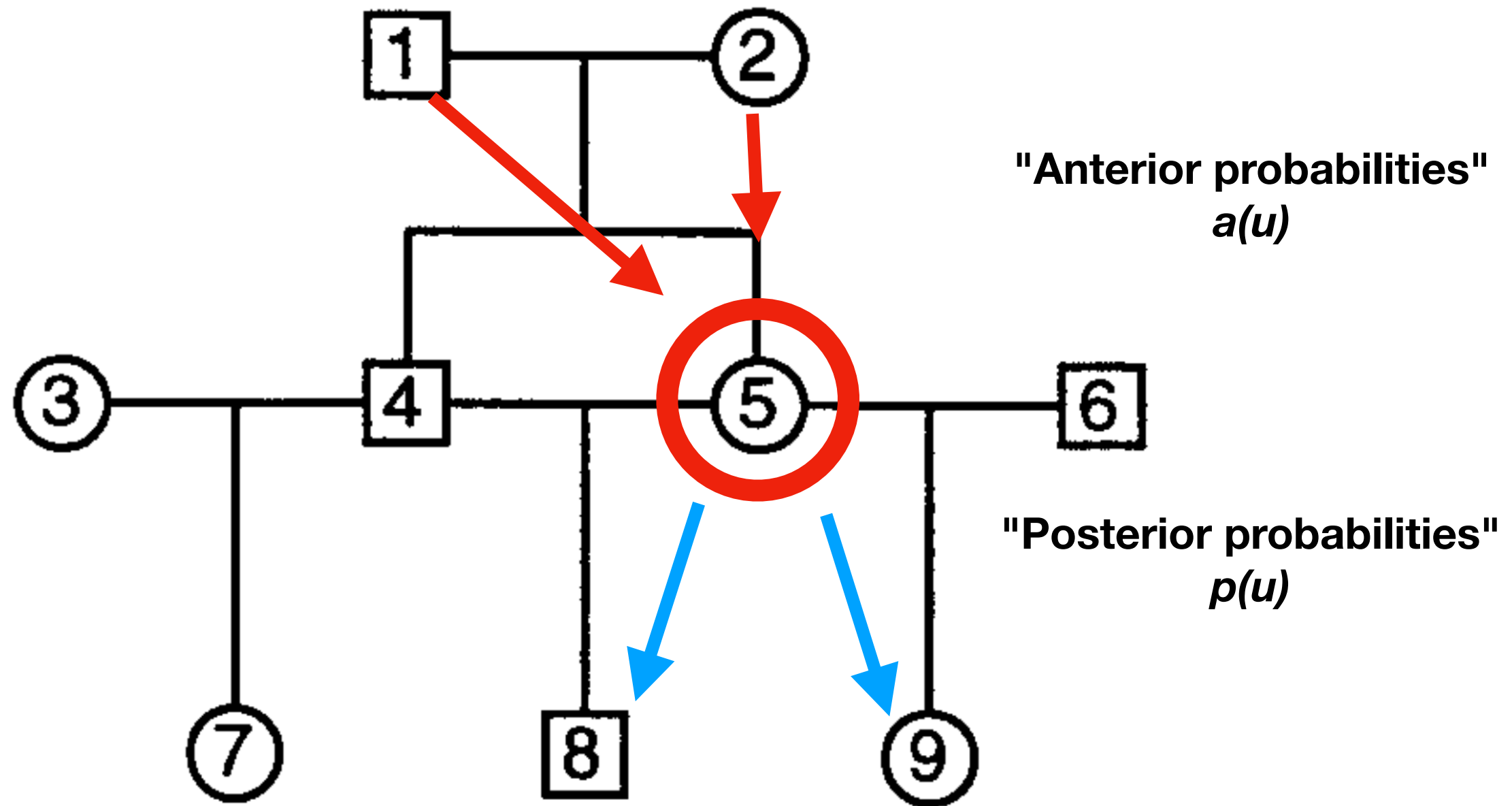
A Visual Display of Peeling

A Visual Display of Peeling

"Anterior probabilities"
a(u)

A Visual Display of Peeling

"Anterior probabilities"
a(u)

"Posterior probabilities"
p(u)

Kerr and Kinghorn 1996, *J Anim Breed Genet*

A Visual Display of Peeling

"Penetrance"
g(y|u)

"Anterior probabilities"
a(u)

"Posterior probabilities"
p(u)

Kerr and Kinghorn 1996, *J Anim Breed Genet*

A Visual Display of Peeling

"Penetrance"
$g(y|u)$

"Anterior probabilities"
$a(u)$

"Posterior probabilities"
$p(u)$

$$Pr(u_i \mid y) = \frac{a_i(u_i)g(y_i \mid u_i)\prod\limits_{j \in S_i} p_{ij}(u_i)}{L}$$

Kerr and Kinghorn 1996, *J Anim Breed Genet*

Important Caveats

- Peeling looks attractive for pedigreed populations, but has caveats:

  - Expensive. It is a slow process, and there are a lot of computations.

  - No exact solution for pedigrees with loops (inbreeding)

  - Treats marker loci as independent, which was fine in the mid-90s when a lot of the work was first done

AlphaPeel

- Uses "hybrid peeling" - multilocus iterative peeling to identify IBD in a pedigree, then single locus peeling to impute genotypes
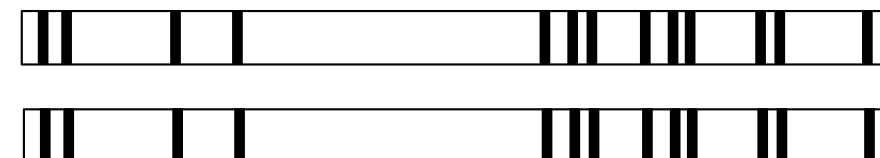
AlphaPeel

- Uses genetic map, observed genotypes, and known pedigree structure to propagate genetic information through pedigree

- Designed specifically for scenarios in which some individuals are resequenced, many are genotyped, and some may be unobserved
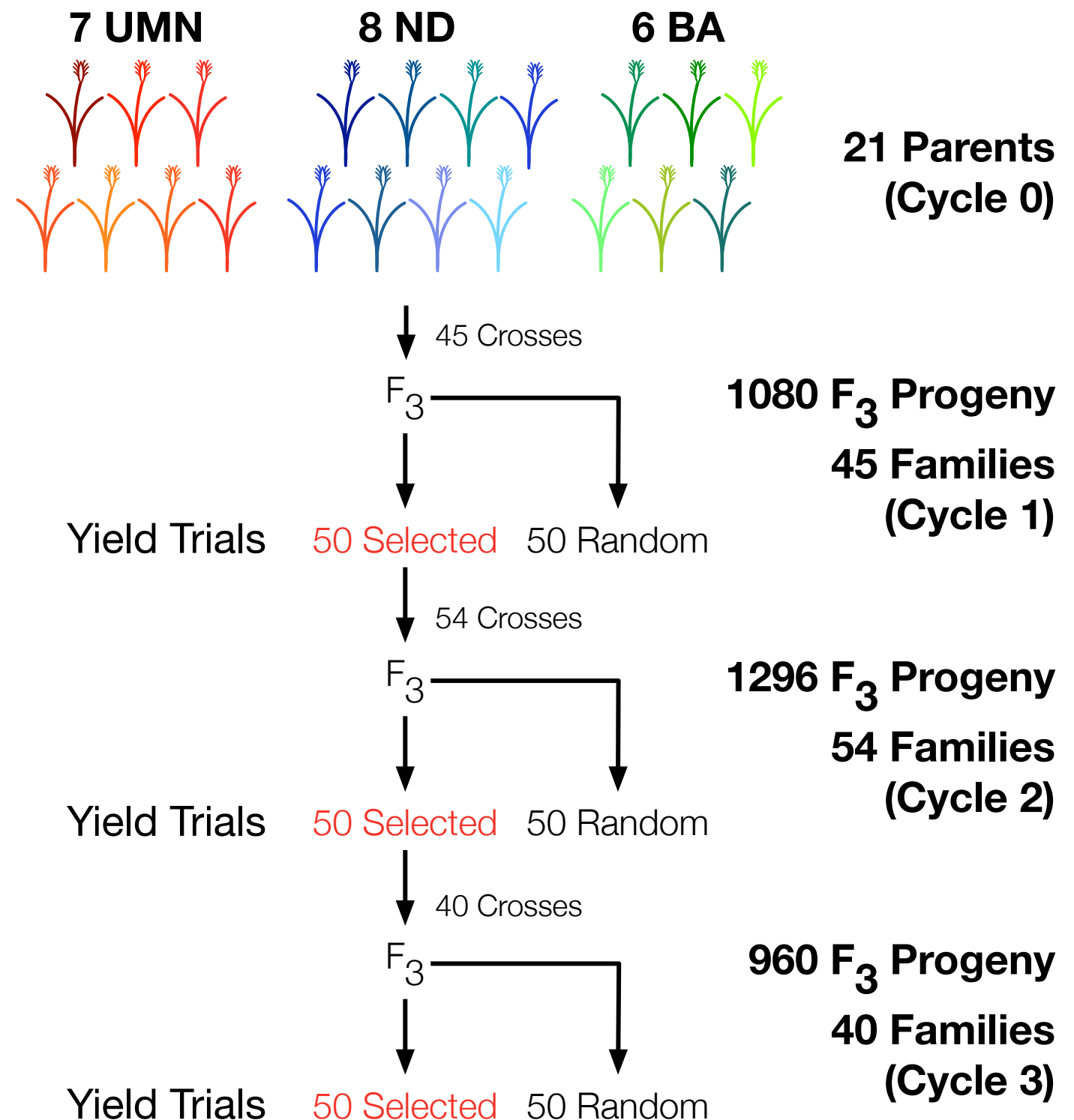

Parents With Genotypes and Sequences

Non-genotyped Progeny

Genotyped Progeny

Example: Barley GS Population

- All individuals were genotyped with 384 fixed SNP platform

- 21 parental individuals were resequenced

  - ~450,000 SNPs

- Crosses were recorded, so pedigree is known

7 UMN **8 ND** **6 BA**

21 Parents (Cycle 0)

45 Crosses

F_3

1080 F_3 Progeny

45 Families (Cycle 1)

Yield Trials   50 Selected   50 Random

54 Crosses

F_3

1296 F_3 Progeny

54 Families (Cycle 2)

Yield Trials   50 Selected   50 Random

40 Crosses

F_3

960 F_3 Progeny

40 Families (Cycle 3)

Yield Trials   50 Selected   50 Random

Pedigrees Are Important

- Pedigrees were advanced by "single seed descent"

  - One individual was self-fertilized

  - Family was reared, one individual was self-fertilized

- This structure must be explicitly coded for AlphaPeel
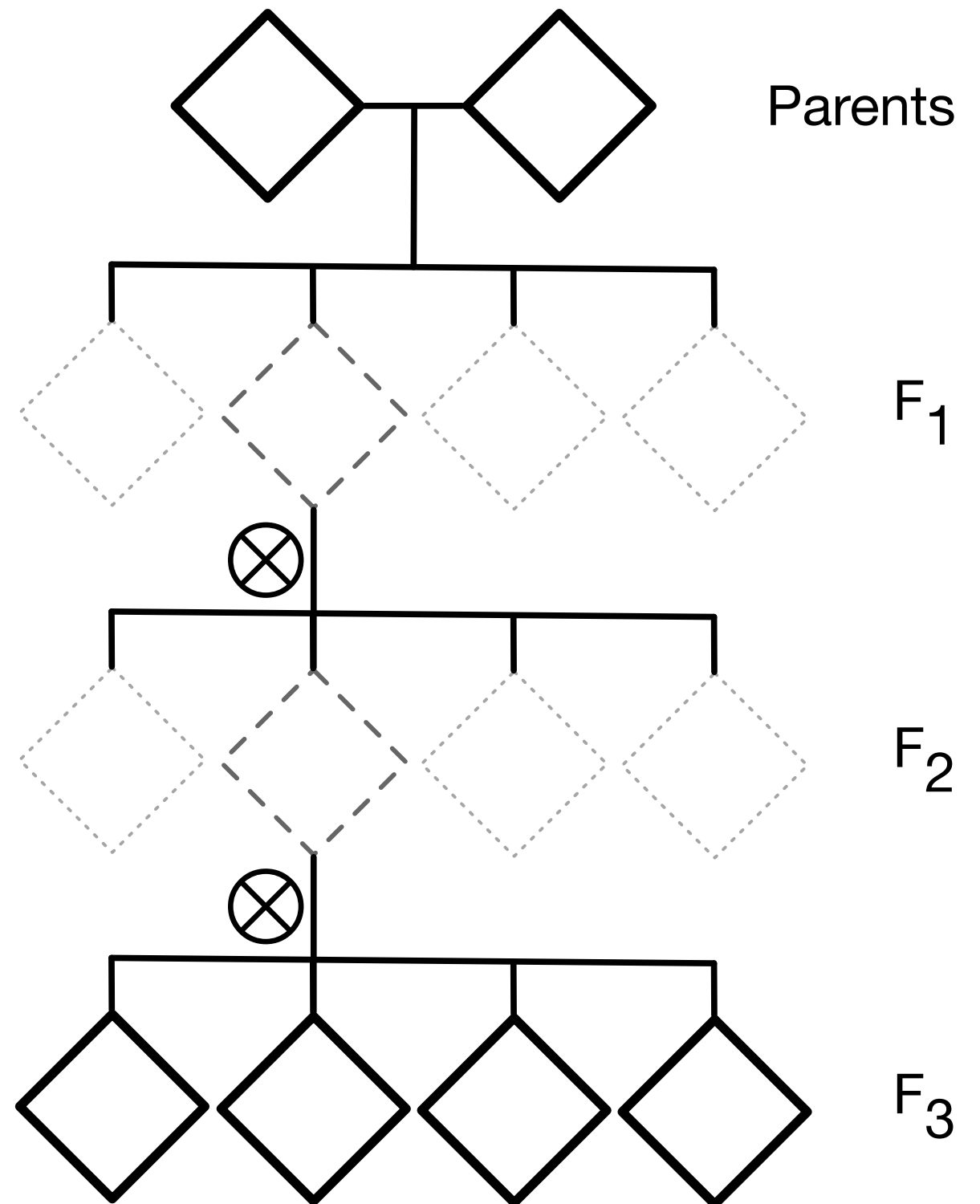
Genetic Maps, Too

- AlphaPeel genetic maps are coded as linear interpolations between markers with known genetic position (i.e., genotyped and mapped markers)

  - We know this isn't what recombination rate looks like, but it's a decent approximation with a dense enough set of markers

```
1 2 0.9996341145
1 2 0.9996391963
1 2 0.9998441599   ←  Resequencing SNPs between
1 2 0.9998509355      known marker positions 1 and 2
1 2 0.9999610400
2 2 0            ←  Known position (genotyped)
2 3 0.0000166892
2 3 0.0000199545
2 3 0.0000216476
2 3 0.0000637334   ←  Resequencing SNPs between
2 3 0.0000769155      known marker positions 2 and 3
2 3 0.0001073914
2 3 0.0001205735
```

Useful Links

- AlphaPeel:
  http://www.alphagenes.roslin.ed.ac.uk/alphasuite-softwares/alphapeel/

- Preprint:
  https://www.biorxiv.org/content/early/2017/12/04/228999

- IMPUTE2 Manual (LD-based imputation):
  http://mathgen.stats.ox.ac.uk/impute/impute_v2.html