Extensions of the coalescent

Matthieu Foll
21.11.2011
Population Genomics course
Helsinki

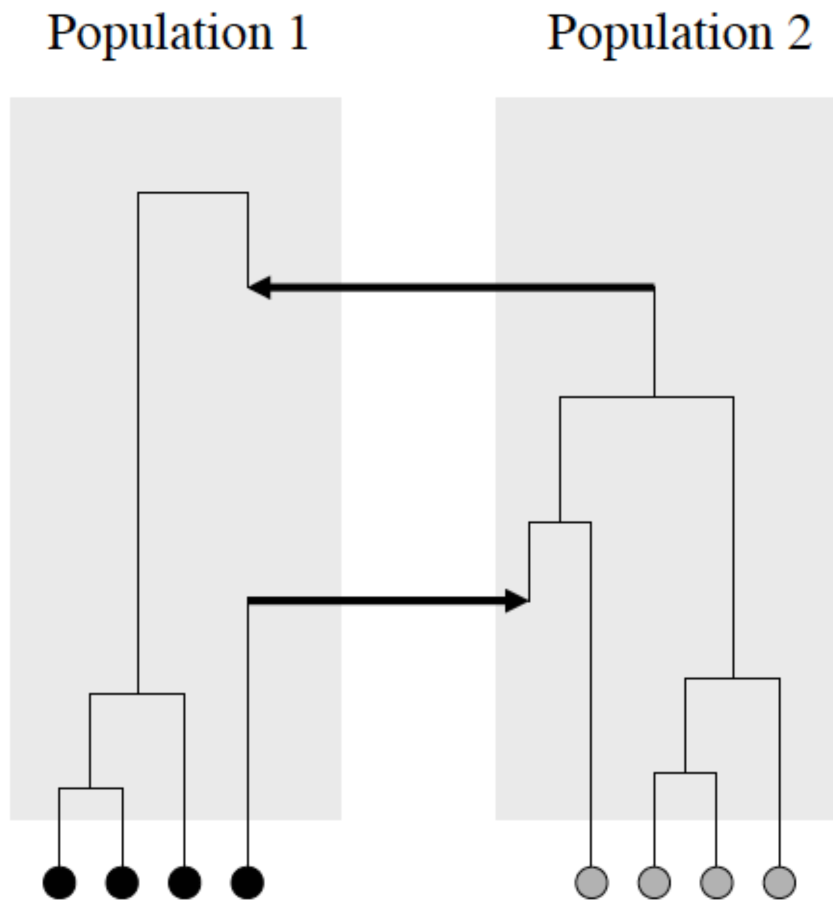
Extensions of the coalescent

- The coalescent with migration
  - Island models
  - Splitting models
  - Spatial models
- Gene tree vs. species tree
- Recombination

Coalescent with migration

- Population subdivided in demes/subpopulations
- Standard coalescent event within each deme
  - Geometric distribution
- At each generation, each individual has a probability m to migrate to another deme
  - Geometric distribution for time to migration

Genealogies in subdivided populations



$$P(coalescence) = \frac{n(n-1)}{4N}$$

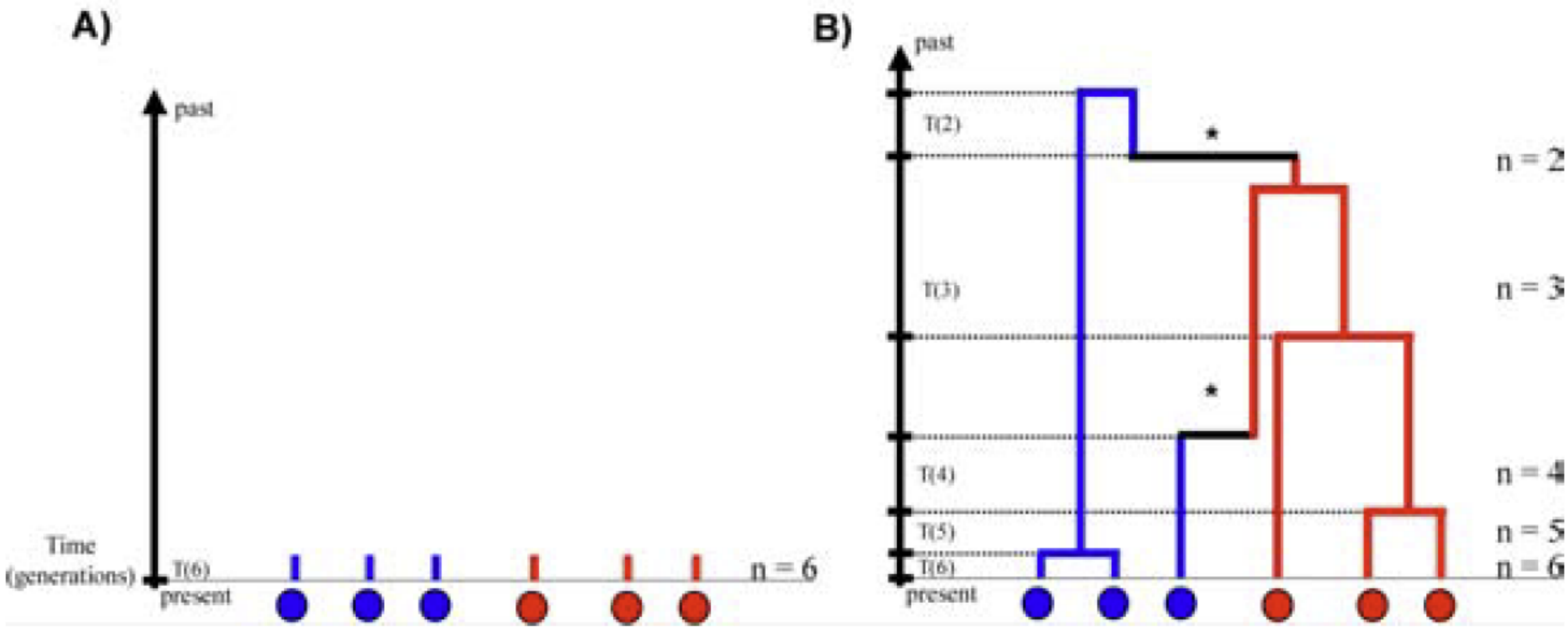
$$P(migration) = n \cdot m$$

Basic algorithm with migration

- Sample n genes from two geographical locations
- Label these individuals by sample location

- Draw a (geometric/exponential) coalescent time with a rate being the sum of the coalescent and migration probabilities
- Choose the kind of event that it will be proportional to the rates of the different events.
  - If the event is a coalescent, then randomly draw two individuals from the same population to coalesce.
  - If the event is a migration event, re-label one randomly chosen individual.

Building the tree



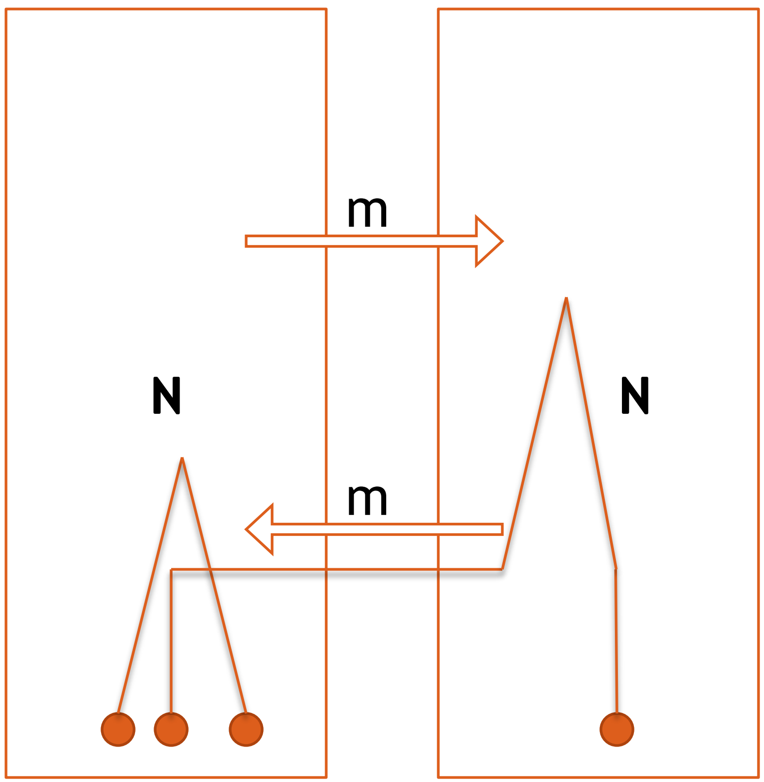
Fst and coalescence time

- We can think of F_{ST} as a measure of the difference in coalescent times within subpopulations relative to the coalescent times between subpopulations:

$$F_{ST} = \frac{t_1 - t_0}{t_1} \qquad \text{(Slatkin 1991)}$$

  - t_0: average coalescent time for a pair of alleles chosen randomly within subpopulations
  - t_1: average coalescent time for a pair sampled at random in different subpopulations

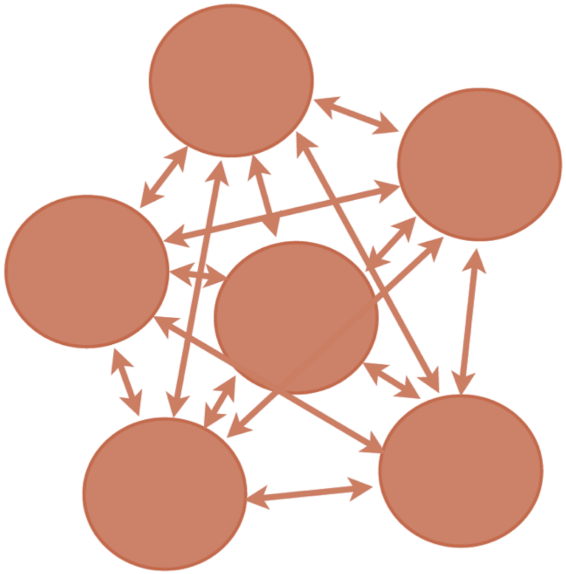
Two populations with migration



- 2 genes within a population:
  - t_0=4N (Strobeck 1987)
  - Does not depend on m!
- 2 genes between populations:
  - P(migration)=2m
  - E[time to migrate]=1/2m
  - t_1=1/2m+4N

$$F_{ST} = \frac{1}{1+8Nm}$$

Generalization



- Finite island model
  - d demes of size N
  - Migration rate m

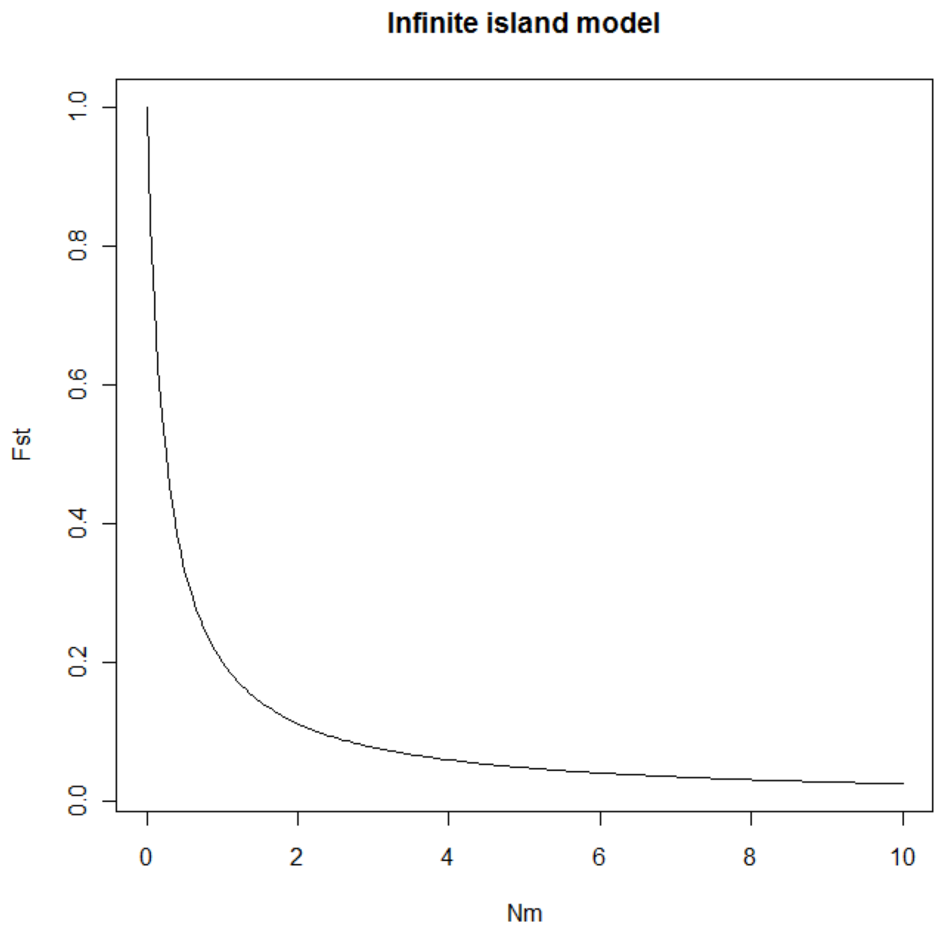
$$F_{ST} = \frac{1}{1 + \frac{4Nmd}{d-1}}$$

  - With d=2 we recover the previous result

- Infinite island model
  - $d \to \infty$
  - $d/(d-1) \to 1$

$$F_{ST} = \frac{1}{1+4Nm}$$

Demo

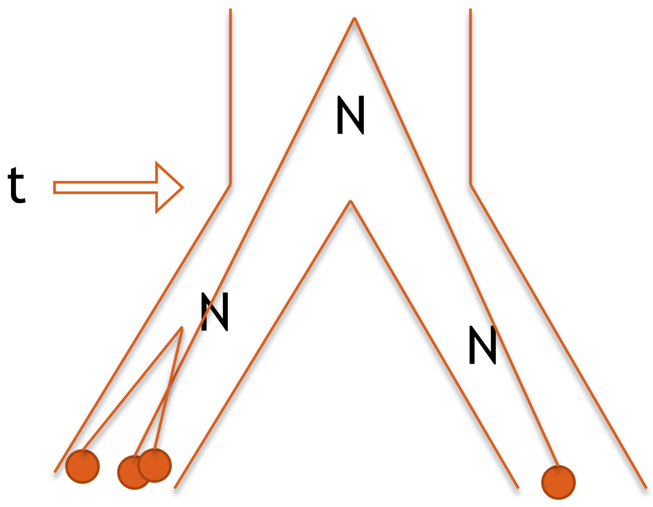



```
./fastsimcoal -i island_model_20_0.005_10_100.par -n 100
./LaunchArlSumStatDirMac.sh island_model_20_0.005_10_100 SettingsDNAStats.ars stats.txt
```

In R:

```
res=read.table("stats.txt",header=T)
mean(res$FST)
1/(1+4*100*9*0.005*10/9)
```

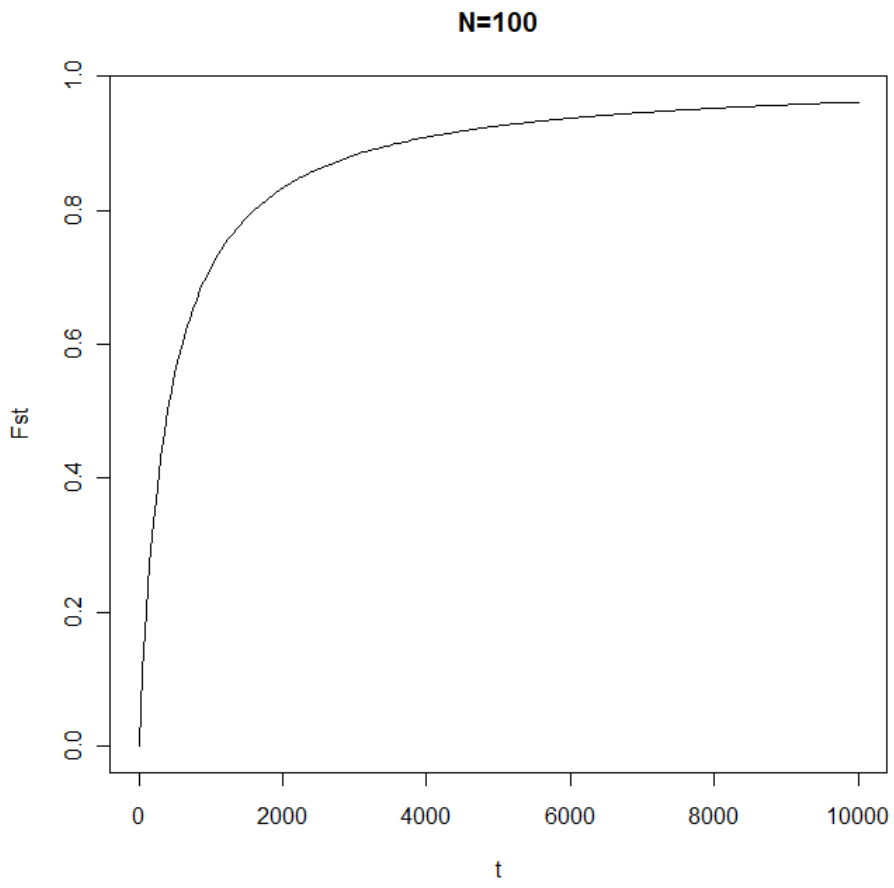
Population split



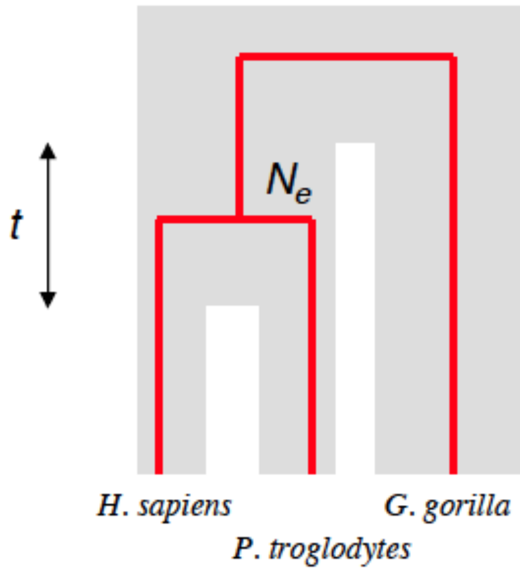
- A population of size N split in 2 populations of size N t generations ago
- 2 genes within a population:
  - $t_0 = 2N$
- 2 genes between populations:
  - $t_1 = t + 2N$

$$F_{ST} = \frac{t_1 - t_0}{t_1} = \frac{t}{t + 2N}$$

Demo

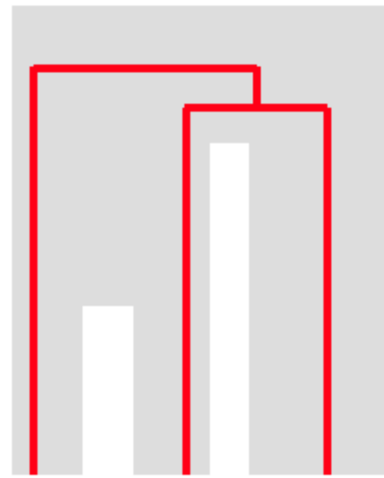


Gene tree vs. species tree



Gene tree and species tree congruent

$$P = 1 - \frac{2}{3} e^{-t/2N_e}$$

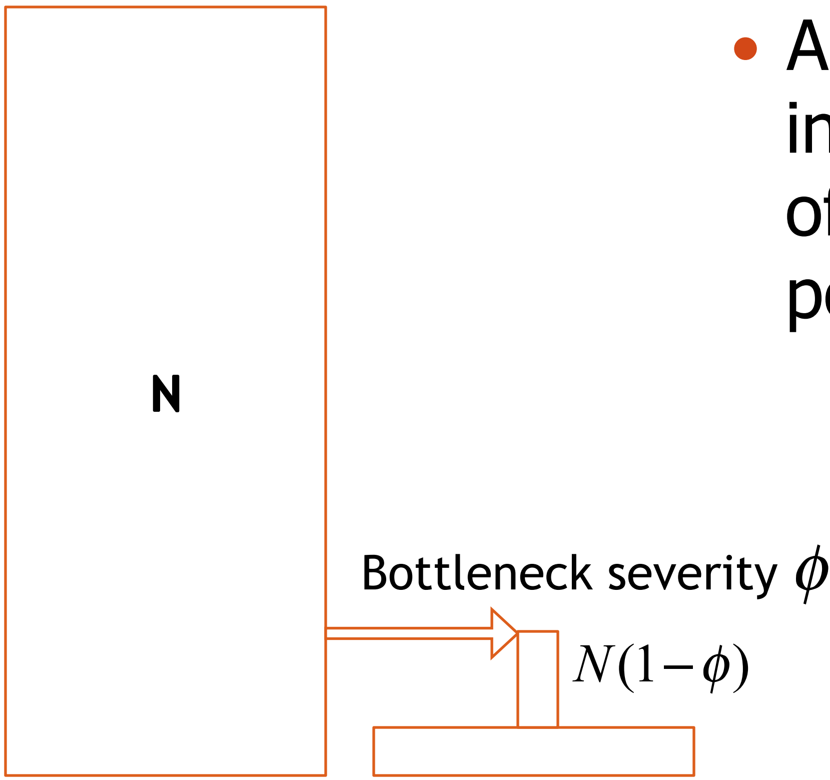


Gene tree and species tree incongruent

$$P = \frac{2}{3} e^{-t/2N_e}$$

- Can estimate divergence times and ancestral population sizes
- Chen and Li (2001) found with 53 autosomal regions 68% of congruent trees
- t = 0.766*2Ne

Founder effect



- A small proportion of individuals from a population of size N found a new population of size $N(1-\phi)$

$$t_{01} = 2N$$

$$t_{02} = 2N(1-\phi)$$

$$t_0 = N + N(1-\phi)$$

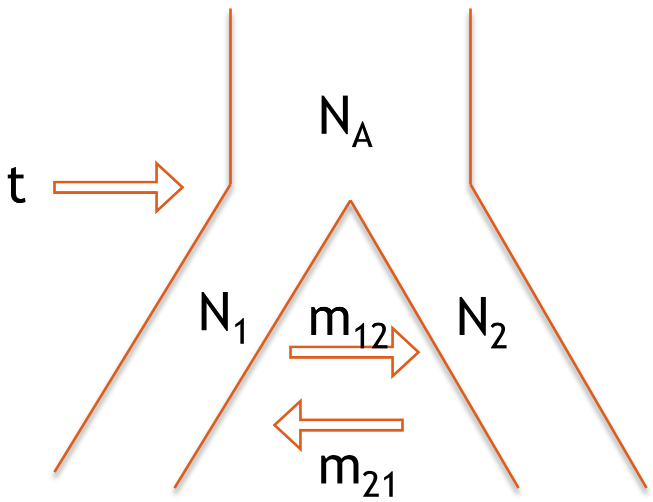
$$t_1 = 2N$$

$$F_{ST} = \frac{\phi}{2}$$

What does Fst measure?

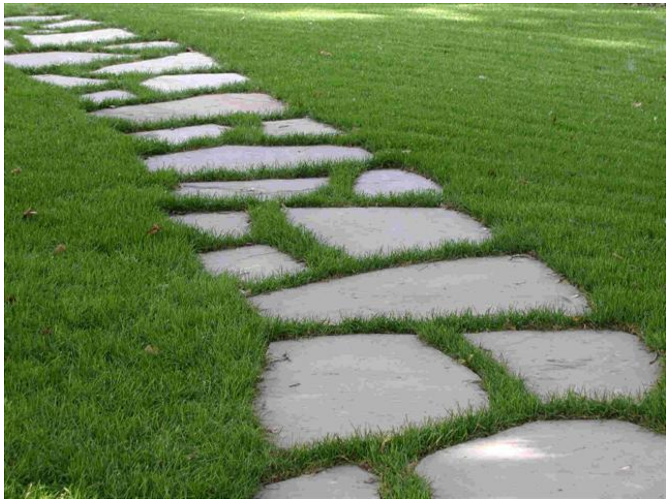
- Fst can be used as a statistic to summarize patterns of differentiation between populations
- However, the interpretation of Fst depends critically on which model applies to the populations of interest
  - Migration rates
  - Time since separation
  - Founder events
- Explicit modeling of population histories allows us to distinguish between different demographic scenarios: see ABC course

Isolation with migration model

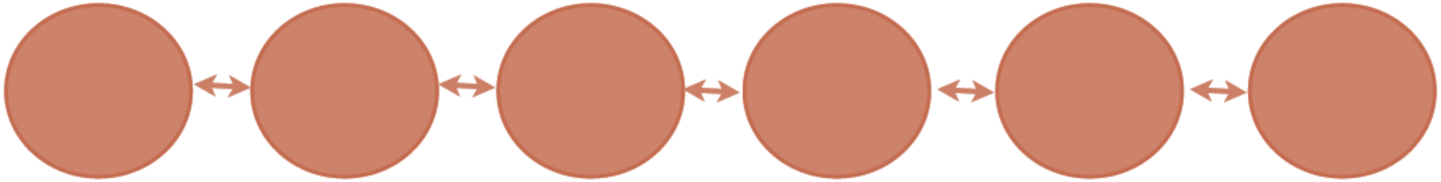


- A complex model with 6 parameters
- Still easy to simulate but…
- Difficult to infer parameters
  - Old split with high migration reassembles to a recent split with low migration

Stepping stone model



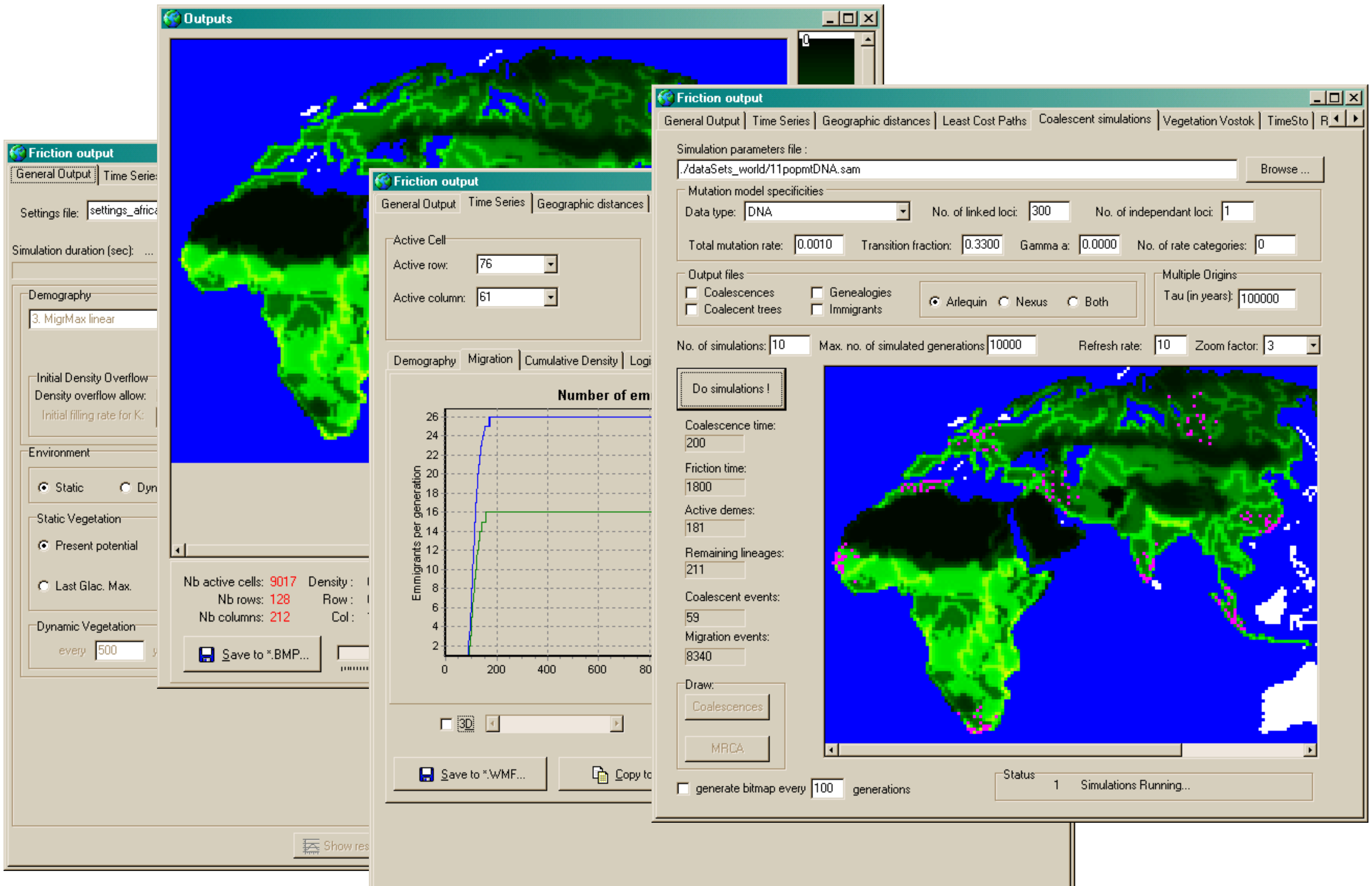
- Grid representation of a continuous habitat in 1D or 2D
- Migration occurs only between adjacent demes in the grid
- Limited analytical results in 1D (Wilkins and Wakeley 2002):
  - Expected coalescence time increases with geographical distance between genes
  - Ancestor biased toward the center
  - Border tend to isolate genes



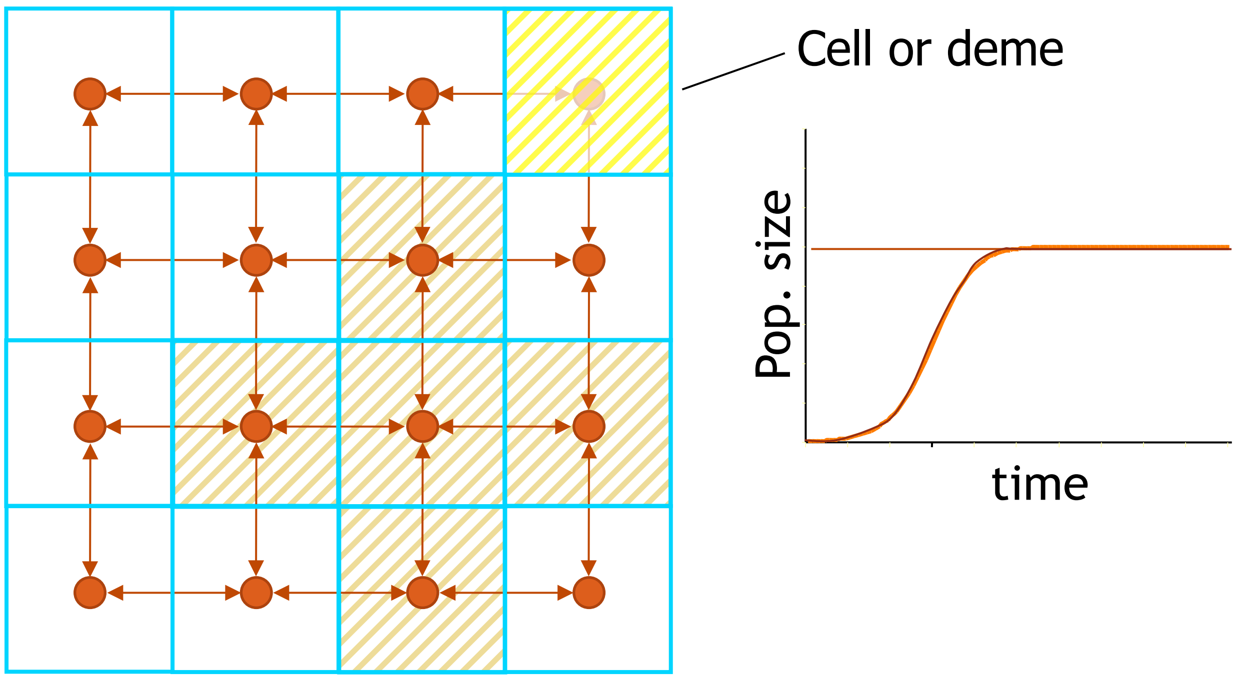
SPLATCHE

SPatia**L** **A**nd **T**emporal **C**oalescences in **H**eterogeneous **E**nvironment

http://www.splatche.com/



Demographic simulations

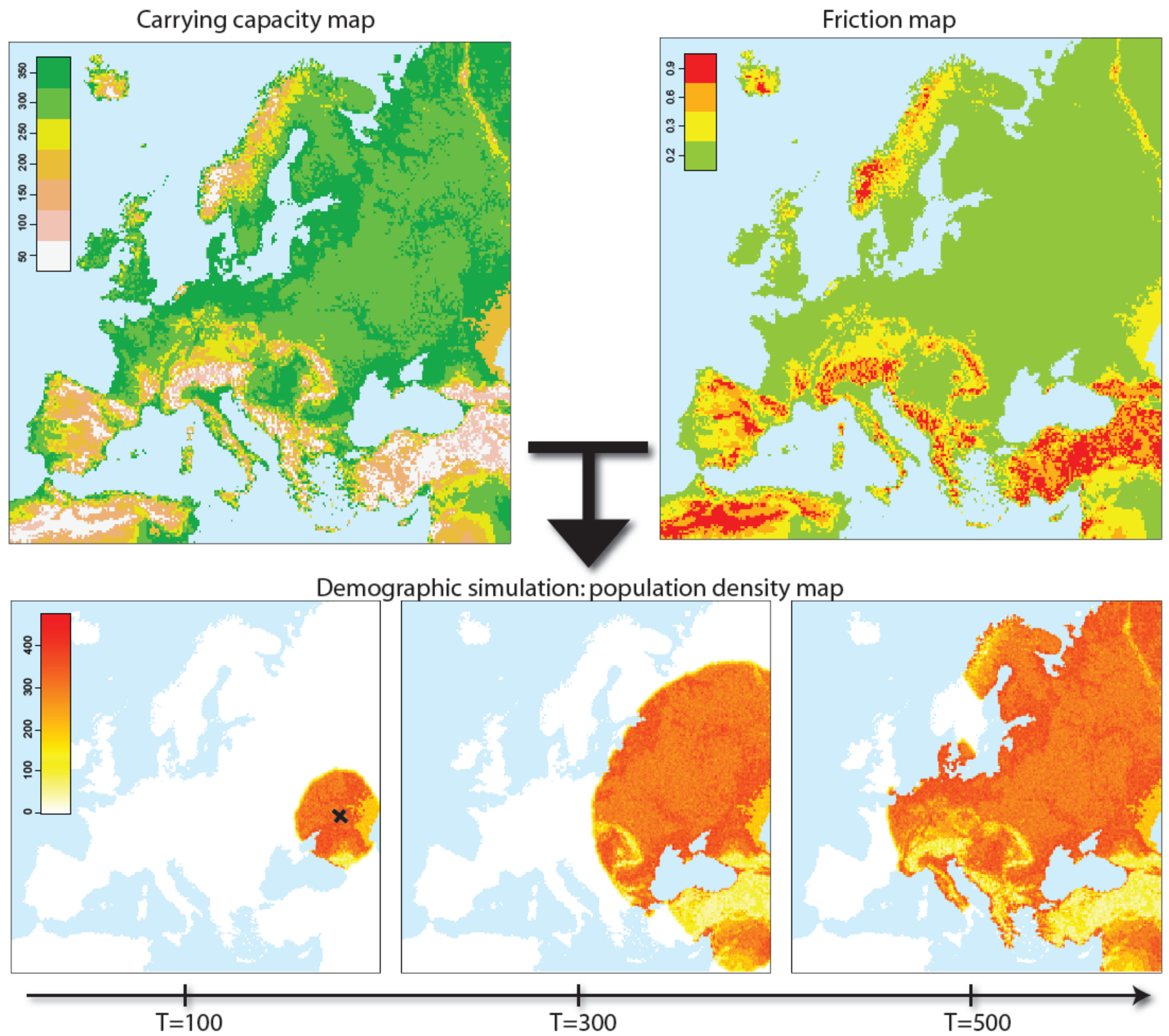


Emigration and growth

- N_t = size of deme at time t
- Distribute m N_t emigrants to 4 nearest neighboring demes.
  - Controlled through friction values (f_i), for each deme.
  - **Relative** difficulty of moving through a deme
  - Multinomial with parameters:

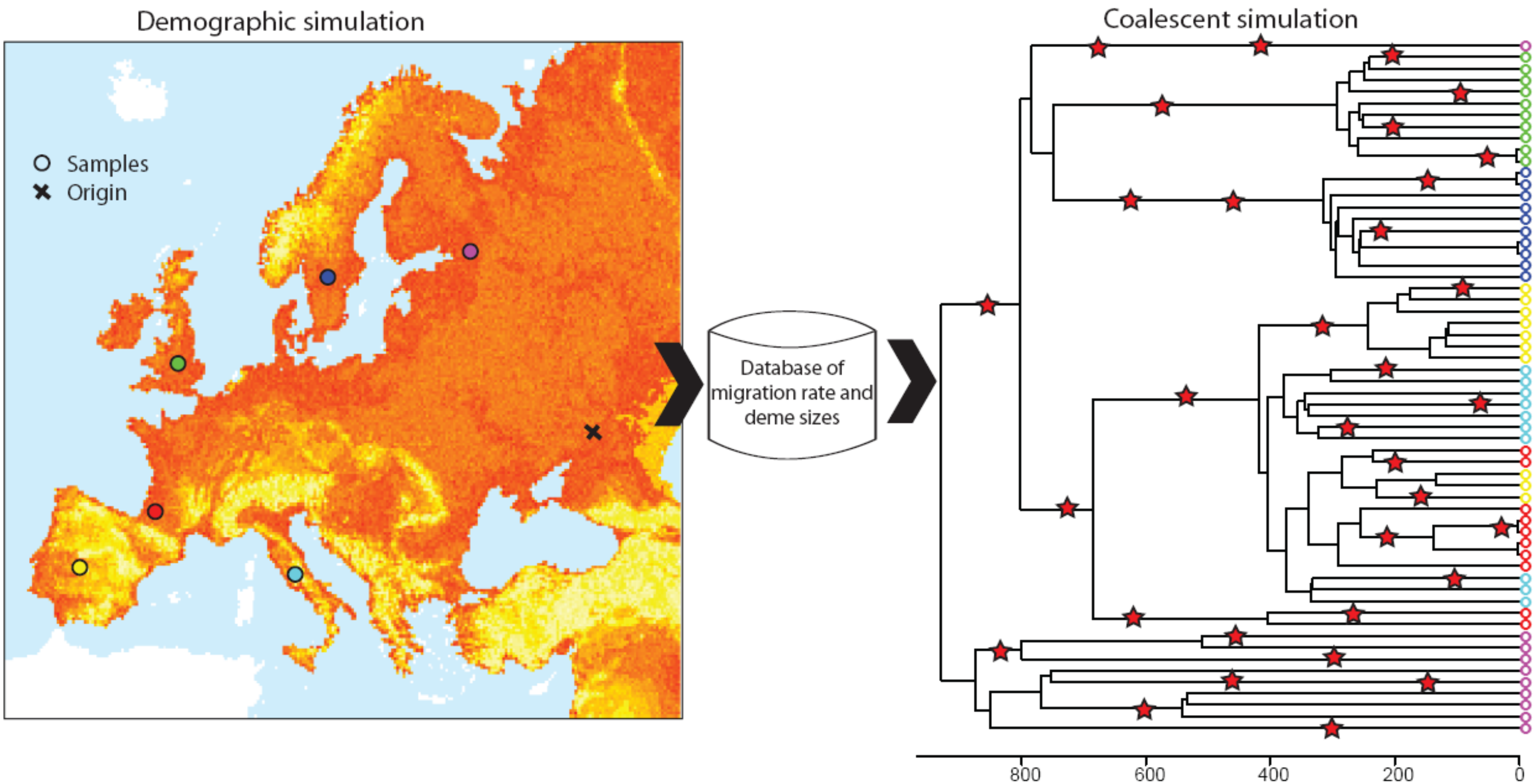
$$p_i = \frac{1}{f_j \sum_{j=1}^{n} \frac{1}{f_j}}$$

- Logistic growth for each deme: $N_{t+1} = N_t \left(1 + r \frac{K - N_t}{K} \right)$
  - K: carying capacity
  - r: growth rate

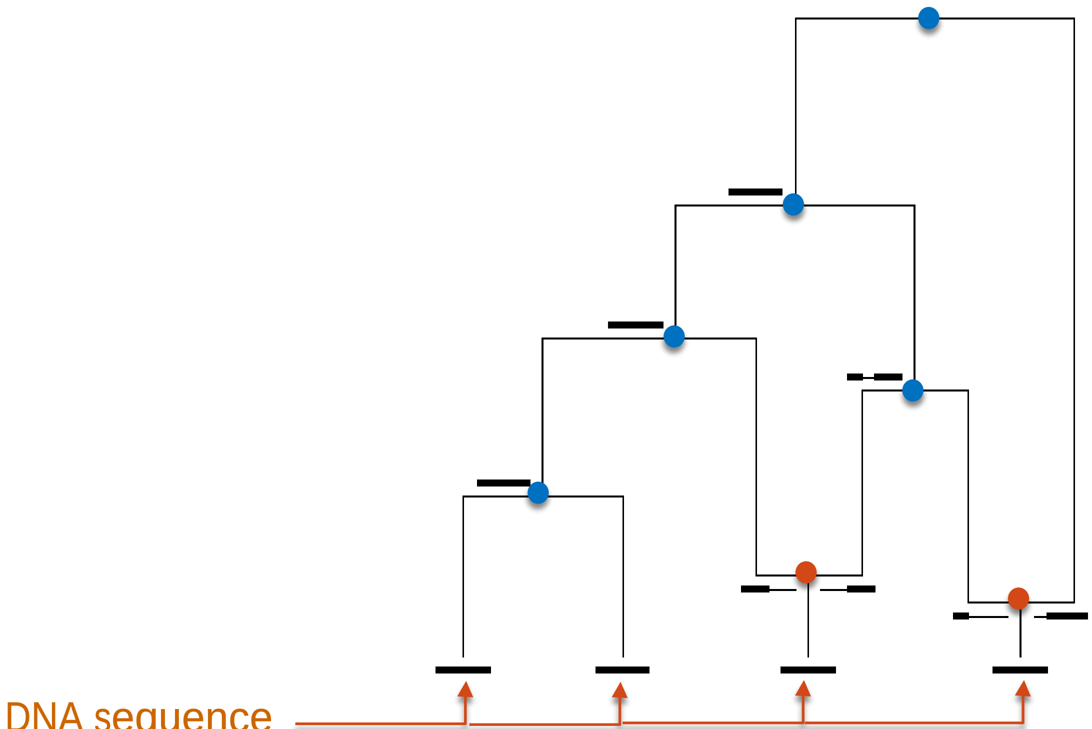


Carrying capacity map
350
300
250
200
150
100
50
Friction map
0.9
0.6
0.3
0.2
Demographic simulation: population density map
400
300
200
100
0
T=100
T=300
T=500

Genetic coalescent simulation



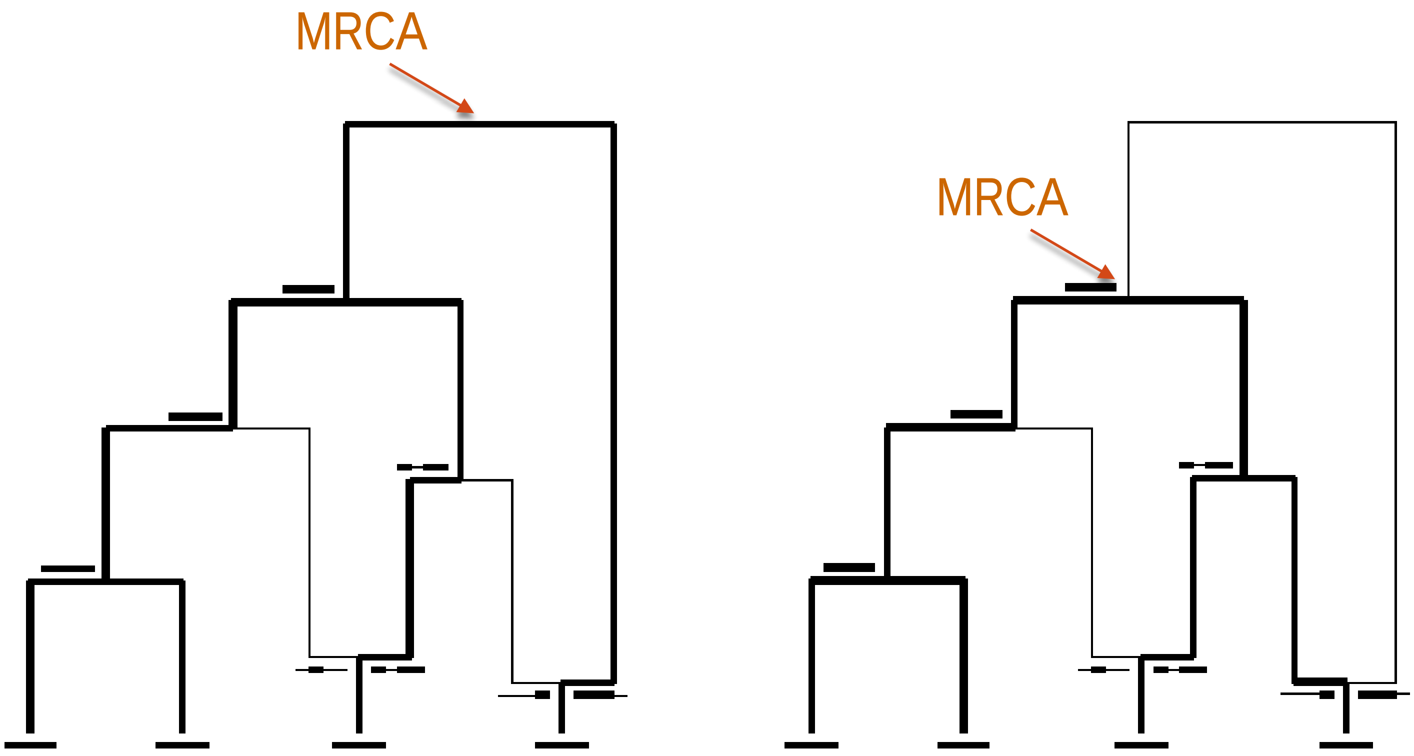
Ancestral Recombination Graph (ARG)



The ancestral recombination graph records **all** recombination and **all** coalescent events having occurred in the ancestry of some observed gene lineages (thick lines).

The probability of a recombination event at any time is $j(L-1)r$ where *j* is the number of remaining lineages, *L* is the sequence length (bp) and *r* the recombination rate between adjacent nucleotides.

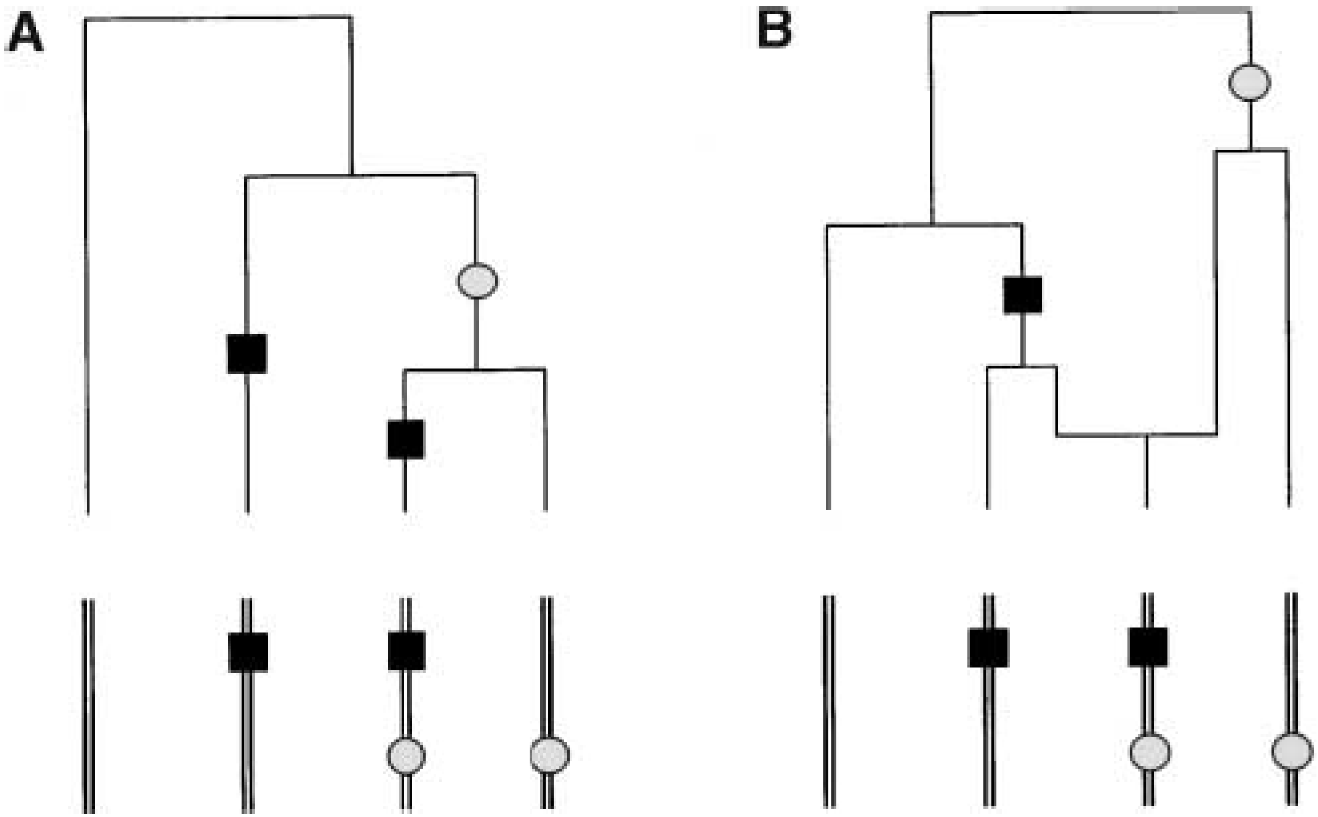
Different coalescent trees for different positions



Coalescent tree at first position

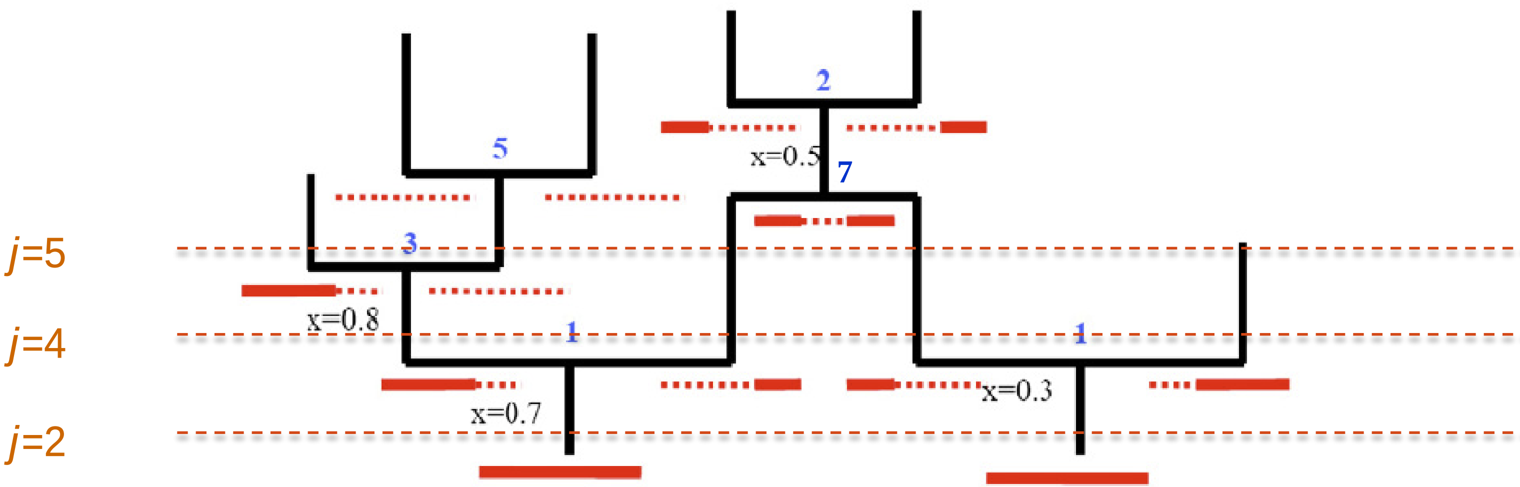
Coalescent tree at last position

Recurrent mutations and recombinations can lead to the same pattern of polymorphism



McVean, Awadalla, Fearnhead 2002 Genetics

Not all events have effects in an ARG



Event types:

1. Recombination in ancestral material
2. Recombination in non-ancestral material that has ancestral material to both sides
3. Recombination in non-ancestral material that has ancestral material only to the left
4. Recombination in non-ancestral material that has ancestral material only to the right
5. Recombination in an individual that carries no ancestral material
6. Coalescent event including a chromosome with only non-ancestral material
7. Coalescent between chromosomes carrying ancestral material

Source: Marjoram and Wall (2006)

ARG is not a very efficient way to model recombination

- We model (follow) many lineages that have no impact on current levels of diversity
- The number of gene lineages to follow can explode if recombination rate is large.
- This leads to high memory requirements and is computationally inefficient

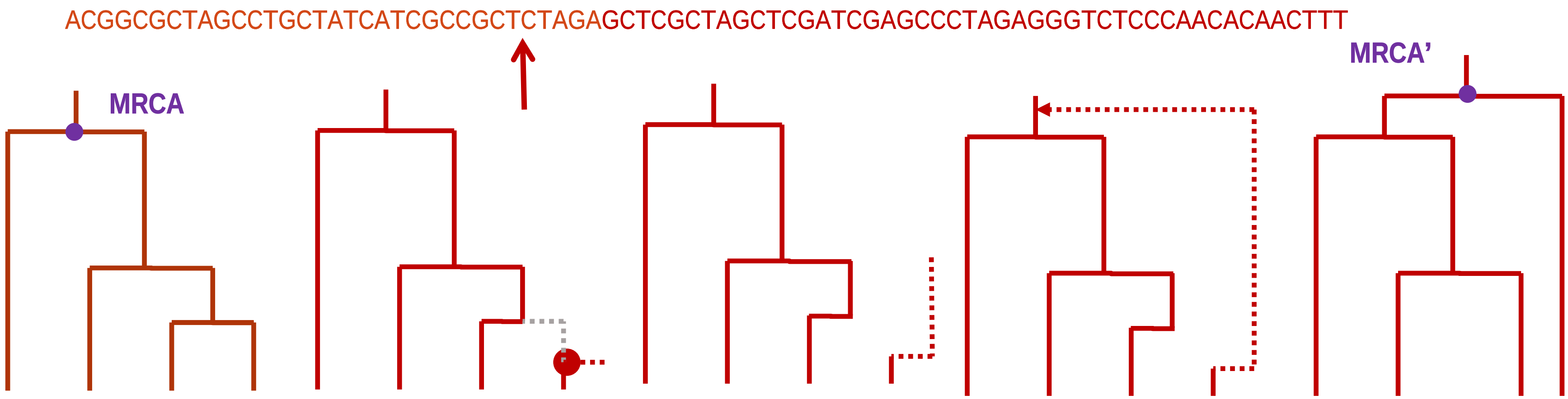
Need of alternative and more efficient algorithms:

- Sequentially Markov Coalescent (McVean and Cardin 2005)

  The idea is to simulate a different tree for each non-recombining segment instead of the whole ARG

SMC algorithm

- Generate a tree at the leftmost end of a DNA sequence of *L* nucleotides

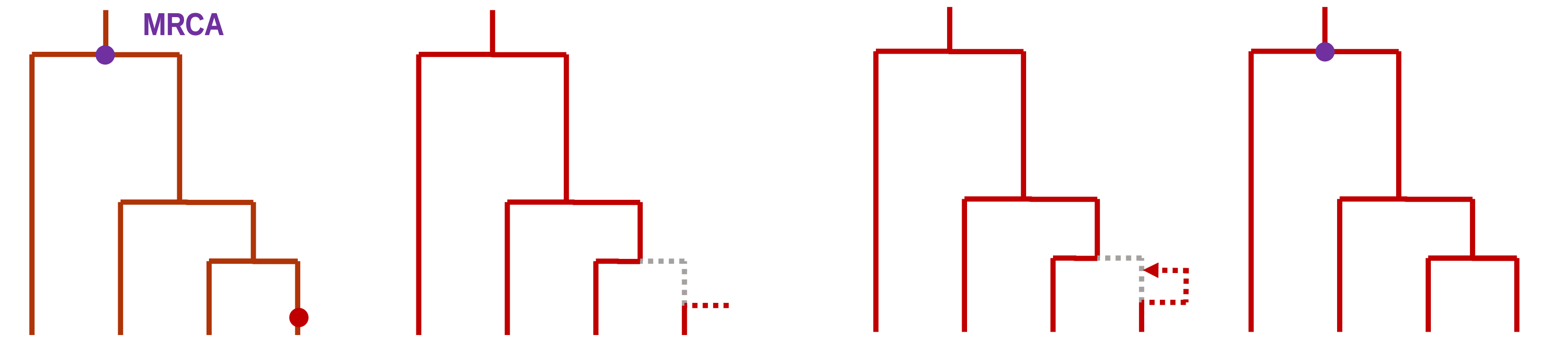


- A recombination event can occur between any adjacent nucleotide with probability $r\,T_1$, where r is the recombination rate per nucleotide and T_1 is the total length of the first genealogy.
- Find the position of the next recombination by drawing a random exponentially distributed number $exp(r\,T_1)$
- Generate a recombination event on the tree by drawing a random number uniformly on 1.. T_1
- Remove the lineage belonging to the left tree and implement a normal coalescent process for this new recombining lineage until it coalesces with another lineage
- Note that the new tree has a potentially different topology and MRCA

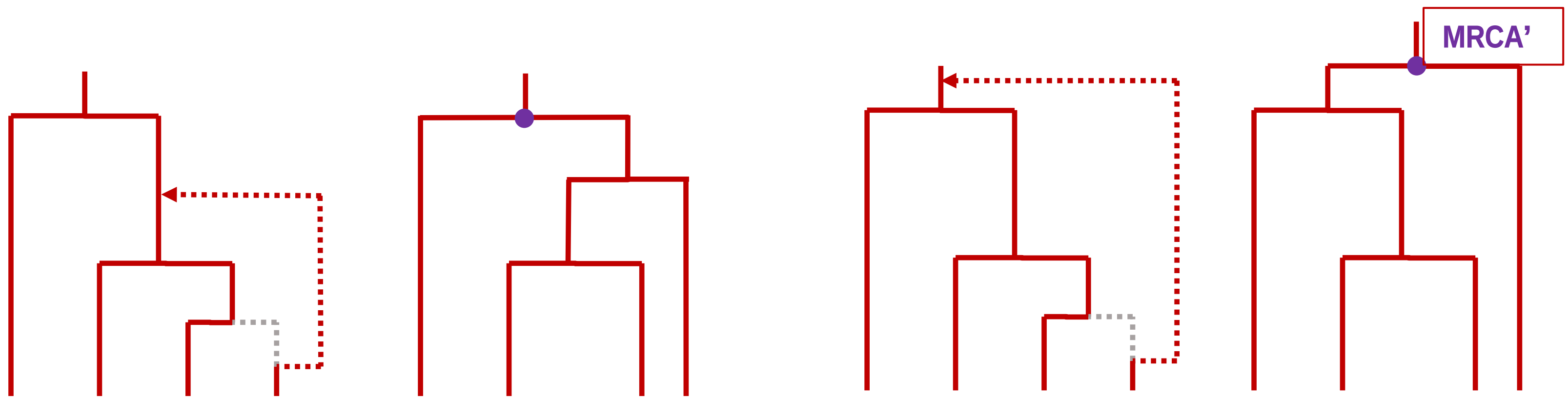
Recombination approximations

Modified sequentially Markov coalescent (SMC')

Marjoram, P. and Wall, J.D. 2006. Fast "coalescent" simulation. *BMC Genet. 7: 16.*



Same topology, same MRCA

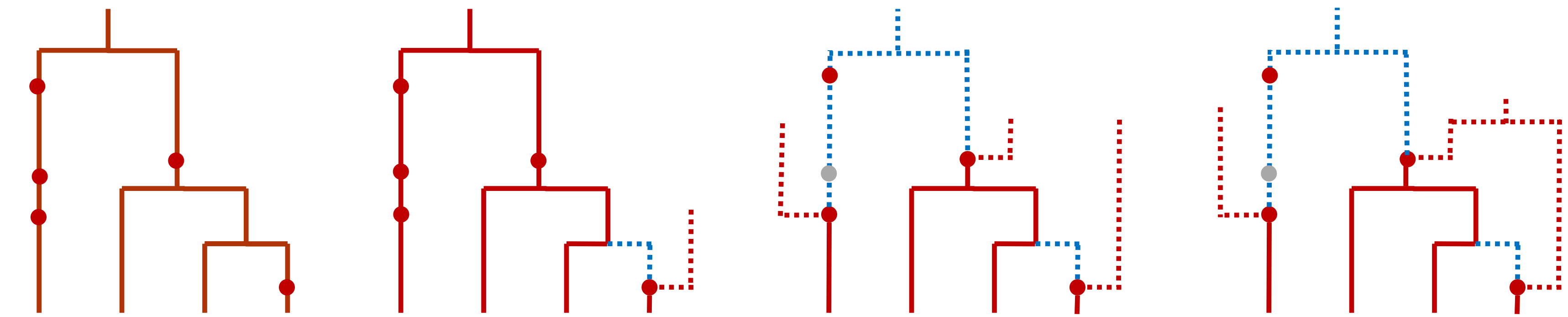


Different topology, same MRCA

Different topology, different MRCA

fastsimcoal

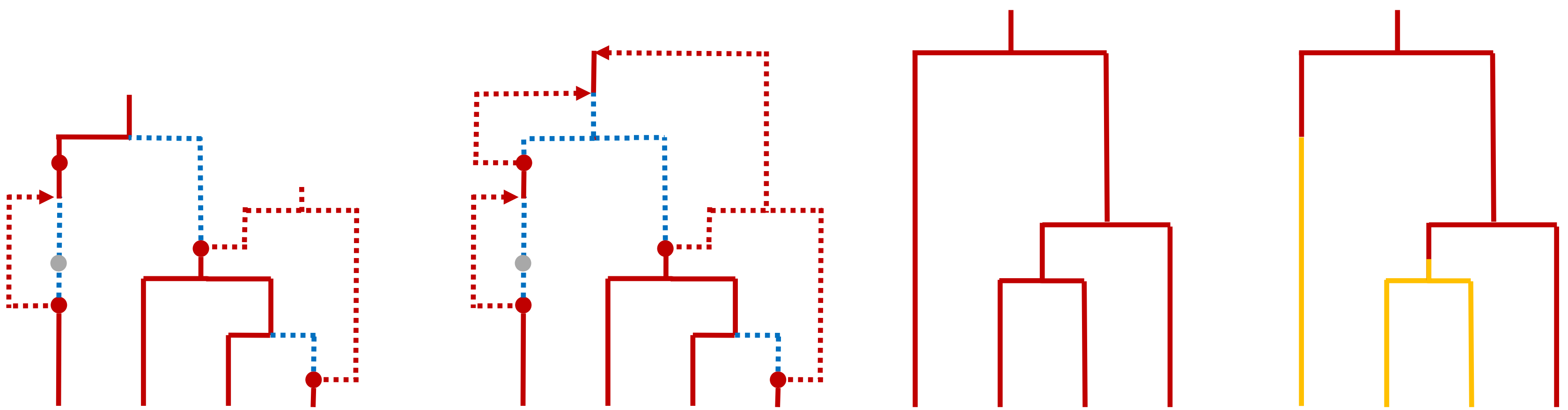
- Uses the SMC' algorithm for DNA sequences
- Uses a multiple recombination approach for other data types with loci simulated at variable recombination distances:
    - Allows for several recombinations events per tree and per branch



First recombination event

Recombination events on discarded lineages are ignored

Coal between two recombined lineages



Coal between a recombined lineage and a discarded lineage

New tree

Shared lineages between the two trees