Multi-generational pedigrees with Sequoia

Jisca Huisman

04-04-2017

THE UNIVERSITY of EDINBURGH

Pedigree reconstruction > paternity assignment

Often not all candidate parents are genotyped
$\rightarrow$ cluster siblings & assign dummy parent to each sibship

But: those sibships are unconnected to the rest of the pedigree.

- Sometimes a sibship can be matched to a non-genotyped parent using field data. If this field parent has field parents itself, it can provide a link to the rest of the pedigree
- If those field-grandparents are genotyped, they should be related to all siblings by 0.25, and so it should be possible to infer them genetically

Additionally: Colony needs to be run cohort-by-cohort; finding sibships that span multiple cohorts is a pain

Full pedigree reconstruction

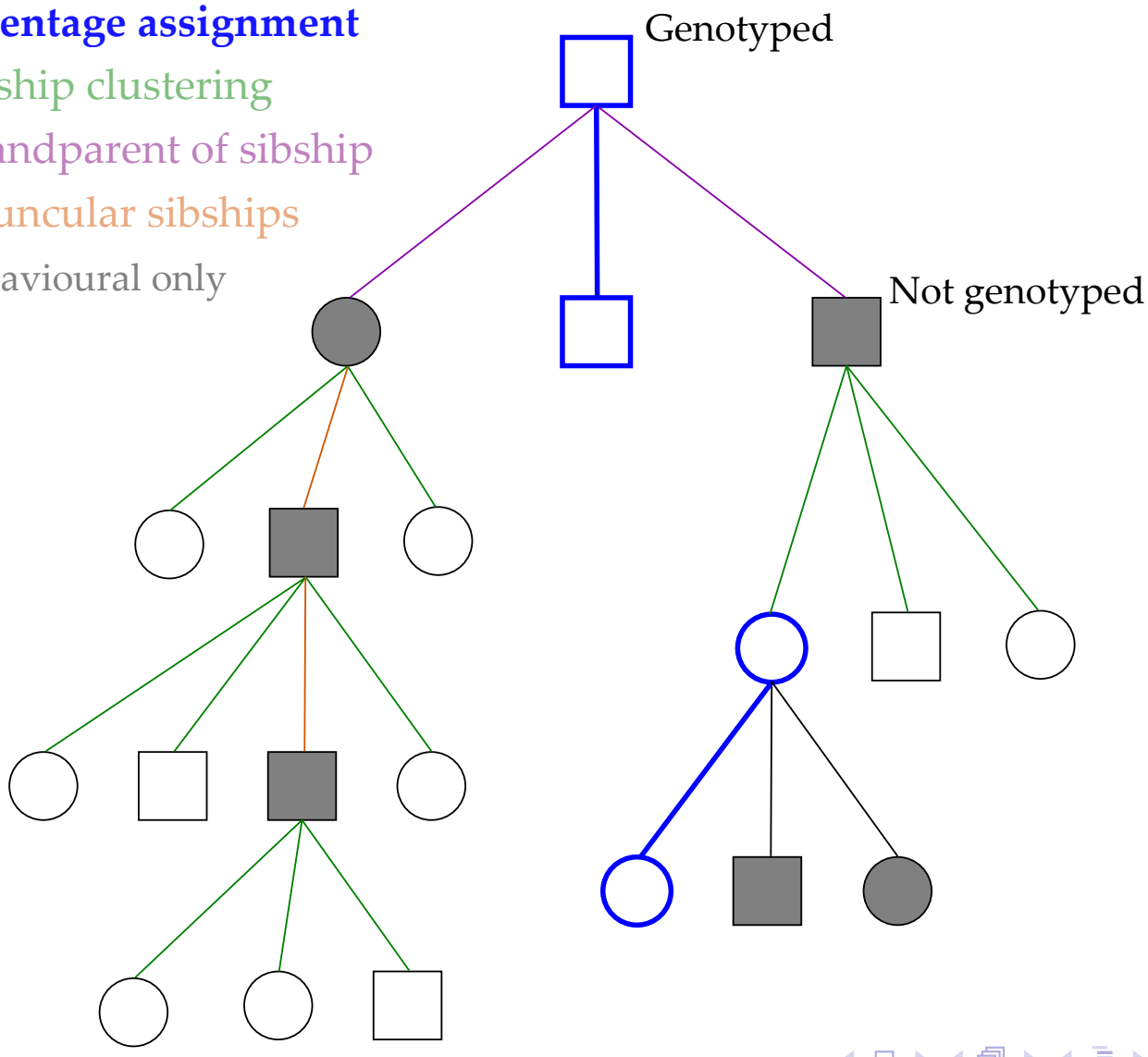
Parentage assignment

Sibship clustering

Grandparent of sibship

Avuncular sibships

Behavioural only



Full pedigree reconstruction

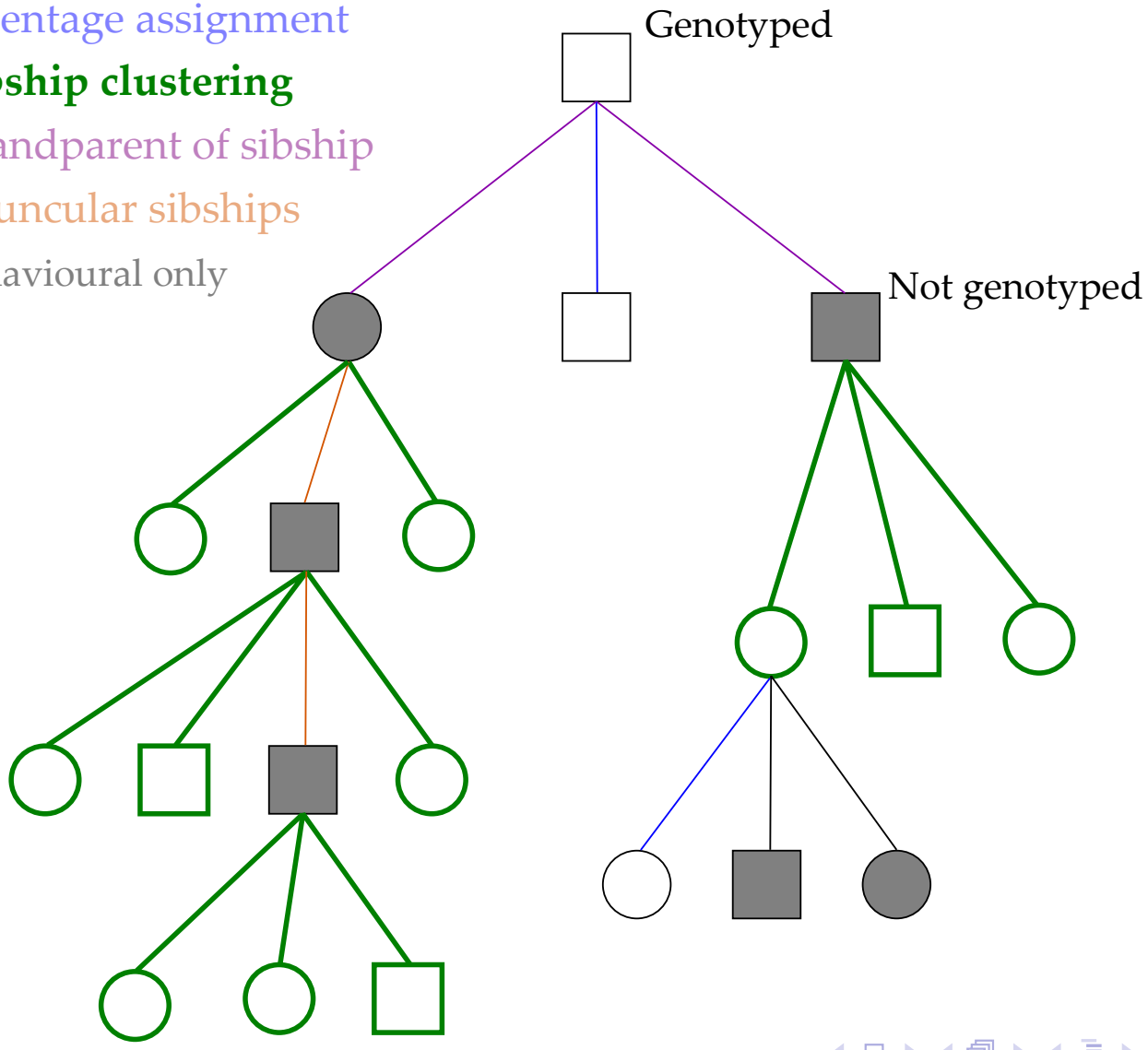
Parentage assignment

Sibship clustering

Grandparent of sibship

Avuncular sibships

Behavioural only



Full pedigree reconstruction

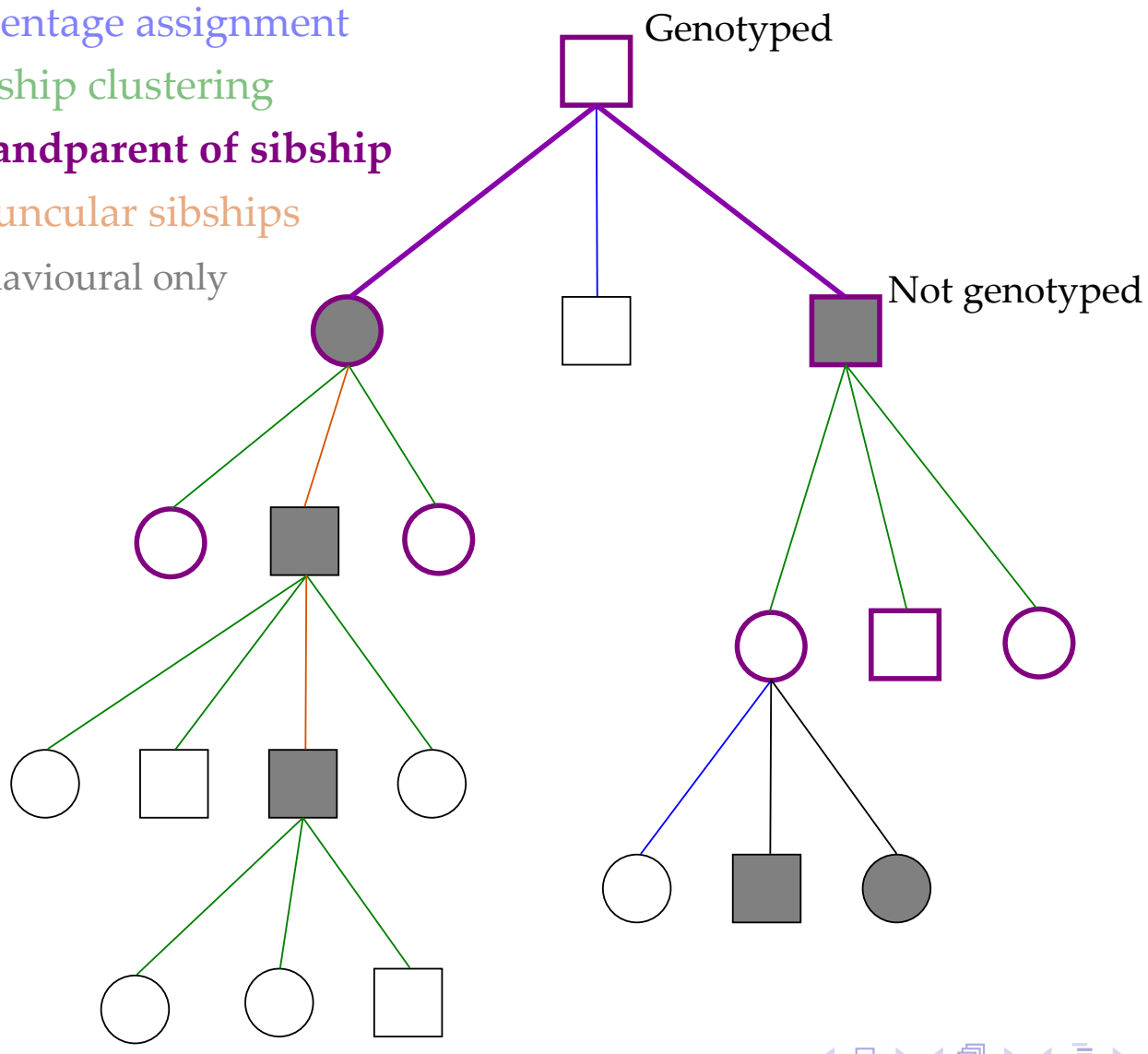
Parentage assignment

Sibship clustering

Grandparent of sibship

Avuncular sibships

Behavioural only



Full pedigree reconstruction

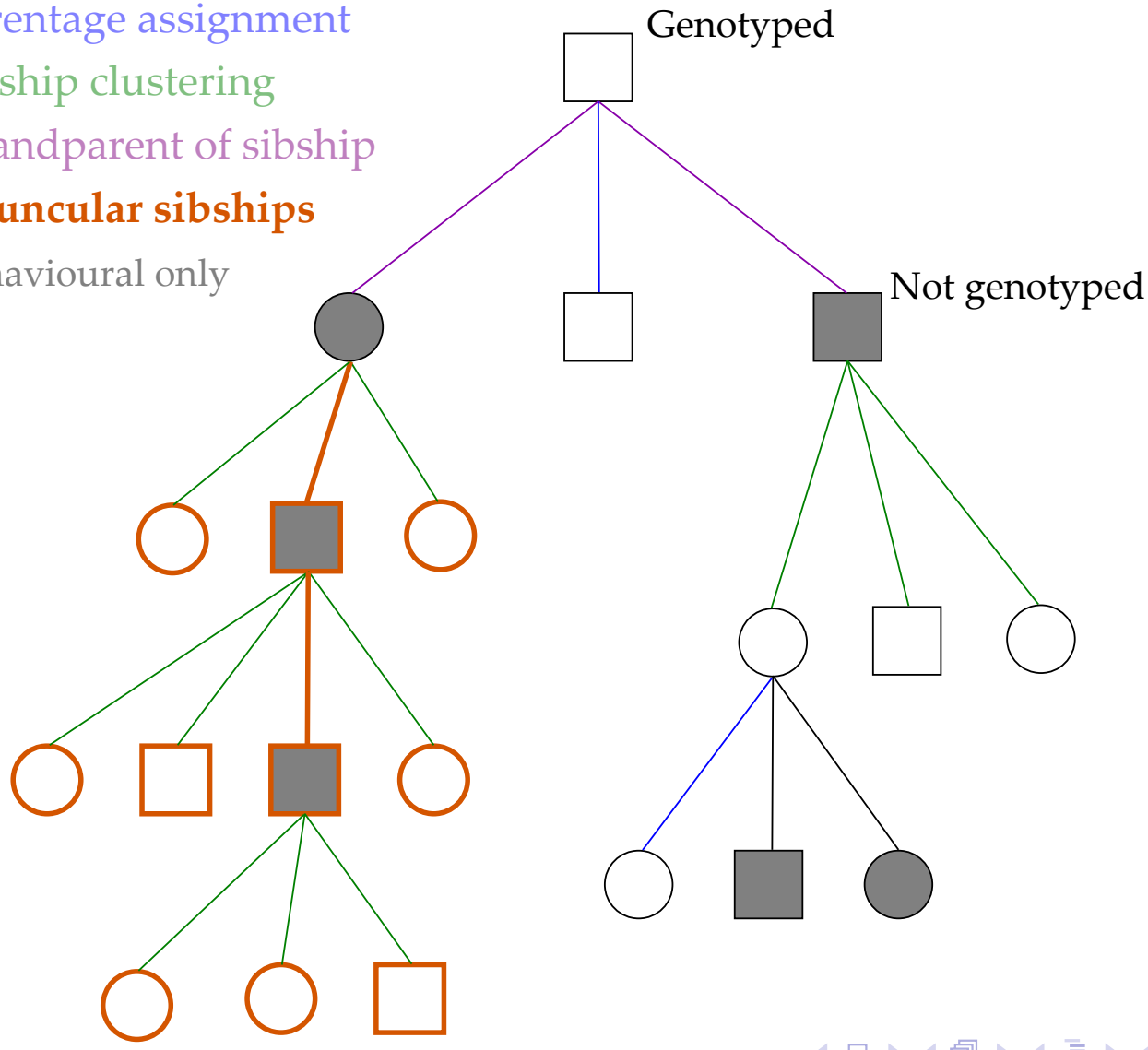
Parentage assignment

Sibship clustering

Grandparent of sibship

Avuncular sibships

Behavioural only



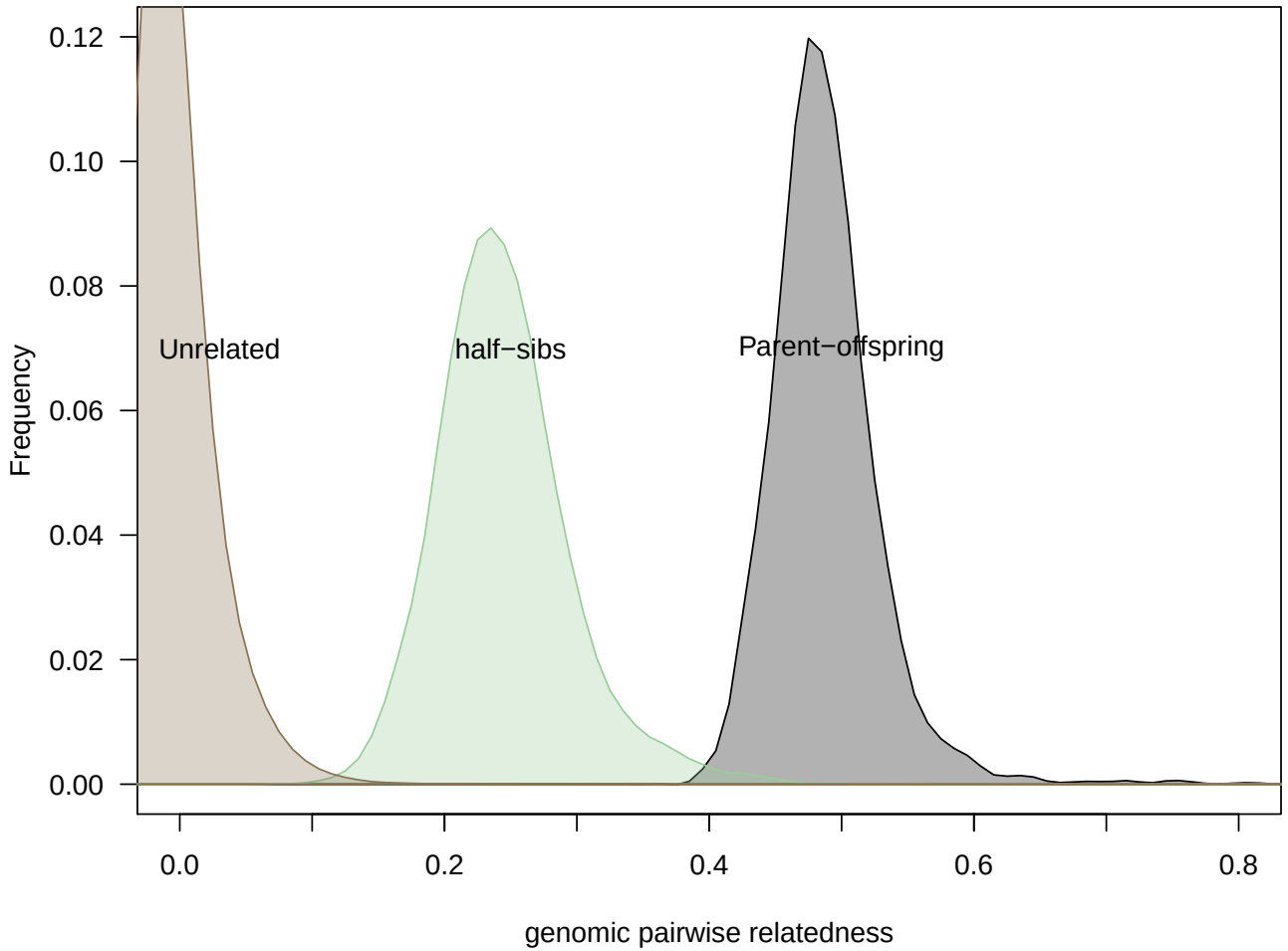
How do you make a pedigree from genomic data?

- Just exclude all impossible candidate parents (/relatives)
  - Provides no solution if > 1 candidate parent is non-excluded

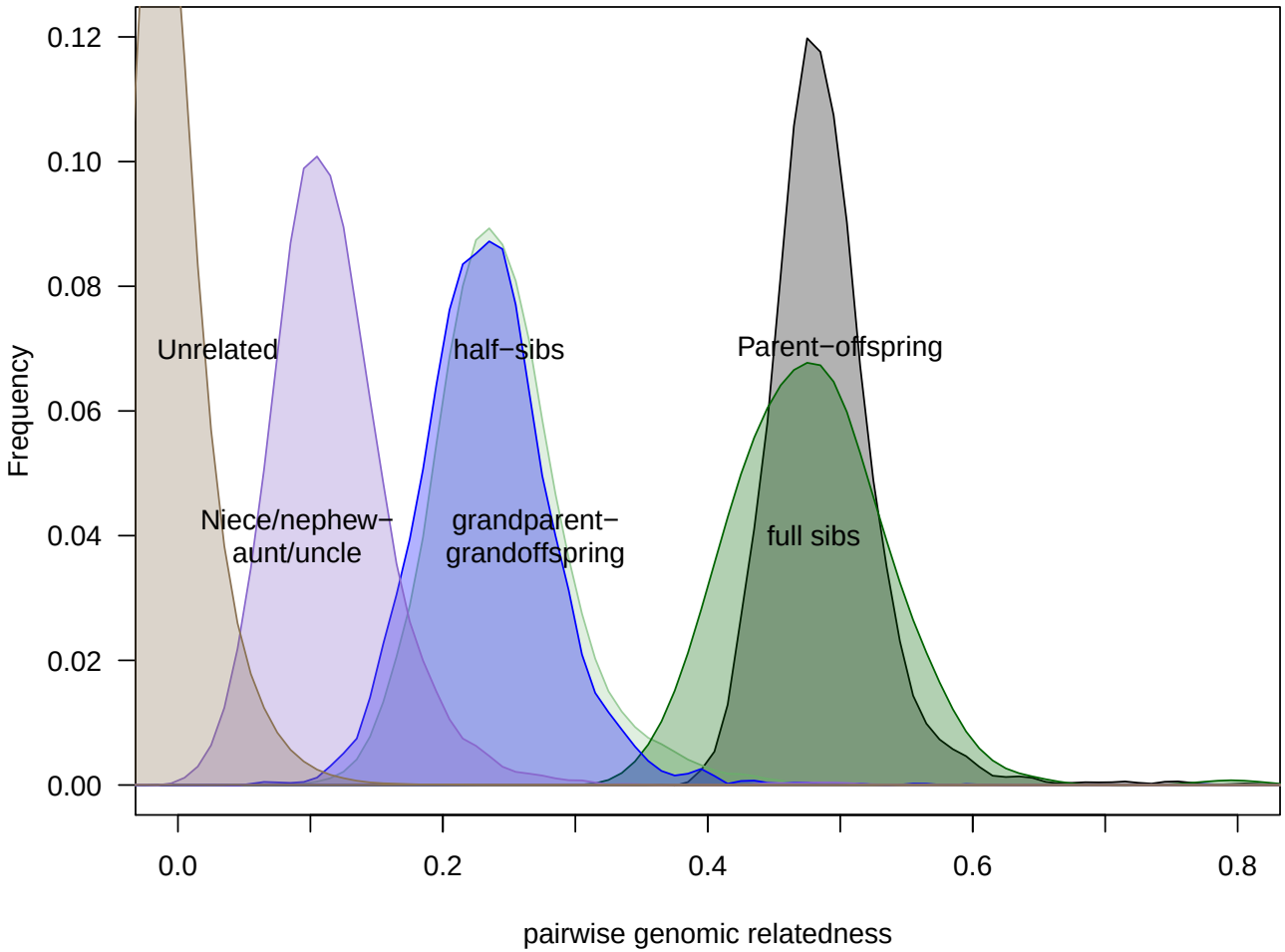
How do you make a pedigree from genomic data?

- Just exclude all impossible candidate parents (/relatives)
  - Provides no solution if > 1 candidate parent is non-excluded
- Use genomic pairwise relatedness

How do you make a pedigree from genomic data?



How do you make a pedigree from genomic data?



How do you make a pedigree from genomic data?

- Just exclude all impossible candidate parents (/relatives)
  - Provides no solution if > 1 candidate parent is non-excluded
- Use genomic pairwise relatedness
- Likelihood approach

Likelihood approach

Likelihood of a pedigree = probability of observing the genotypes, given that pedigree

- The probability that you have observed genotyped G, given that you have actual genotype Z ($1 - \epsilon$ etc.)
- Probability that you inherited genotype Z, given that your parents have genotype X and Y (0, 1/2, 1/4, or 1)
- Probability that parent P has actual genotype X given its observed genotype, or allele frequency + HWE, or your siblings's genotypes

Sum over all (3*3*3) possible values of X, Y and Z

(microsat w 10 alleles $\rightarrow$ 10 homs + 9*10 hets, 166 375 possible combo's)

$\rightarrow$ can be extended to many individuals, not just pairs

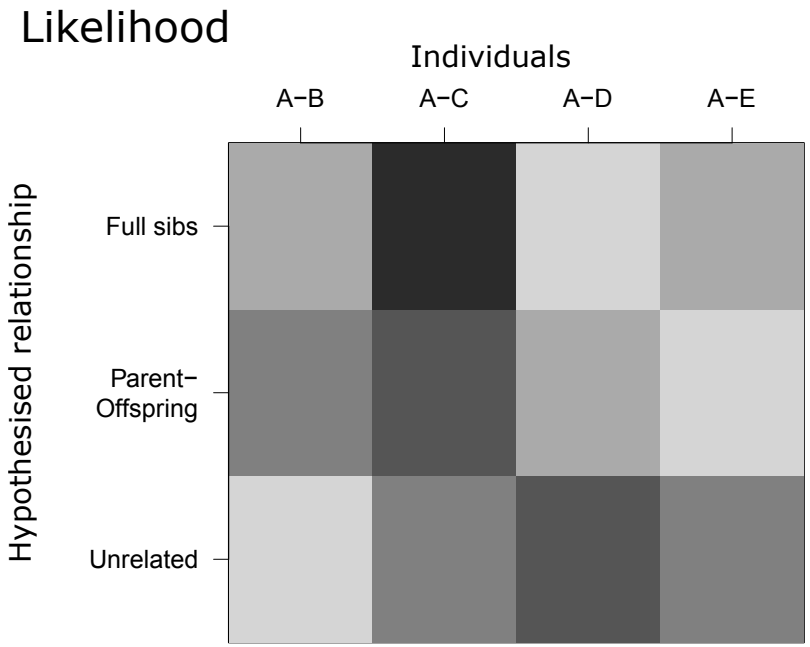
Problems with likelihoods - 1

- the likelihood that A is the parent of B, is identical to the likelihood of B being the parent of A
  - need info on who of the two is the eldest (or candidate parent lists)
  - possible to find 'complementary' parent pairs even w/o age info
- the likelihoods of all 2nd degree relationships (half sib, grandparental, full avuncular) are identical
  - age info can help a lot
  - if both have a parent assigned, the likelihoods do differ (does work with dummy parents too)

Problems with likelihoods - 2

"Although comparison of alternative genealogical relationships between a specified set of individuals is an entirely consistent statistical procedure,

comparison of alternative individuals for a given genealogical relationship is not." (Thompson, 1987)

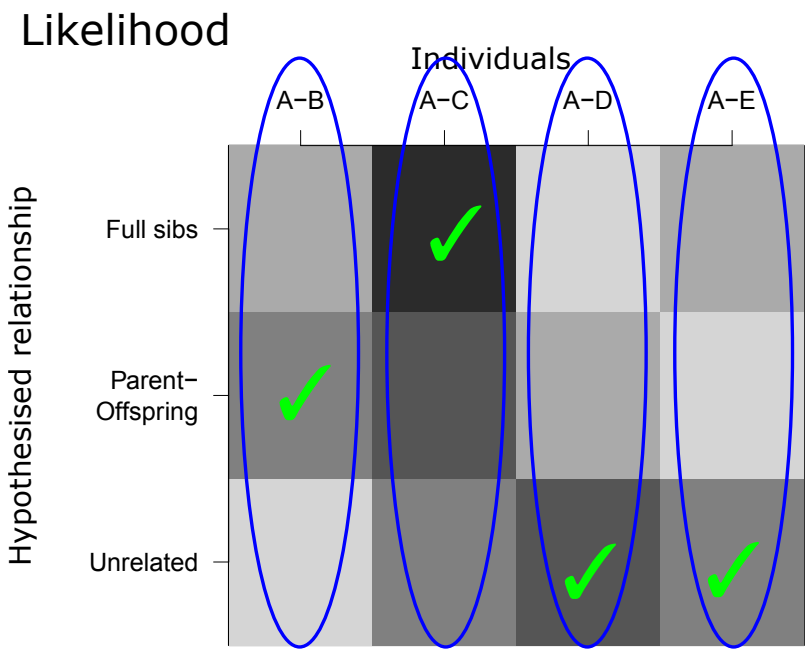


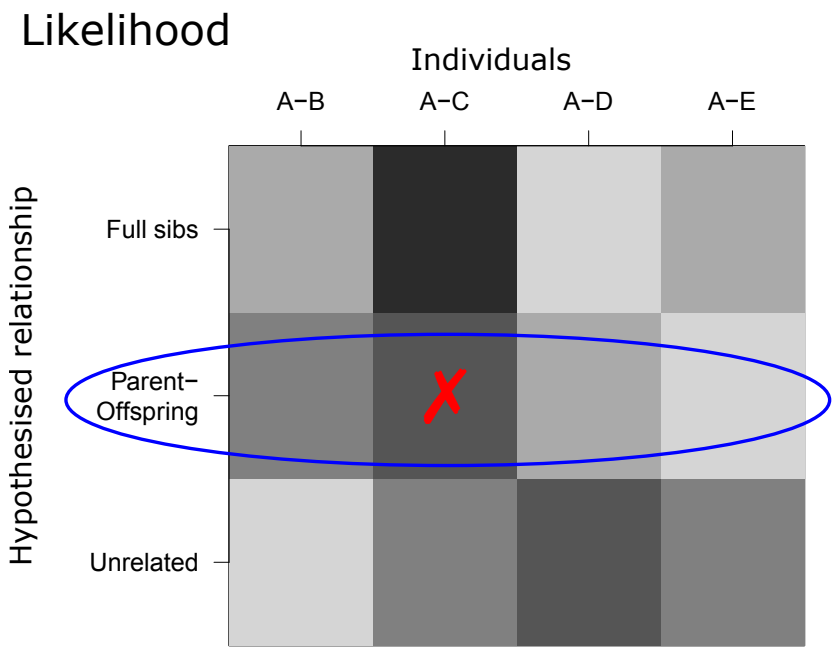
Darker = higher pairwise likelihood

Problems with likelihoods - 2

"Although comparison of alternative genealogical relationships between a specified set of individuals is an entirely consistent statistical procedure,

comparison of alternative individuals for a given genealogical relationship is not." (Thompson, 1987)





Solutions to problem 2

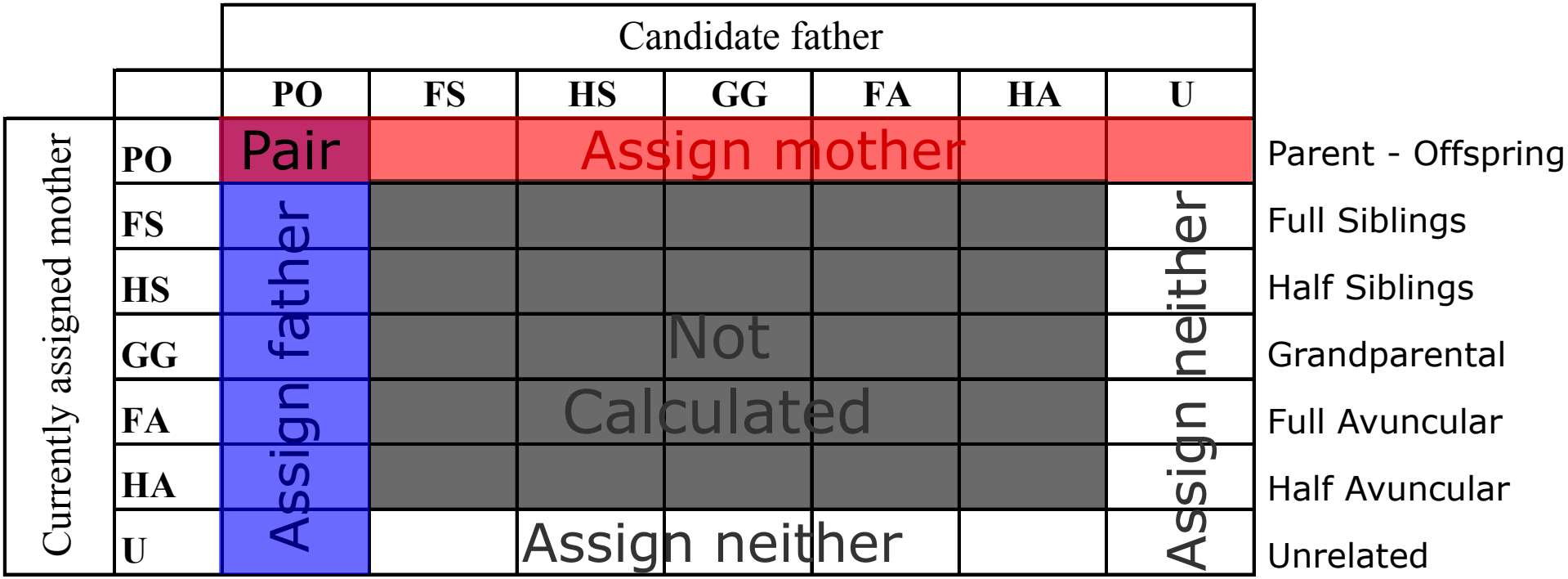
Colony, MasterBayes (?): optimise likelihood over all individuals, rather than each pair

Sequoia - For each pair of candidate relatives, calculate the likelihood of observing their genotypes if they were . . .

1. parent and offspring
2. full siblings
3. half siblings
4. grandparent and grandoffspring
5. full avuncular (aunt/uncle - niece/nephew)
6. 3rd degree relatives (half avuncular, cousins, great-grandparent)
7. unrelated

And assign only if the focal relationship is more likely than all others.

Parent pairs



Calculate the 7 likelihoods conditional on the opposite-sex candidate parent being a parent, and on it being unrelated (including inbred variations)

Repeat for all possible pairs among those that passed filtering (more likely parents will 'take over')

Inbreeding & 'double' relationships

Big disadvantage of this approach: Requires explicit consideration of all possible configurations for best results.

For example, these are easily mistaken for full siblings:

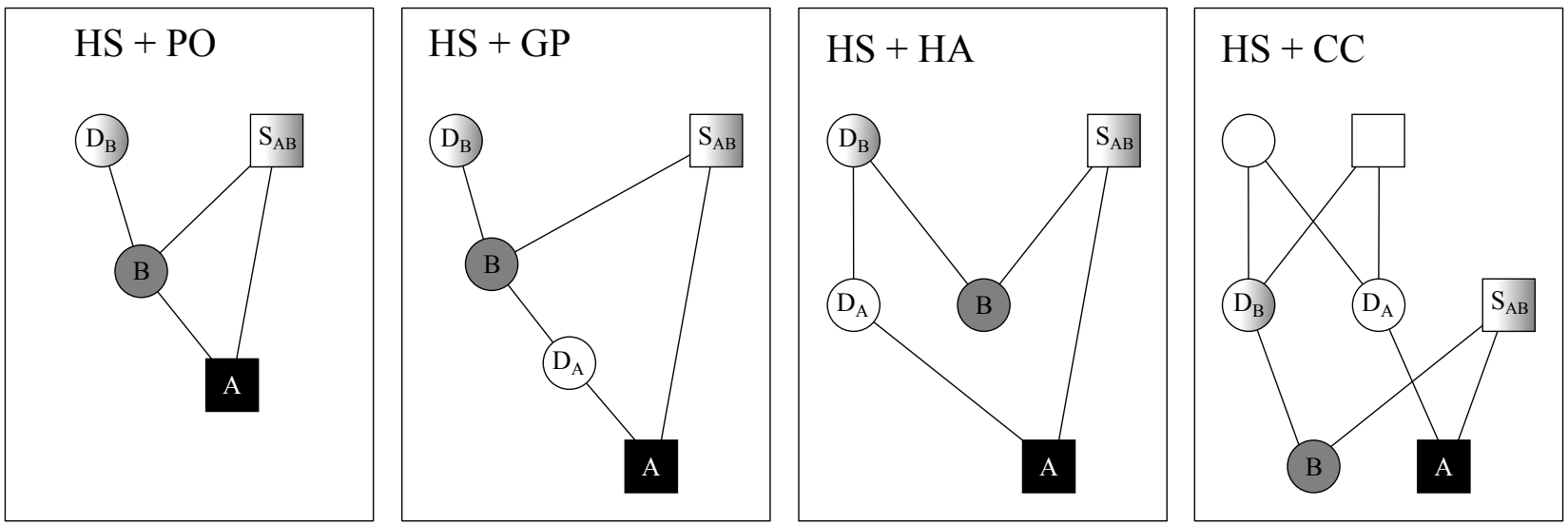


Figure: Examples of double relationships between genotyped individuals A and B, where D_B and S_{AB} may or may not be genotyped, and D_A is not genotyped.

Likelihood optimisation

Calculating all these likelihoods makes checking each potential assignment computationally intensive

- MCMC with tens of thousands of iterations probably would take months
  - $\rightarrow$ no per-individual posterior probability
- Use a conservative hill-climbing algorithm
  - Build upon already made assignments ($\rightarrow$ some pre-set assignment threshold needed)
  - Asymptotes usually in < 10 iterations
  - LLR(Parent / most likely not-parent relationship) in final pedigree provides indication of assignment confidence
- Don't do all calculations for all pairs

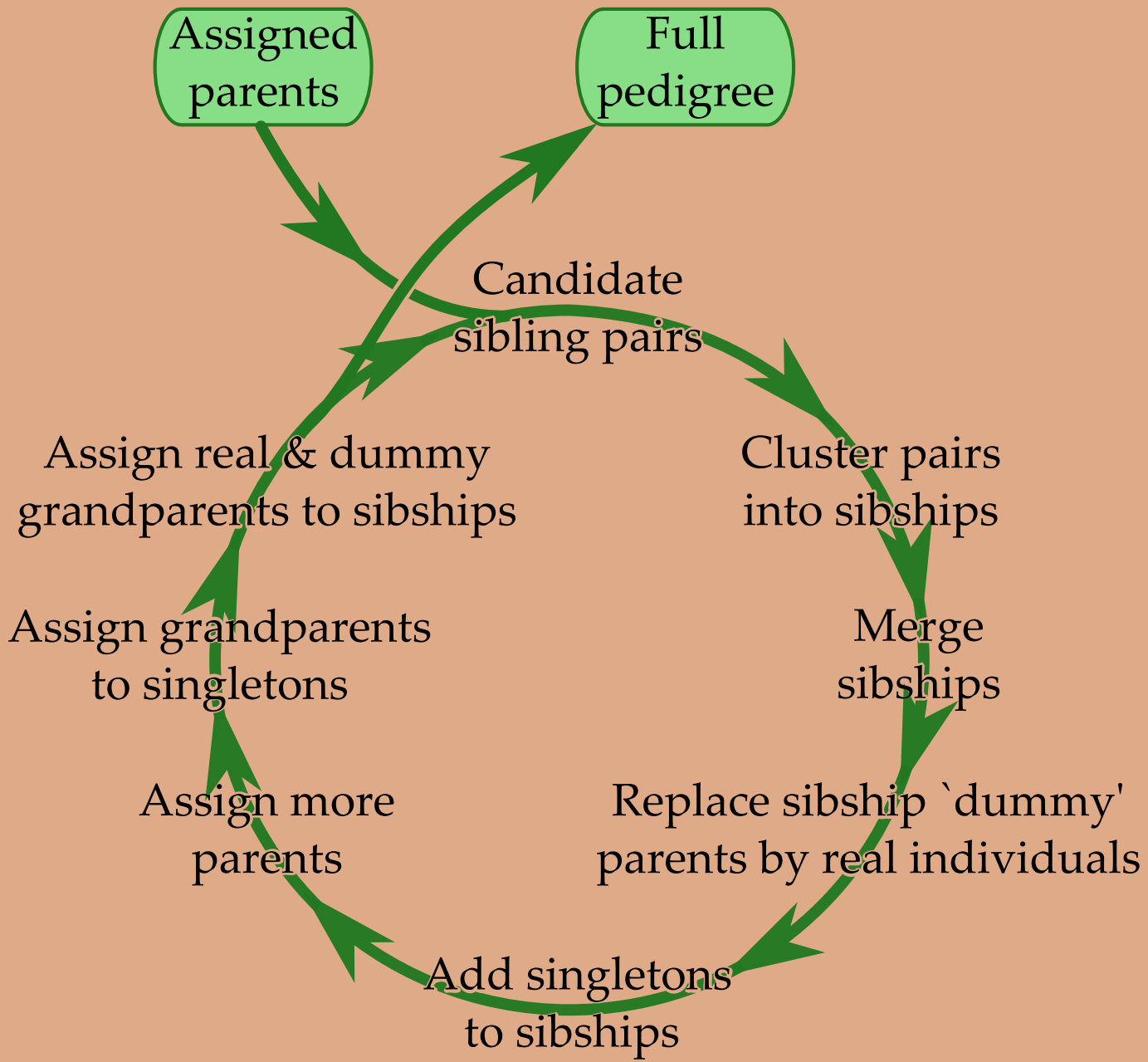
Speeding things up

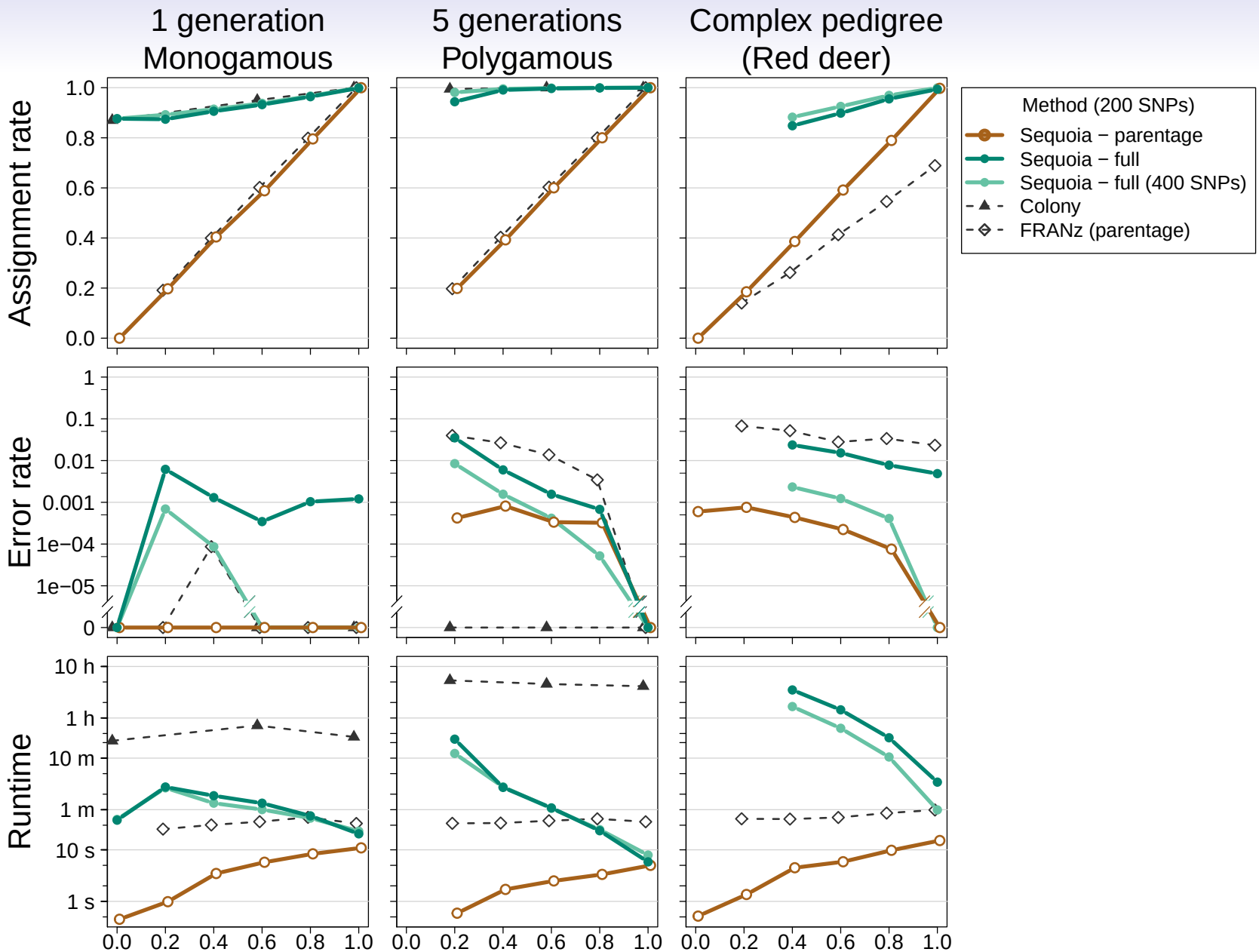
Everybody is a potential candidate parent for everybody else $\rightarrow$ drop all candidates through a series of filters with decreasing 'mesh size' before doing all likelihood computations

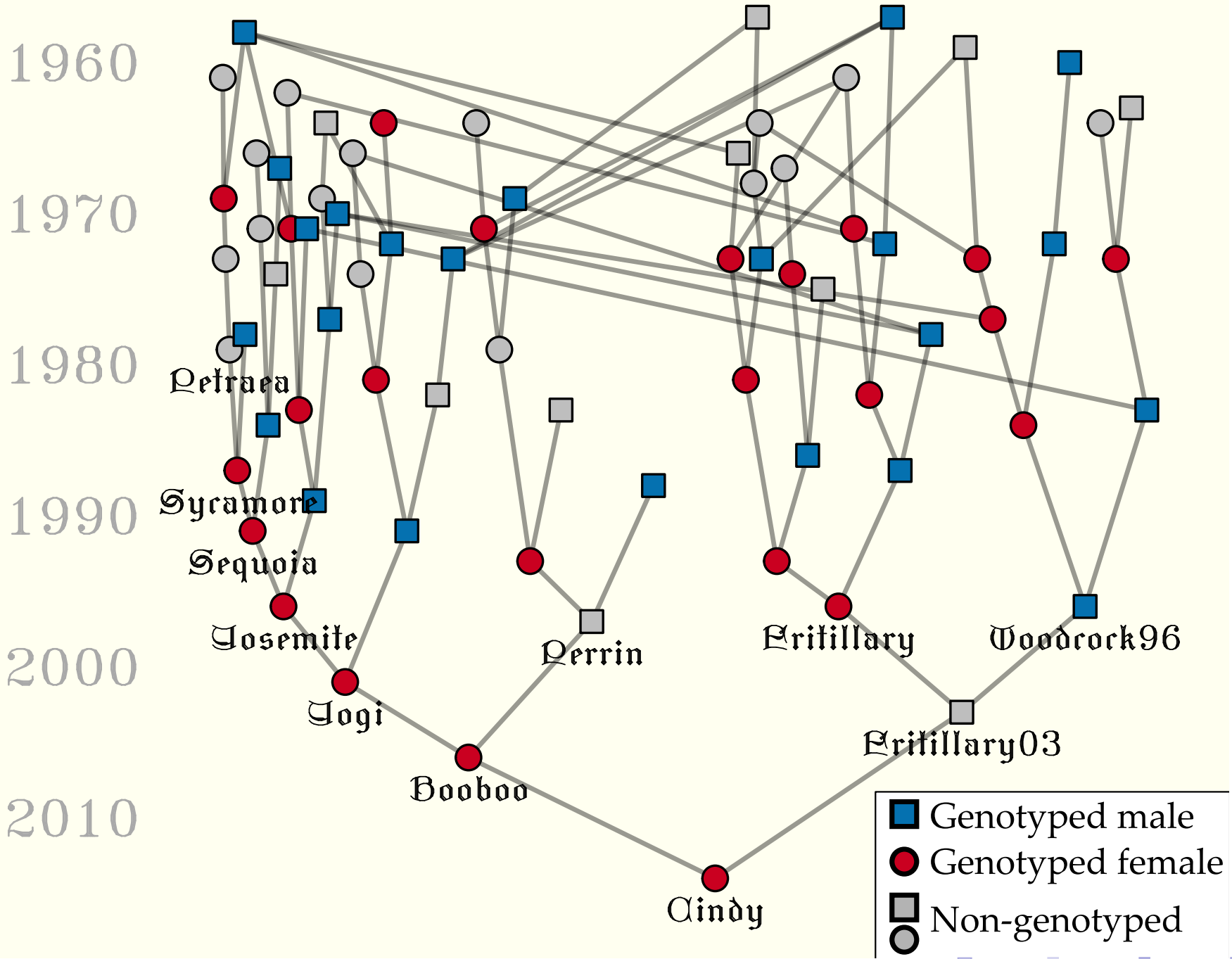
- Exclude any with many opposing homozygous loci
- Exclude any which are younger
- Calculate the approximate LLR(parent/unrelated), without considering any possibly already assigned parents, and exclude if this LLR is lower than T_{filter} (user-set, default -2)

and only then

- calculate the likelihoods of all possible relationships, and assign if parent-offspring is more likely than the next-most-likely by a margin T_{assign} (user-set, default $+0.5$)


Assigned parents
Full pedigree
Candidate sibling pairs
Cluster pairs into sibships
Merge sibships
Replace sibship `dummy' parents by real individuals
Add singletons to sibships
Assign more parents
Assign grandparents to singletons
Assign real & dummy grandparents to sibships


1 generation
Monogamous
5 generations
Polygamous
Complex pedigree
(Red deer)
Assignment rate
Error rate
Runtime
Prop. genotyped parents
Method (200 SNPs)
Sequoia − parentage
Sequoia − full
Sequoia − full (400 SNPs)
Colony
FRANz (parentage)


1960
1970
1980
1990
2000
2010
Petraea
Sycamore
Sequoia
Yosemite
Yogi
Booboo
Perrin
Cindy
Eritillary
Woodcock96
Eritillary03
Genotyped male
Genotyped female
Non-genotyped