Inferring Recombination Graphs with BEAST 2

Ugnė Stolz
cEvo group, D-BSSE, ETH Zürich
Taming the BEAST ONLINE, 11.06.2021

Adapted from previous talks by T. Vaughan!

Mutation vs Recombination

Mutation:

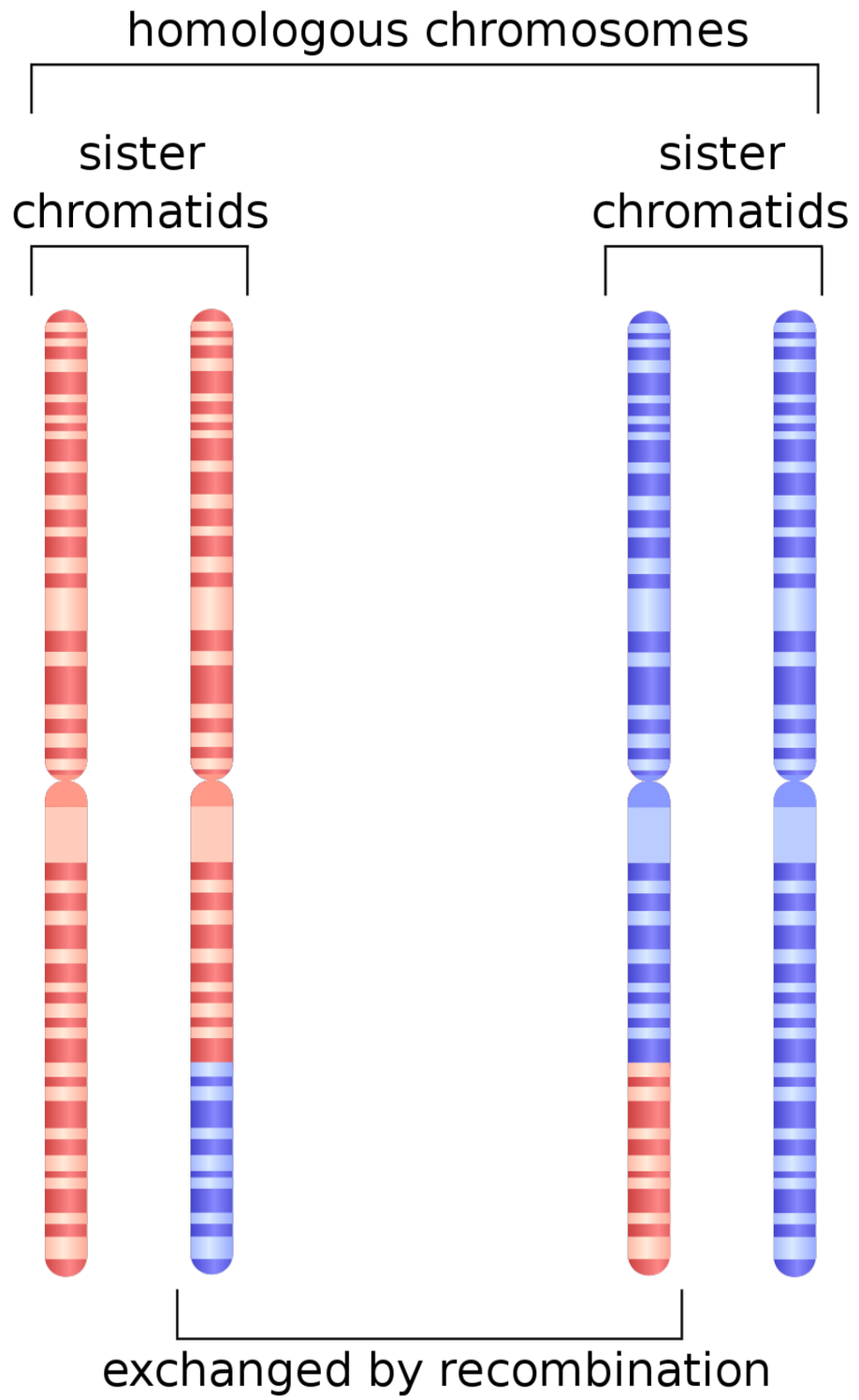
- A change in the nucleotide sequence of a short region of a genome
- Most are mutations are point mutations. Others involve insertion or deletion of one or a few nucleotides.
- Occurs during DNA replication

Recombination:

- The exchange of genetic material either between multiple chromosomes or between different regions of the same chromosome.
- Bring high-scale changes to the genome
- Occurs during gamete production*

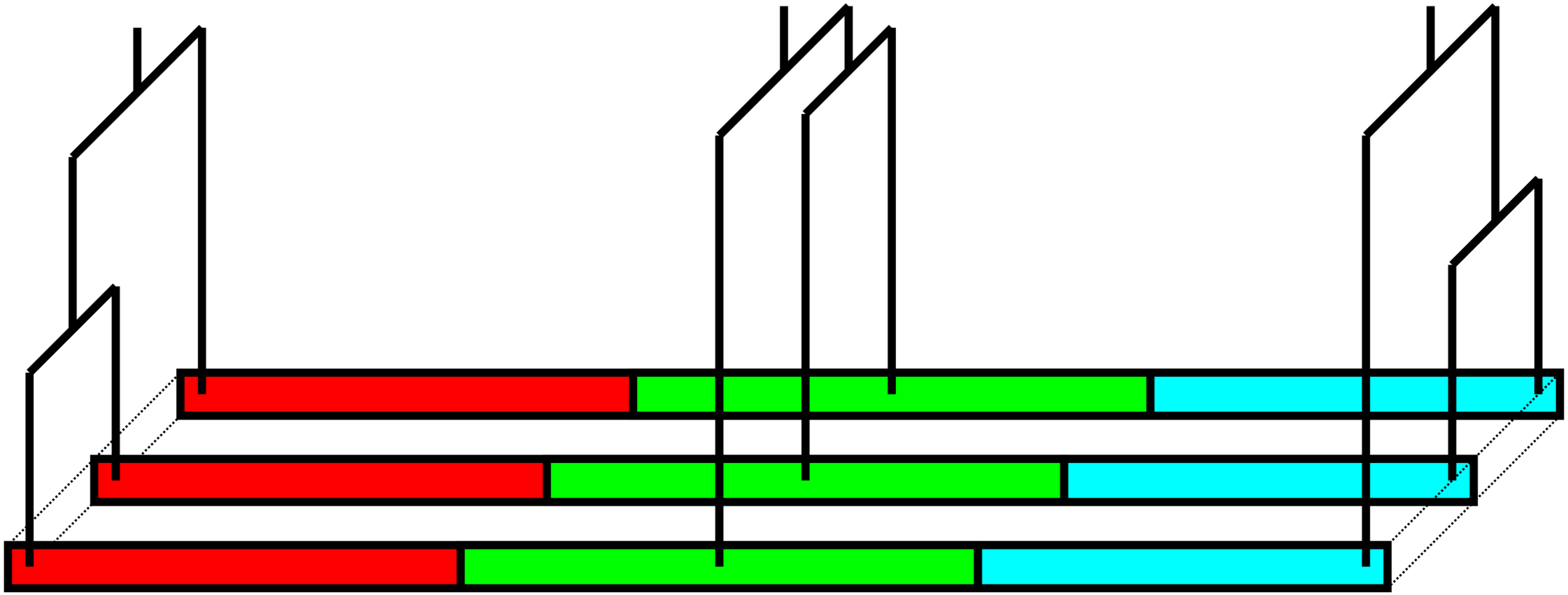
Recombination

- Occurs in eukaryotes, bacteria and viruses.
- Eukaryotic recombination occurs during meiosis via chromosomal crossover.
- Bacterial recombination occurs via
  - ✦ phage-mediated transduction,
  - ✦ natural transformation,
  - ✦ conjugation.
- Viral recombination occurs when multiple strains infect a single cell. For example in RNA viruses:
  - ✦ Reassortment (segmented)
  - ✦ Template switching (non-segmented)



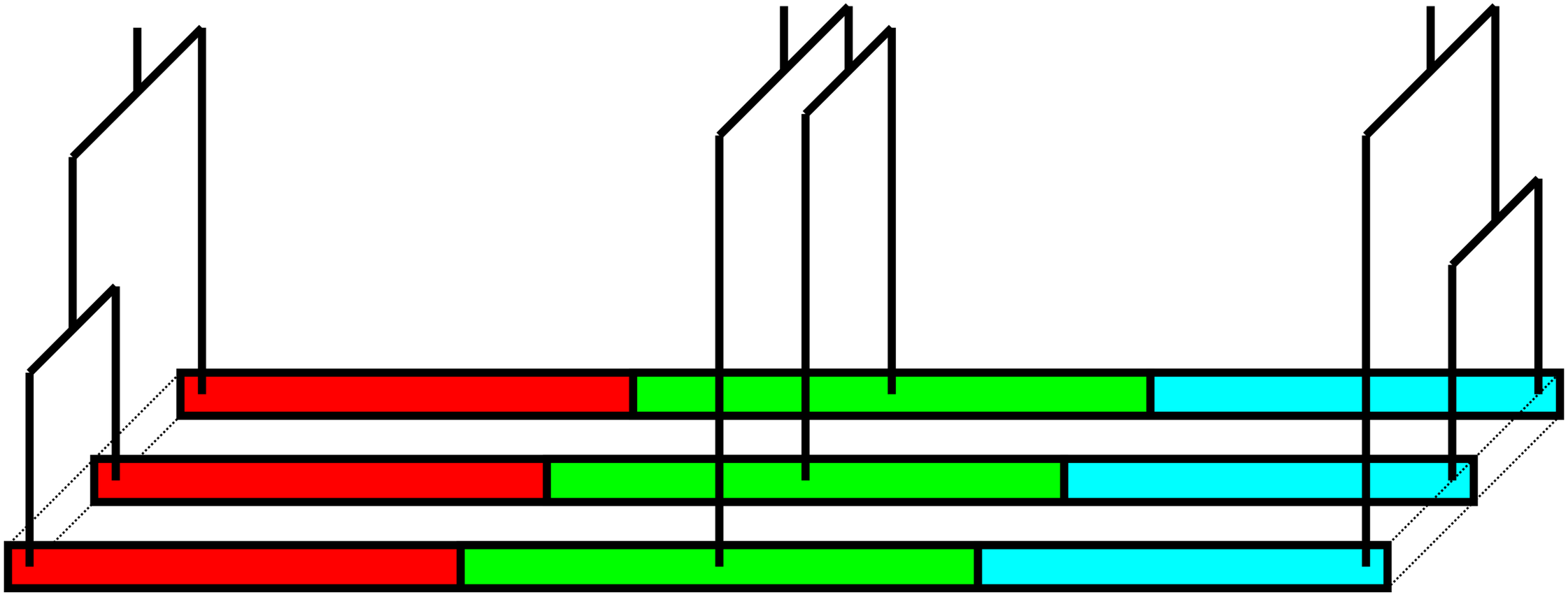
commons.wikimedia.org

Recombination and Phylogenetics



- Different sites correspond to different trees.
- The further away sites are on the alignment, the more likely they are to posses different ancestry.

Recombination and Phylogenetics



Ignoring recombination when present:

- Wrong estimate of ancestral node heights, population size and growth model.
- Overestimation of evolutionary rates.
- Poor mixing (model misspecification).

Schierup and Hein, 2000

Accounting for Recombination

Pre-process the data to identify and remove non-vertically inherited material:

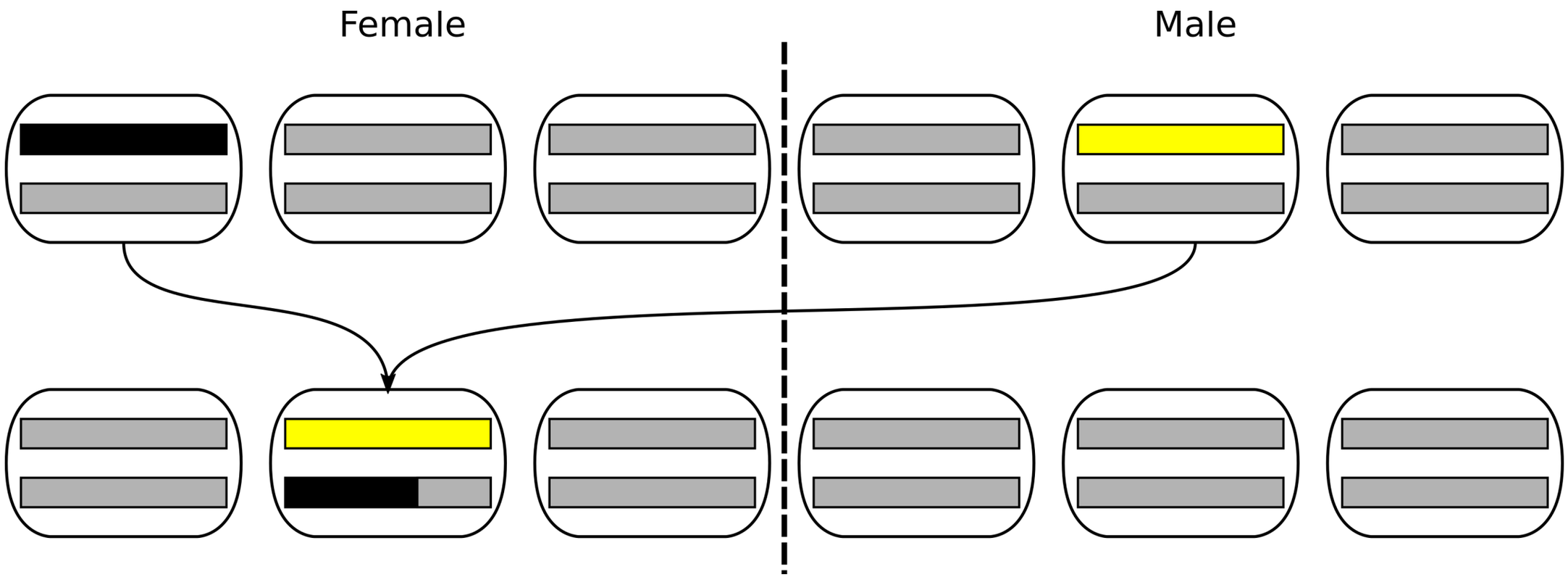
- **Pros**
  - ✦ Can use standard tools for phylogenetic inference.
- **Cons**
  - ✦ Some data is discarded.
  - ✦ May bias results.

Explicitly model recombination:

- **Pros**
  - ✦ Use all data.
  - ✦ Infer recombination rates.
  - ✦ May increase confidence in estimates.
- **Cons**
  - ✦ Complex, many parameters.
  - ✦ Computationally challenging.

Wright-Fisher with Recombination

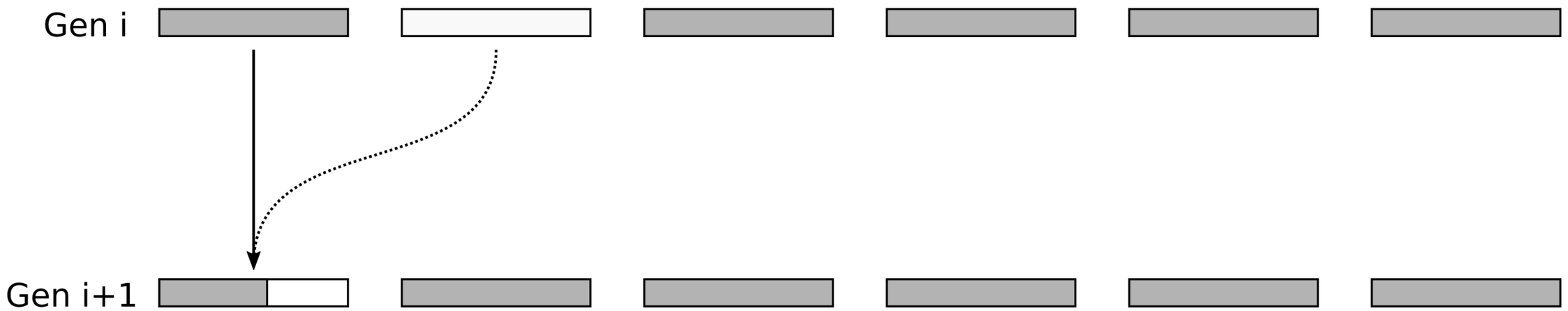
- Consider WF model with male and female diploid individuals.
- Focus on a small segment of the genome.



- Each child selects 1 male and 1 female parent randomly from the previous generation.
- With probability r (which depends on the sequence length) the homologous pair from one of the parents is recombined.

Wright-Fisher with Recombination

- Since the specific pairing of chromosomes only matters over a single generation, in the long term the haploid approximation is very good:

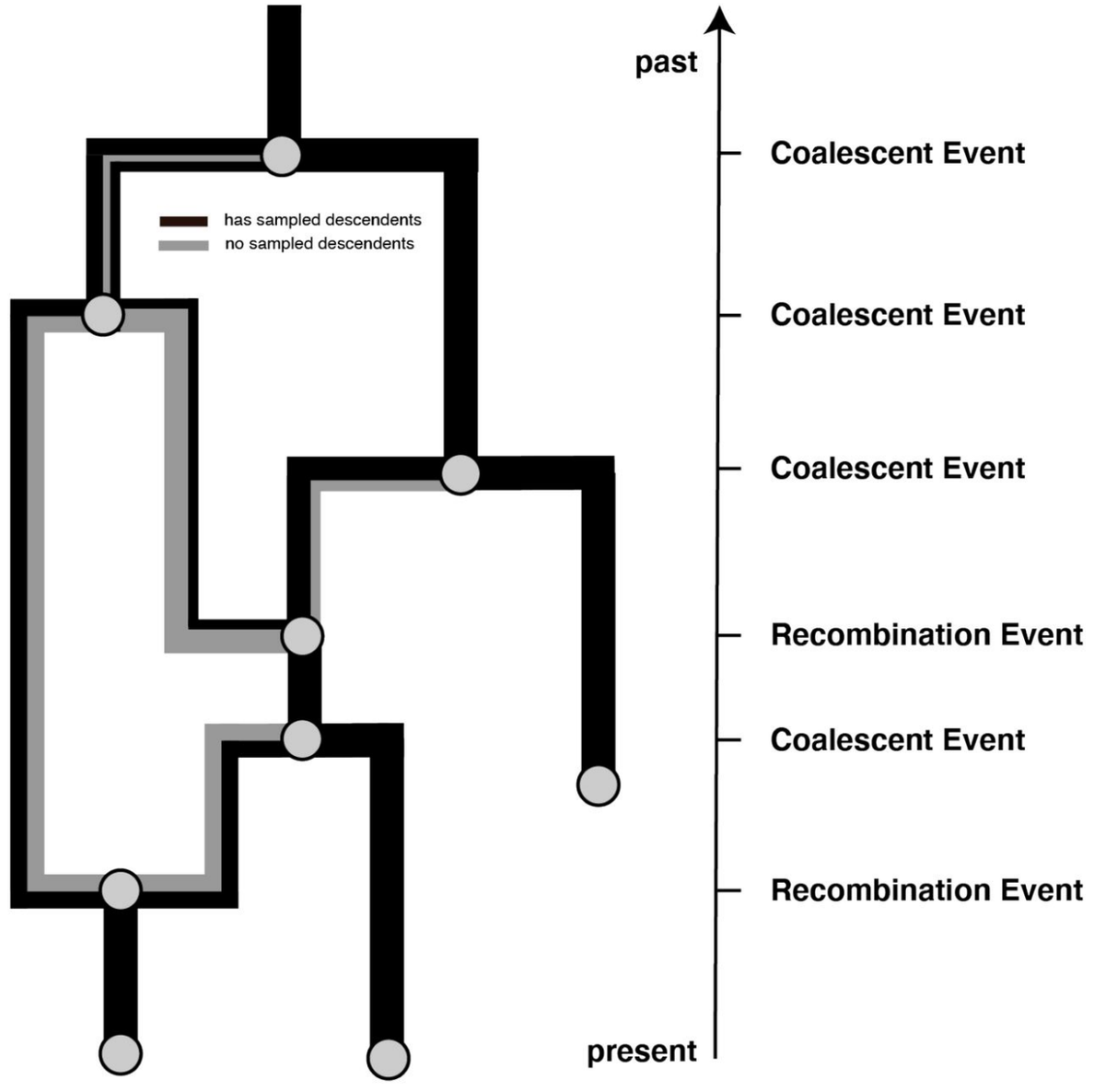


- Each child in i+1 selects a parent at random from generation i.
- With probability r an additional parent is selected.
- In this case, a break-point is chosen randomly on the chromosome, and everything to the right is replaced by the homologous section of the second parent's chromosome.

Coalescent with Recombination

For a fixed recombination rate $\rho = r/g$ in the limit $r \ll 1$, $g \ll 1$ and $N \gg 1$, the genealogical process is the coalescent with recombination (Hudson, 1983):

- **Coalescence rate:** $\binom{k}{2}\frac{1}{Ng}$.
- **Recombination rate:** ρk.
- **Recombination break points:** chosen randomly along sequence: one parent contributes everything to the left, the other everything to the right.



Müller et al. 2021

The result is the ancestral recombination graph (ARG)!

Bayesian phylogenetic network inference

Posterior distribution of a network given a sequence alignment:

$$P(G, \rho, N, Q \mid A) = \frac{1}{P(A)} P(A \mid G, Q) P(G \mid \rho, N) P(\rho, N, Q)$$

- G – recombination graph
- Q – substitution rate matrix
- ρ – recombination rate
- N – effective population size
- A – sequence alignment

Sampling from this distribution is non-trivial (likelihood is invariant under many features of G, surface contains many peaks; huge state space).

Algorithms for Bayesian ARG inference

- **SMARTIE** – MCMC sampler under a "non-informative" network prior rather than CwR. (Bloomquist and Suchard, 2010. BEAST)

- **ARGweaver** – MCMC sampler under a computationally efficient approximation of the CwR. (Rasmussen et al., 2014)

- **ClonalOrigin** – MCMC sampler for Bacterial ARGs under the coalescent with gene conversion (modification of CwR to account for homologous gene conversion). (Didelot et al., 2010)

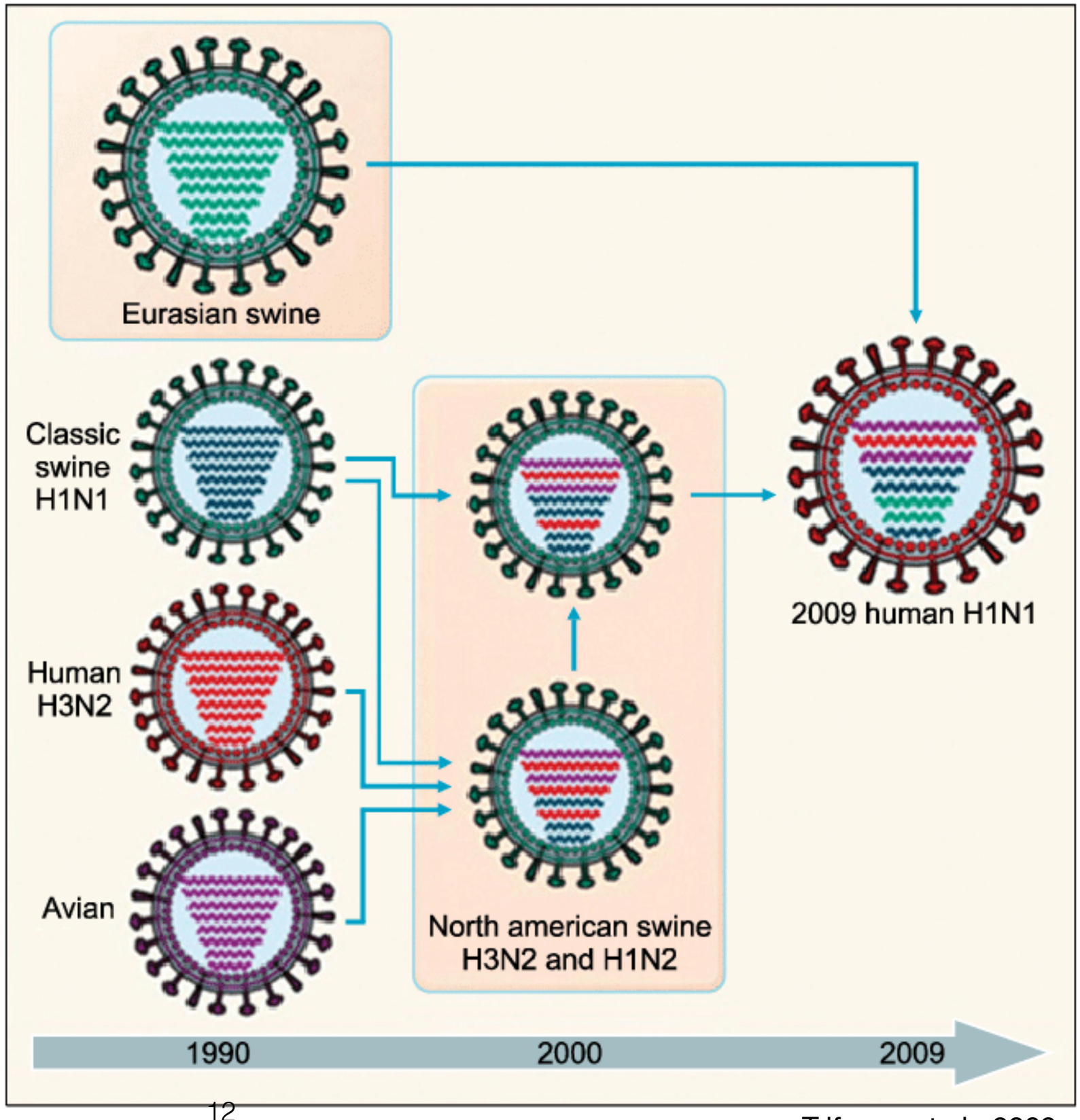
- **Bacter** – Re-implementation of the ClonalOrigin model with fewer restrictions. (Vaughan et al., 2017. **BEAST2**)

- **CoalRe**/**SCoRe** – MCMC sampler for ARGs under coalescent with reassortment/ structured coalescent with reassortment (type of recombination in segmented viruses). (Müller et al. 2020, Stolz et al. 2021. **BEAST2**)

- **Recombination** – MCMC sampler under the CwR. (Müller et al. 2021. **BEAST2**)

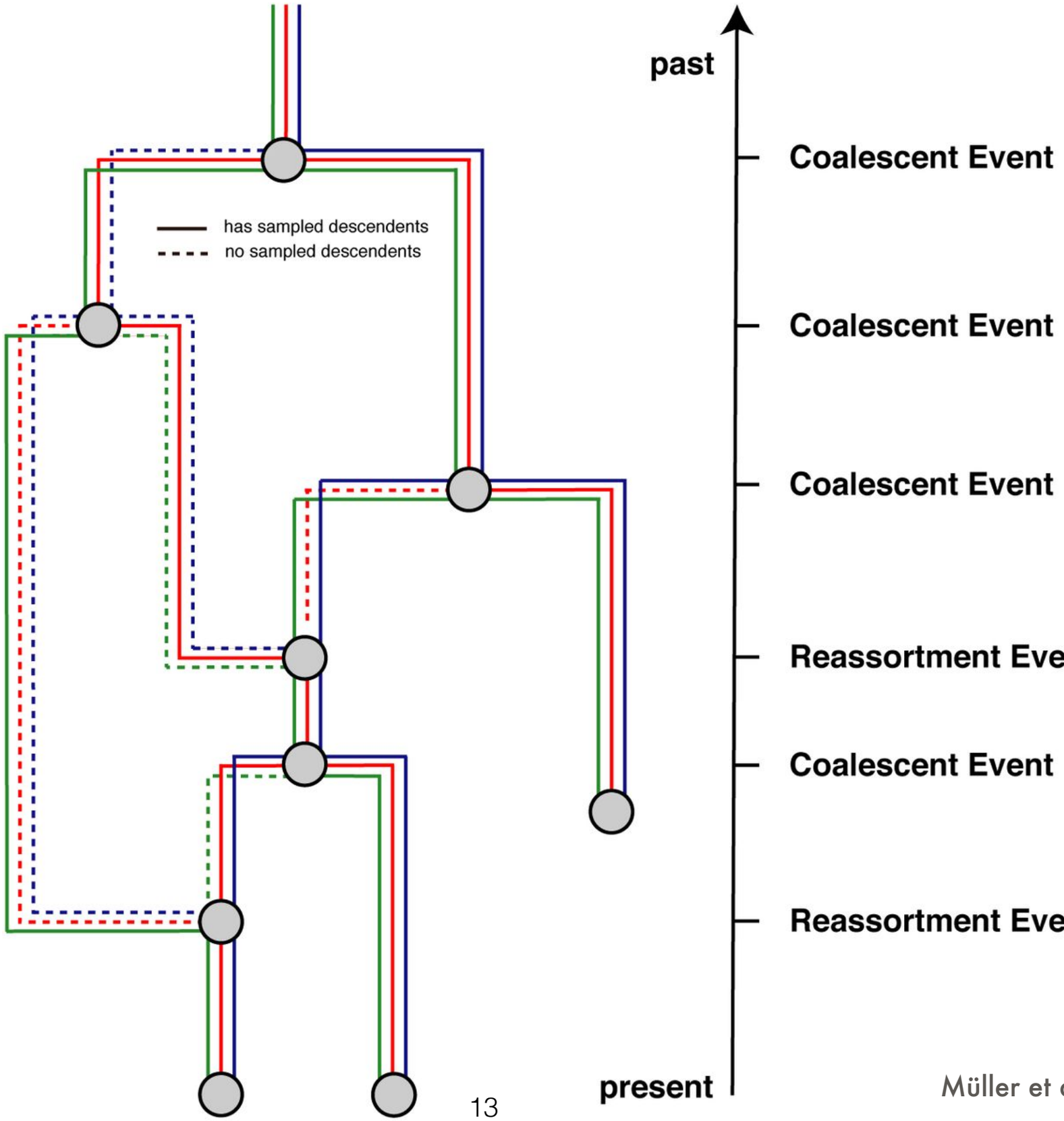
Reassortment

- Happens in segmented RNA viruses, for example influenza, rotaviruses..
- In case of influenza – responsible for previous pandemic strains



Trifonov et al., 2009

Reassortment



Müller et al. 2020

Reassortment tutorial for CoalRe package

https://taming-the-beast.org/tutorials/Reassortment-Tutorial/

If you managed to start the BEAST analysis, use provided output files for further post-processing steps instead of waiting for BEAST inference to complete.

 # t-coalre