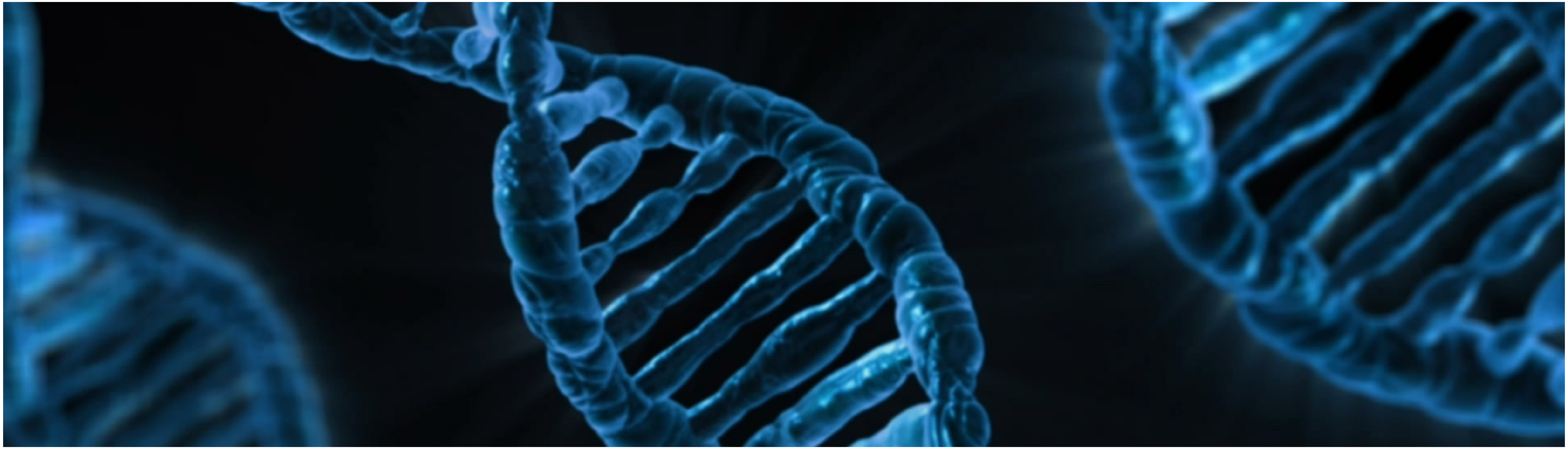
Memoization on Shared Subtrees Accelerates Computations on Genealogical Forests

2024-09-03

Lukas Hübner and Alexandros Stamatakis



Population Genetics

Evolutionary Bioinformatics

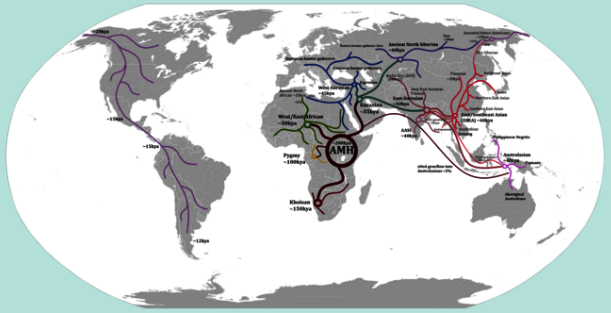
- Living beings organize in a tree modelled based on their genetic code
- Phylogenetics: Evolutionary history among different species
- Genealogy: Evolutionary history among individuals of the same species

applications

understanding evolution

host-parasite interaction

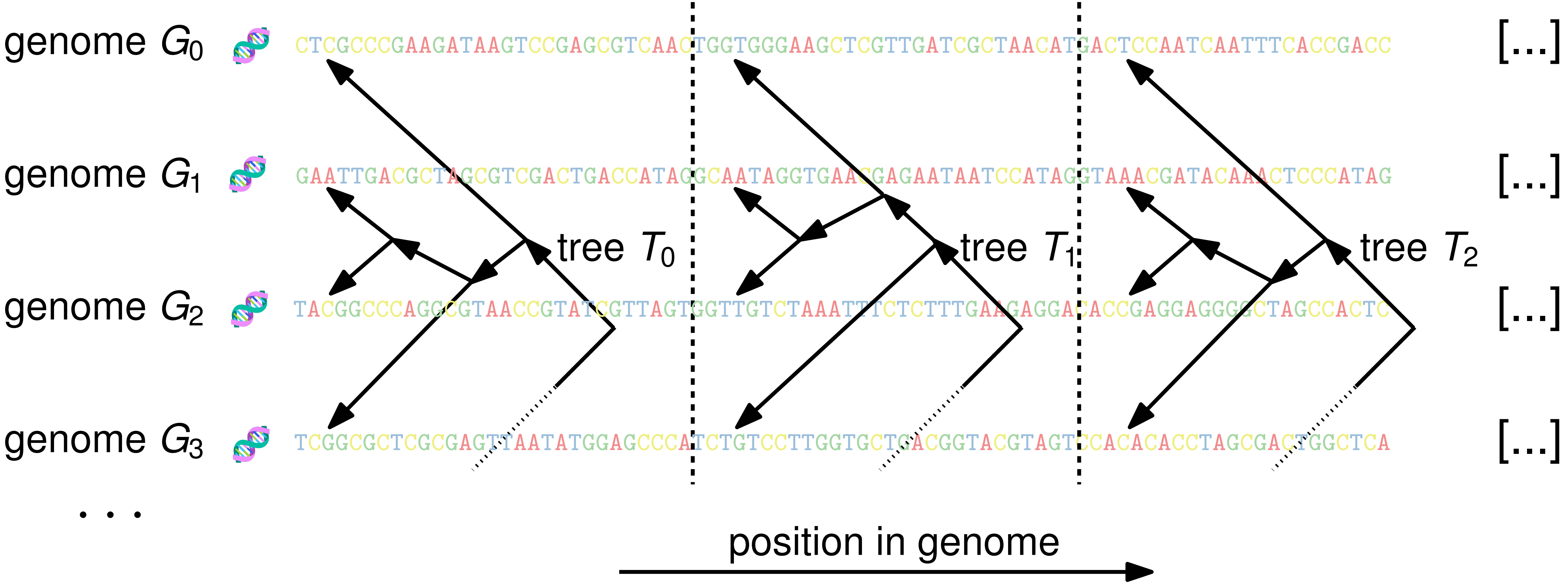
wildlife conservation



human migration patters

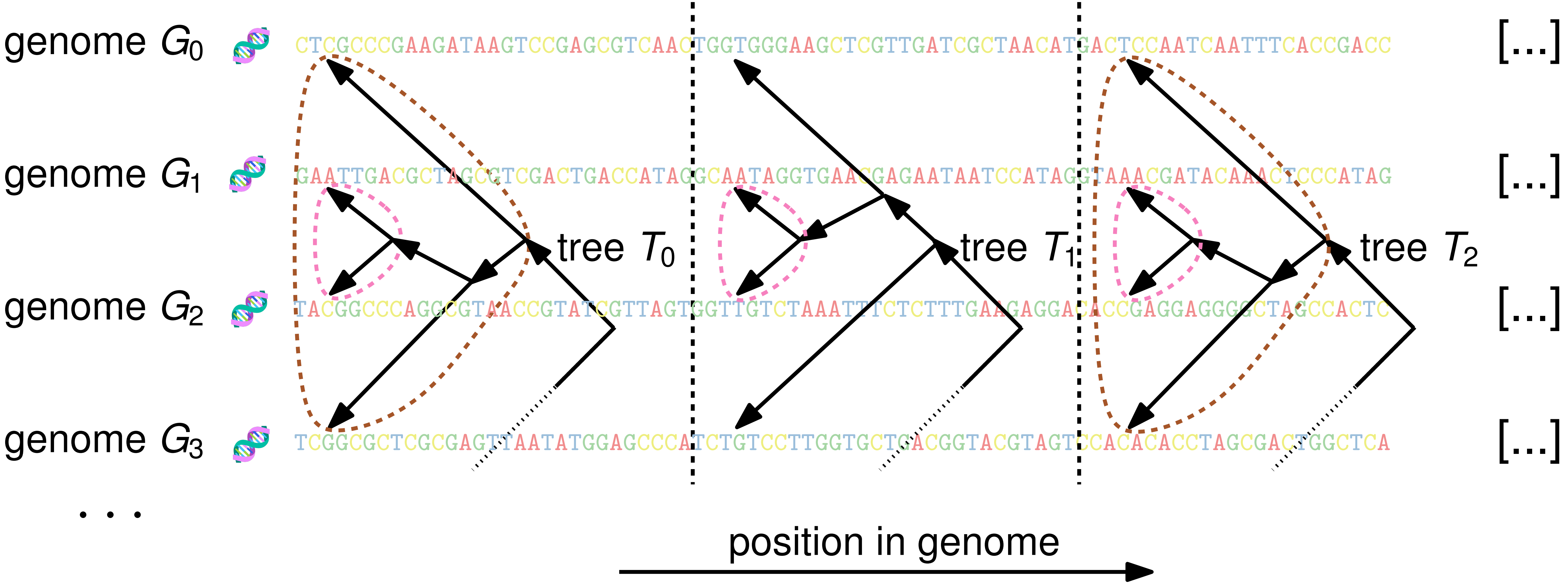
Recombination

- Recombination breaks and recombines genomic code
- Degree of shared evolutionary history between two sites correlates with their distance
- Multiple trees provides a more comprehensive picture

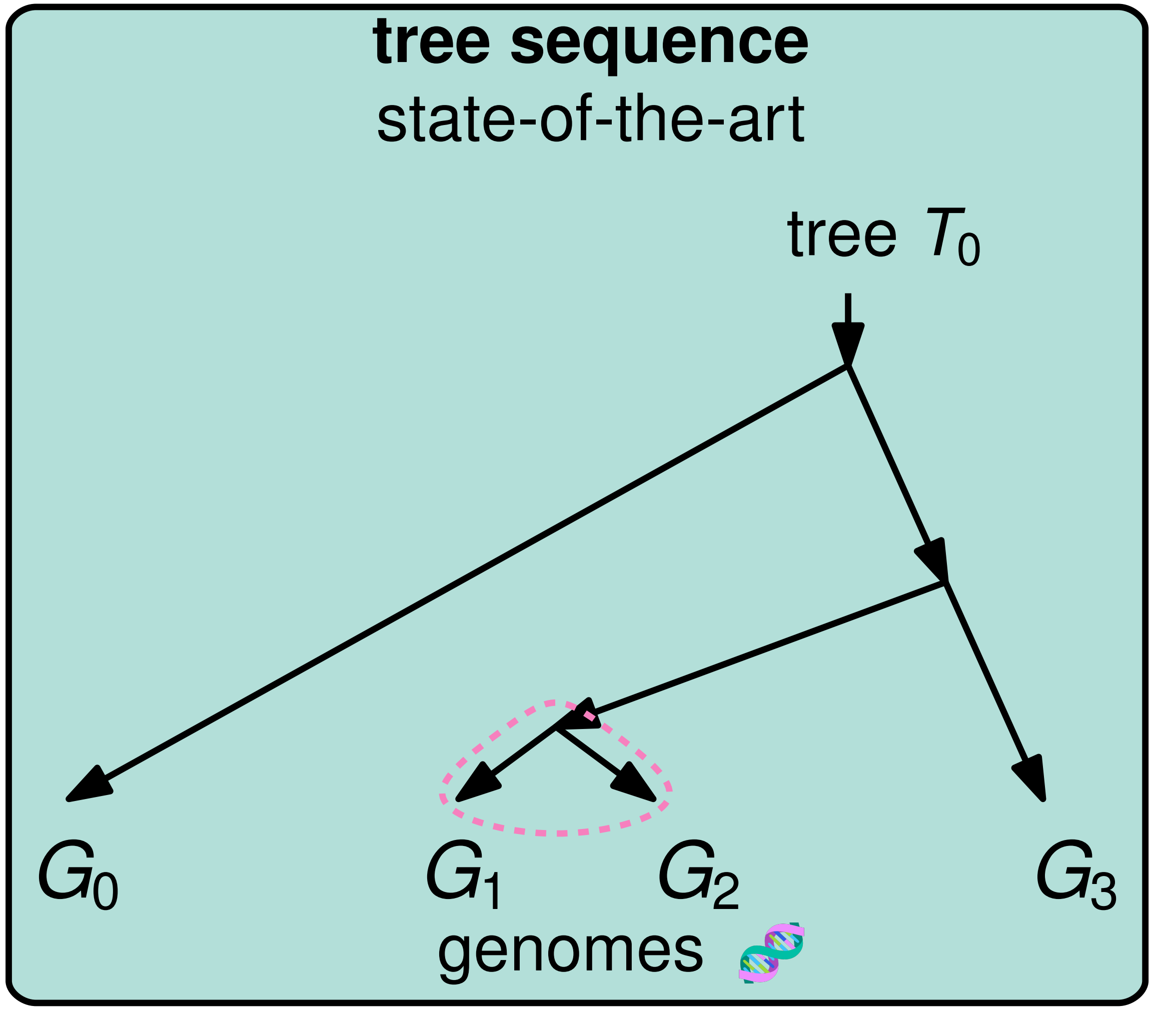

genome G_0
CTCGCCCGAAGATAAGTCCGAGCGTCAACTGGTGGGAAGCTCGTTGATCGCTAACATGACTCCAATCAATTTCACCGACC
[...]
genome G_1
GAATTGACGCTAGCGTCGACTGACCATAGGCAATAGGTGAACGAGAATAATCCATAGGTAAACGATACAAACTCCCATAG
[...]
tree T_0
tree T_1
tree T_2
genome G_2
TACGGCCCAGGCGTAACCGTATCGTTAGTGGTTGTCTAAATTTCTCTTTGAAGAGGACACCGAGGAGGGGCTAGCCACTC
[...]
genome G_3
TCGGCGCTCGCGAGTTAATATGGAGCCCATCTGTCCTTGGTGCTGACGGTACGTAGTCCACACACCTAGCGACTGGCTCA
[...]
. . .
position in genome

Recombination

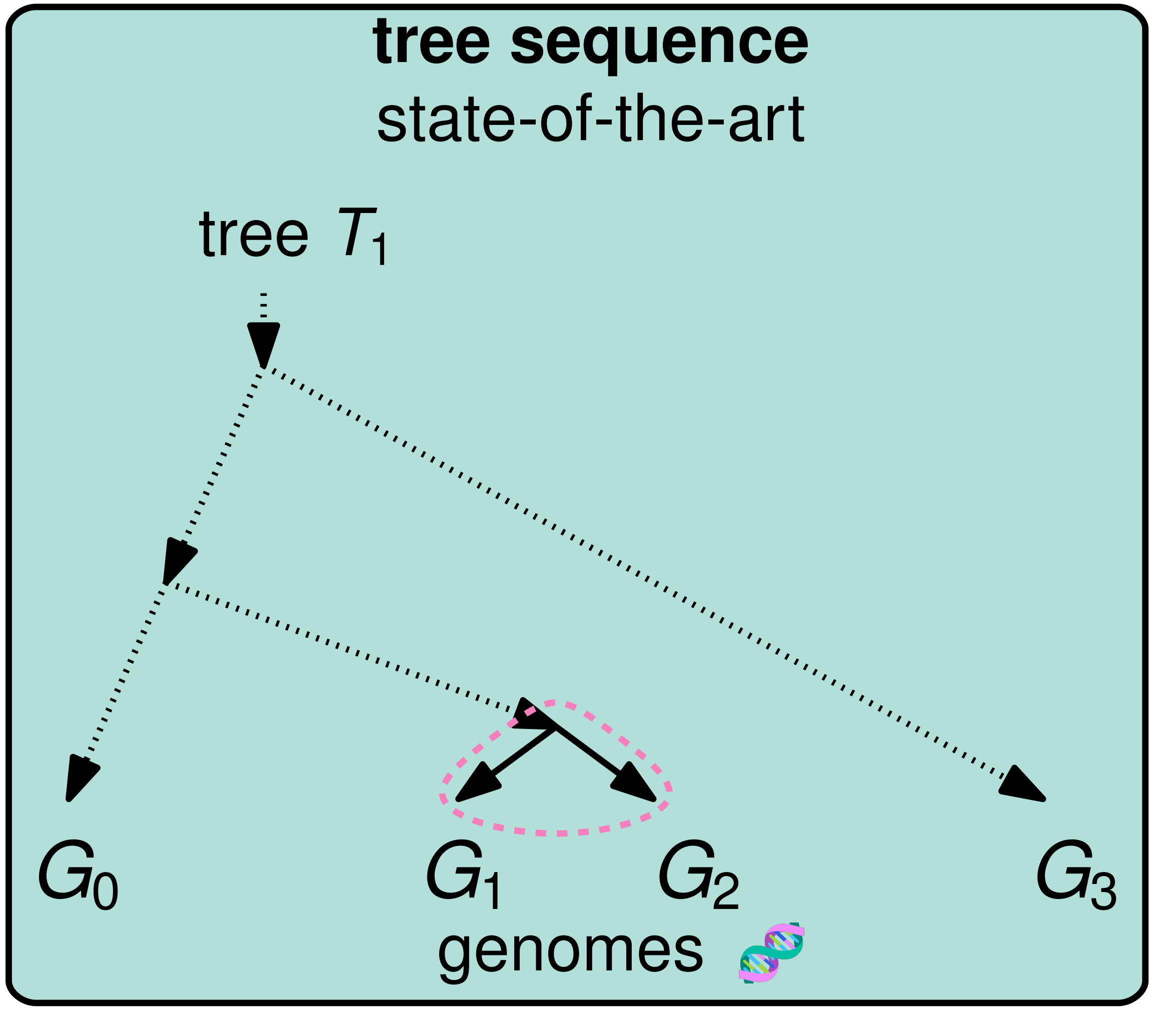
- Recombination breaks and recombines genomic code
- Degree of shared evolutionary history between two sites correlates with their distance
- Multiple trees provides a more comprehensive picture


genome G_0
CTCGCCCGAAGATAAGTCCGAGCGTCAACTGGTGGGAAGCTCGTTGATCGCTAACATGACTCCAATCAATTTCACCGACC
[...]
genome G_1
GAATTGACGCTAGCGTCGACTGACCATAGGCAATAGGTGAATGAGAATAATCCATAGGTAAACGATACAAACTCCCATAG
[...]
genome G_2
TACGGCCCAGGCGTAACCGTATCGTTAGTGGTTGTCTAAATTTCTCTTTGAAGAGGATACCGAGGAGGGGCTAGCCACTC
[...]
genome G_3
TCGGCGCTCGCGAGTTAATATGGAGCCCATCTGTCCTTGGTGCTGACGGTACGTAGTCCACACACCTAGCGACTGGCTCA
[...]
. . .
tree T_0
tree T_1
tree T_2
position in genome

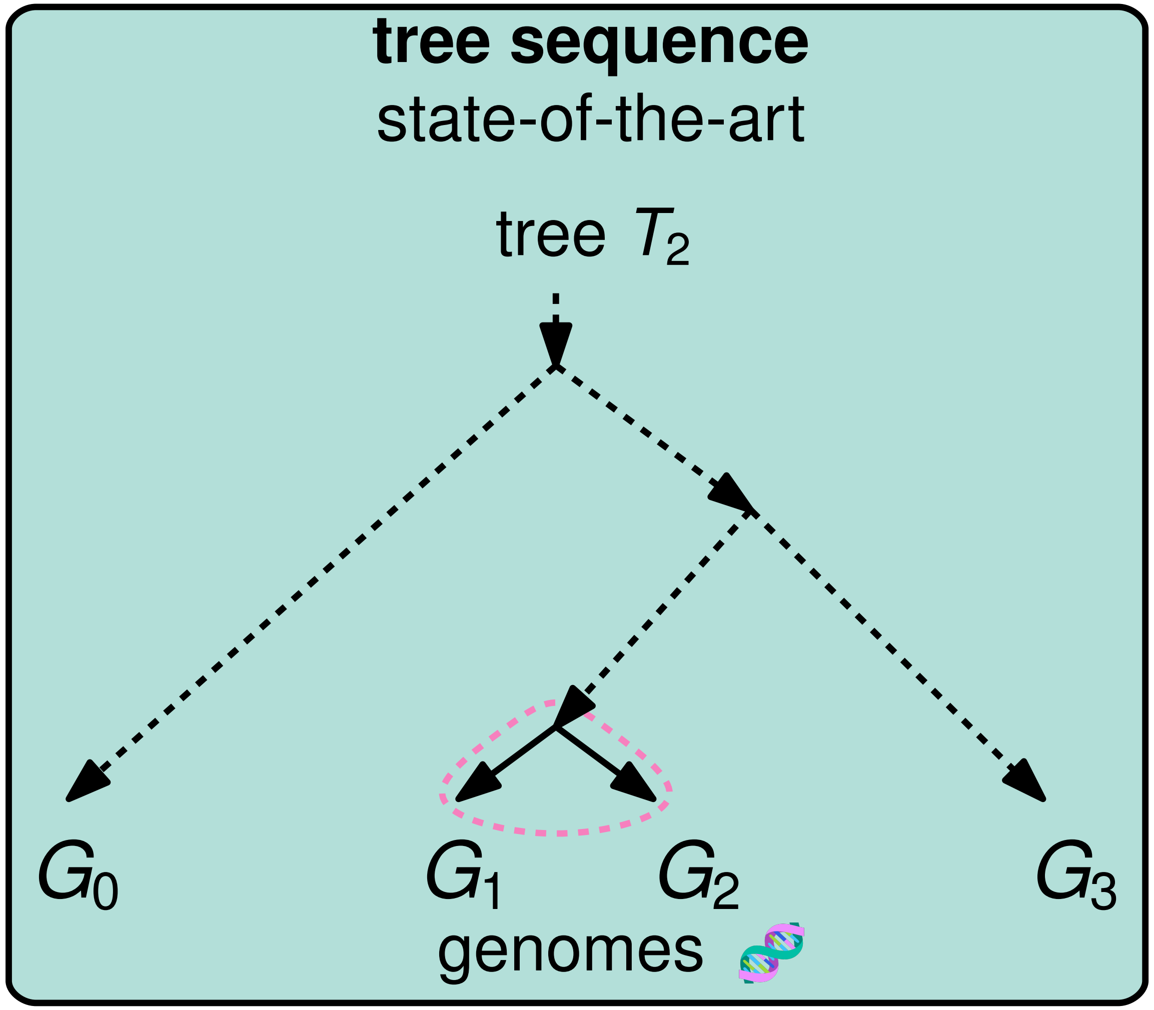
Tree Sequences & Genealogical Forests



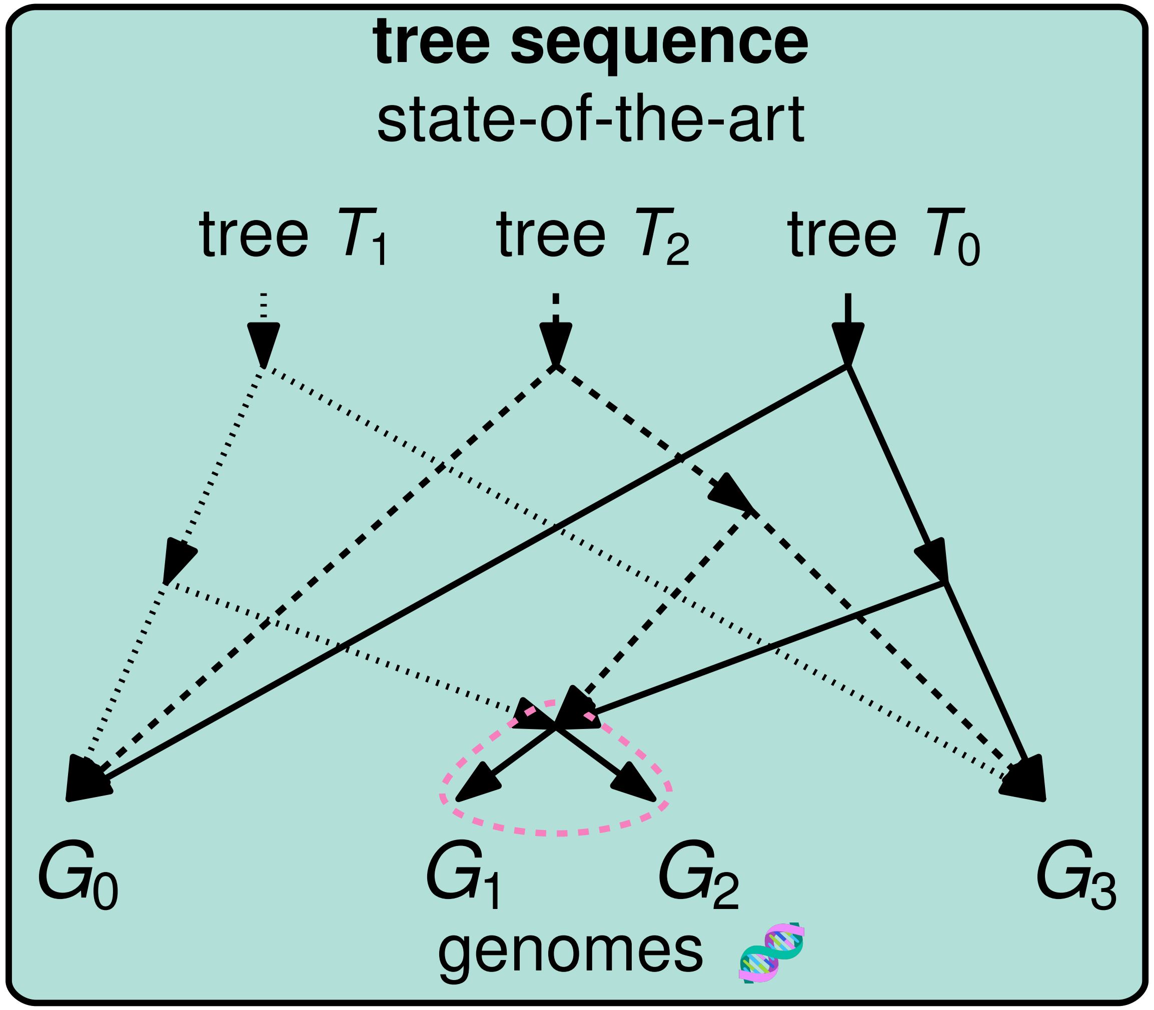
Tree Sequences & Genealogical Forests



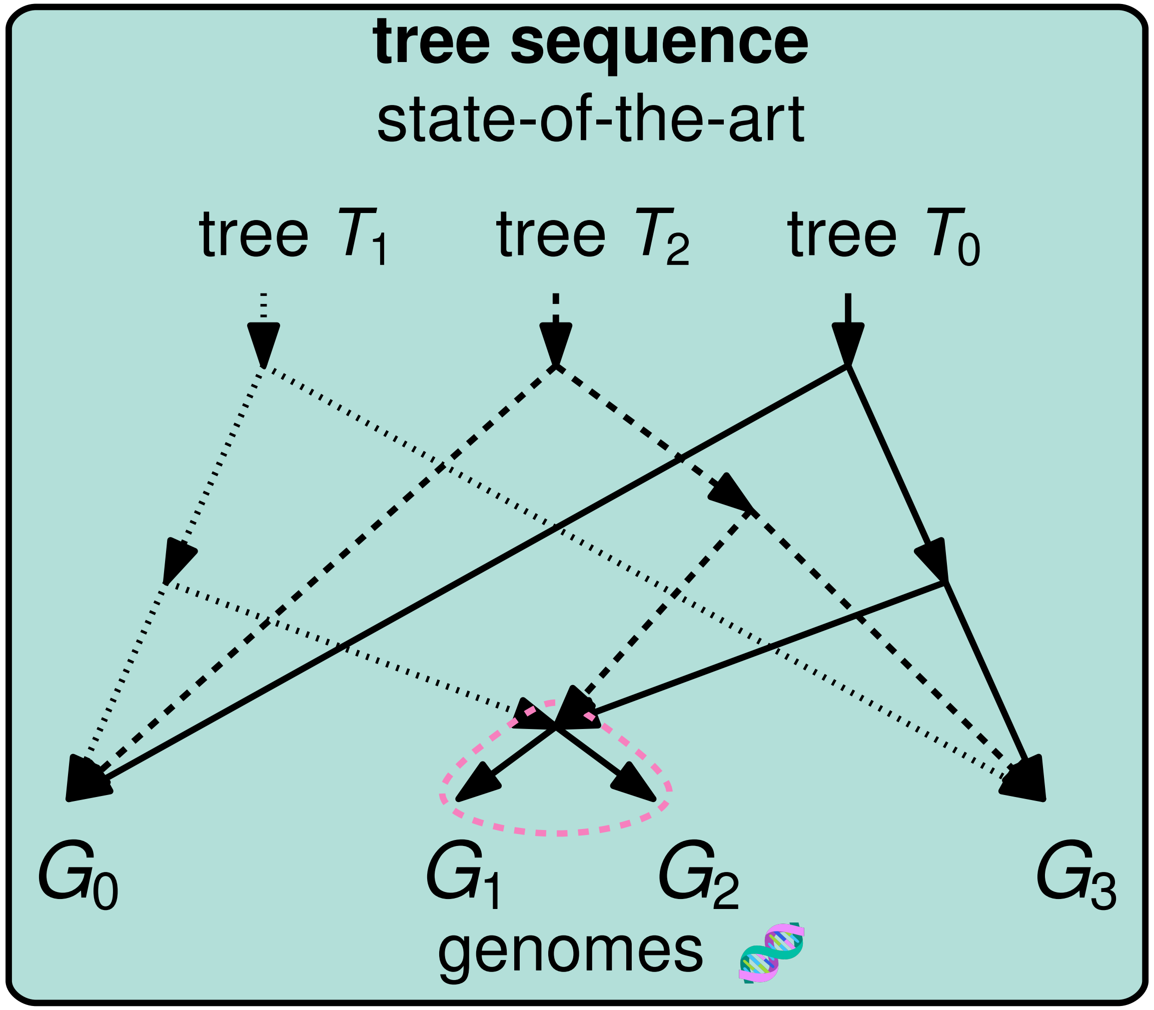
Tree Sequences & Genealogical Forests



Tree Sequences & Genealogical Forests



Tree Sequences & Genealogical Forests



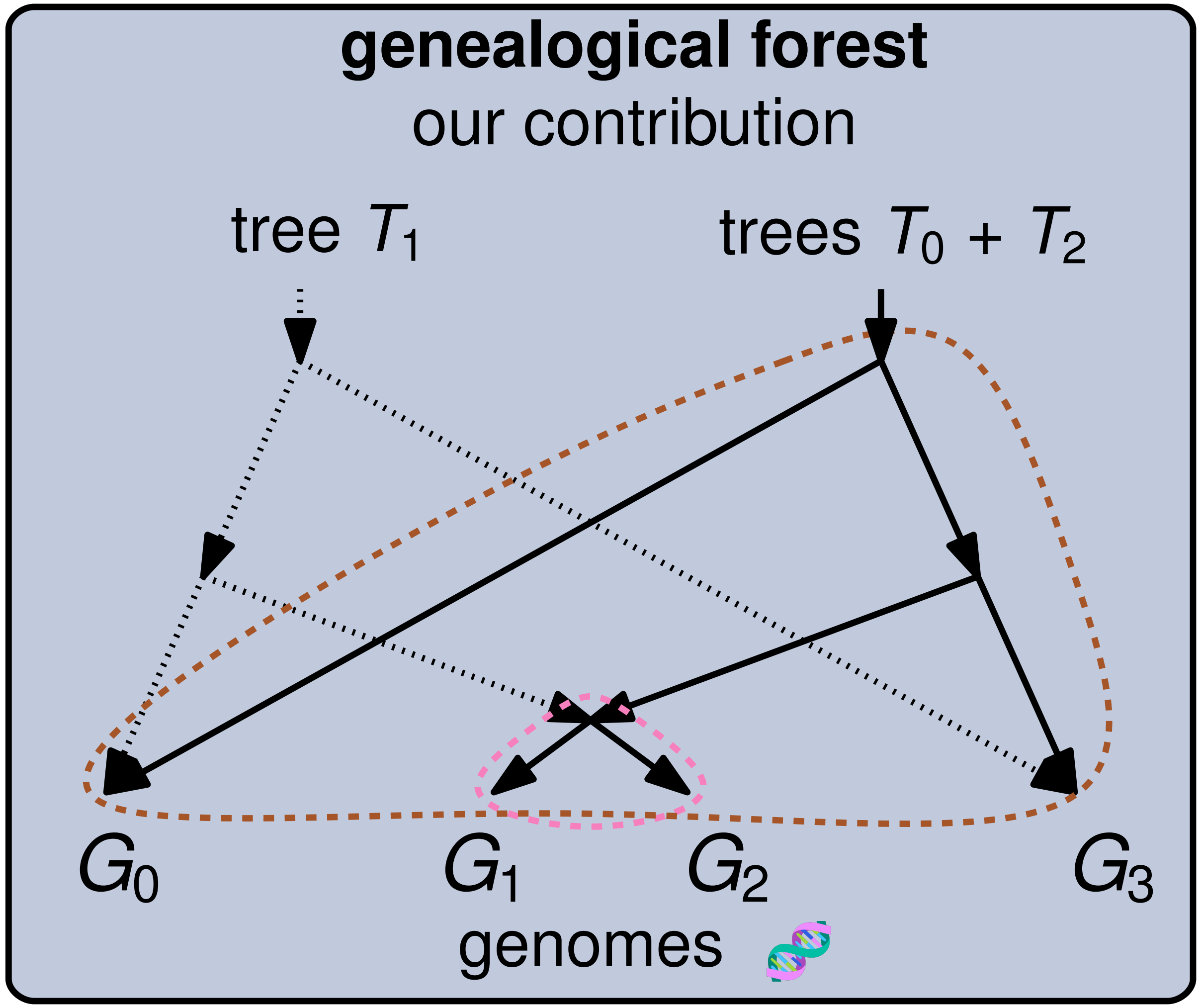
tree sequence

state-of-the-art

tree T_1 tree T_2 tree T_0

G_0 G_1 G_2 G_3

genomes 🧬



genealogical forest

our contribution

tree T_1 trees T_0 + T_2

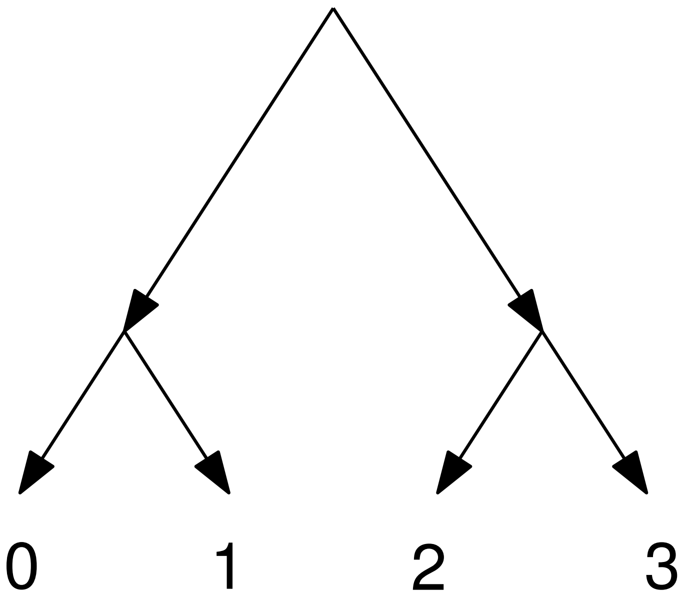
G_0 G_1 G_2 G_3

genomes 🧬

- Advantage: Straigt-forward memoization of intermediate results
- We don’t loose the order of trees

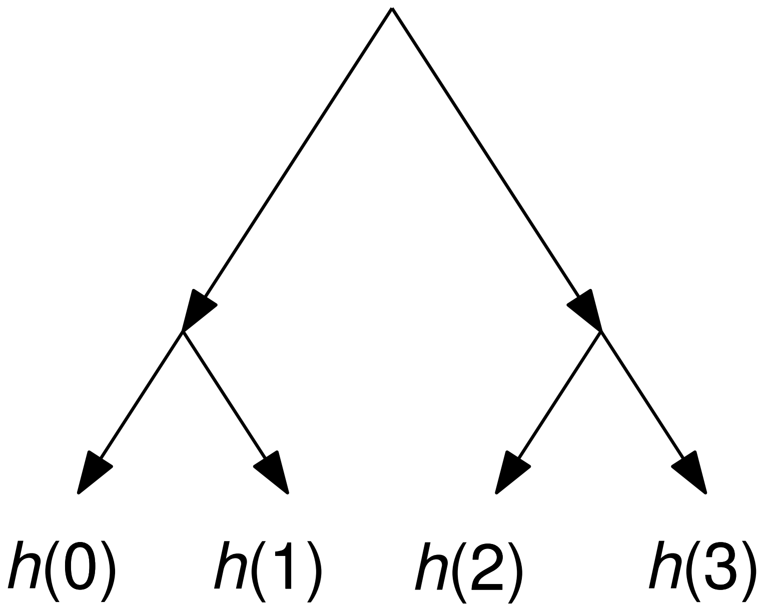
Constructing Genealogical Forests

- For each subtree in each tree in the input tree sequence
- Assign unique IDs to subtrees and represent them as nodes in a DAG



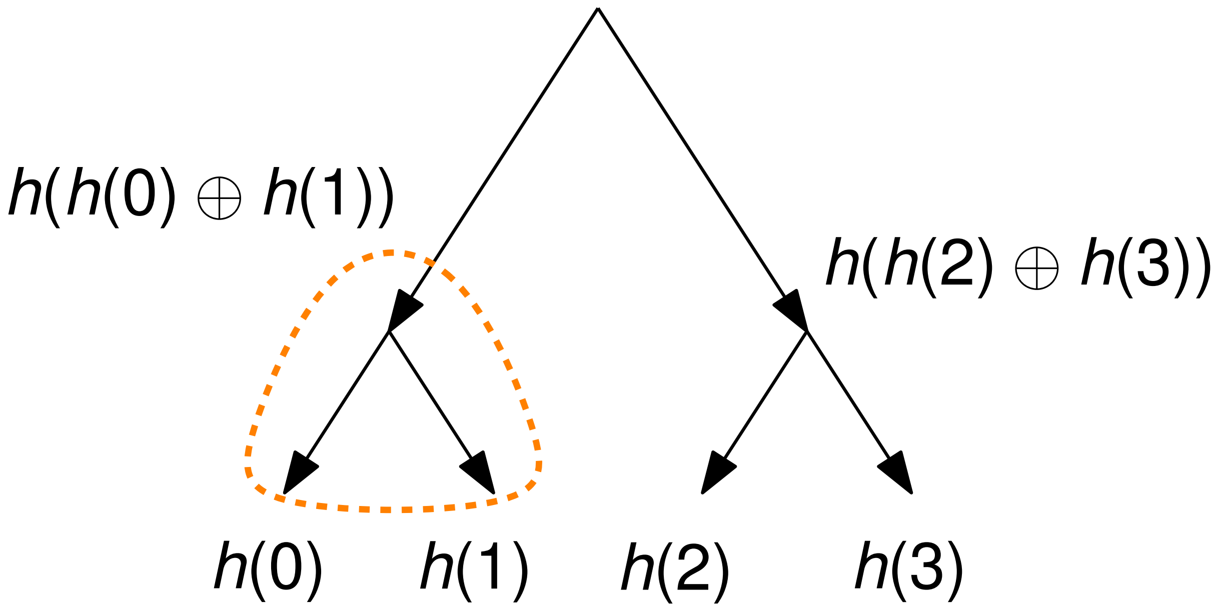
Constructing Genealogical Forests

- For each subtree in each tree in the input tree sequence
- Assign unique IDs to subtrees and represent them as nodes in a DAG



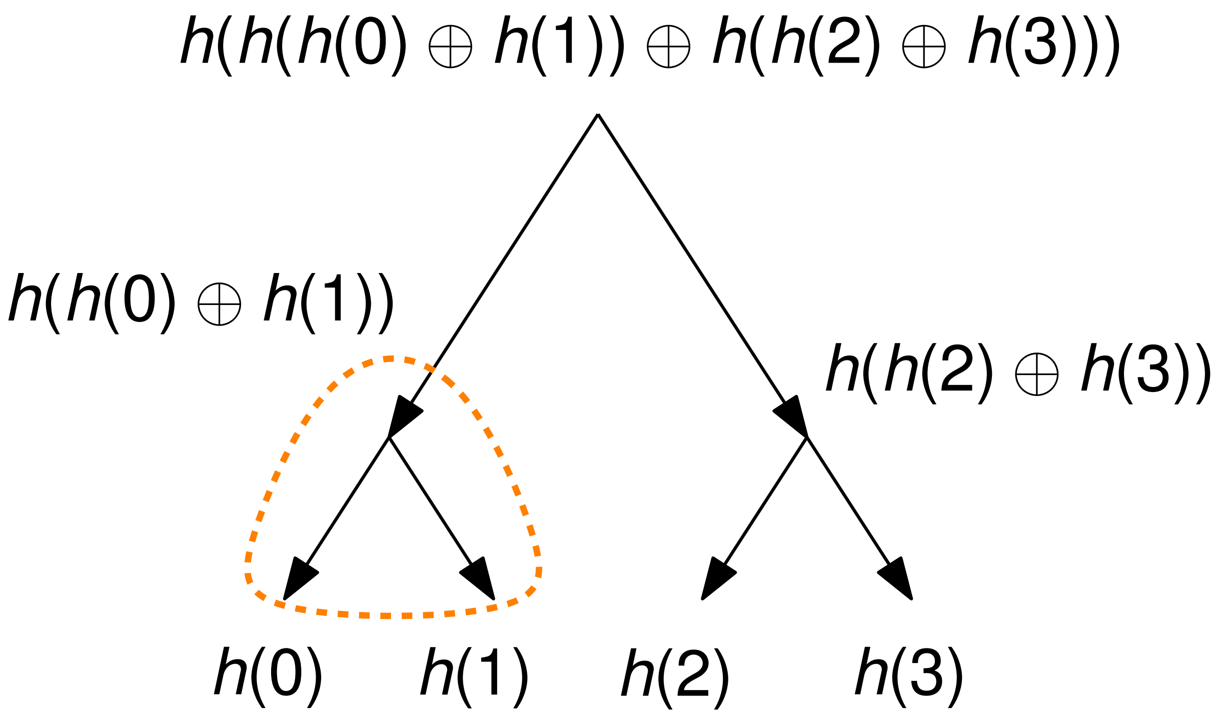
Constructing Genealogical Forests

- For each subtree in each tree in the input tree sequence
- Assign unique IDs to subtrees and represent them as nodes in a DAG



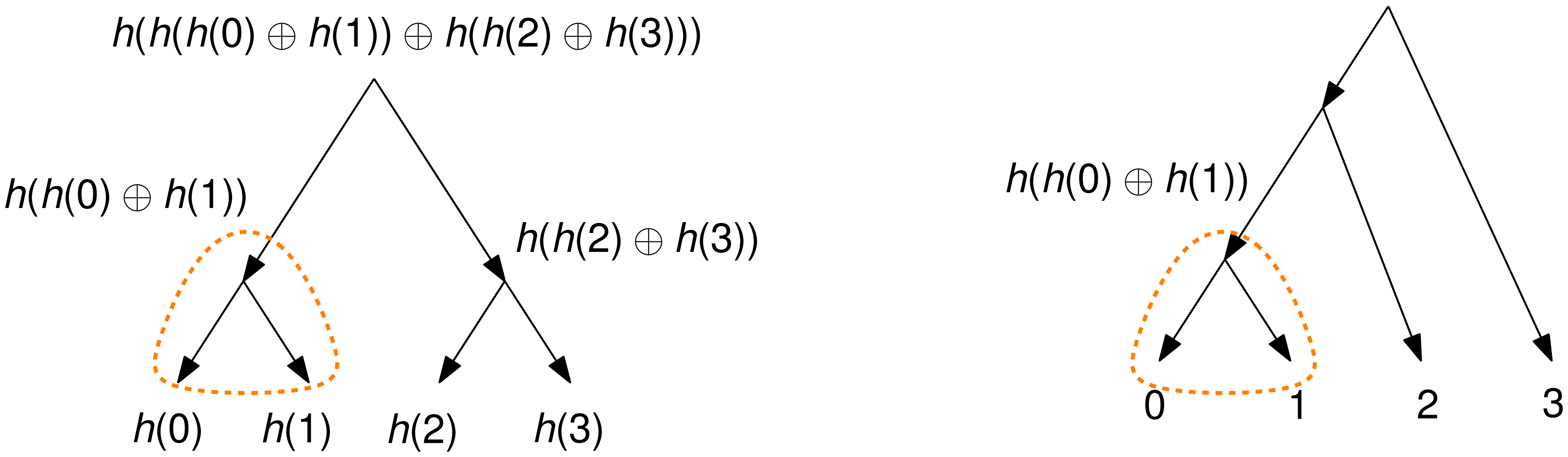
Constructing Genealogical Forests

- For each subtree in each tree in the input tree sequence
- Assign unique IDs to subtrees and represent them as nodes in a DAG



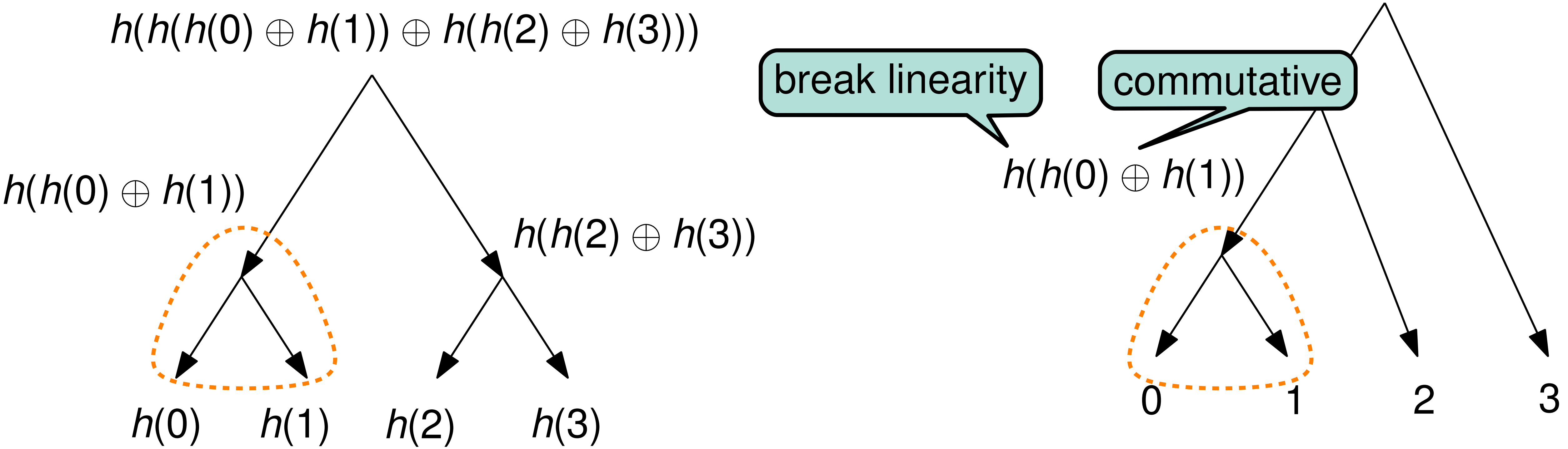
Constructing Genealogical Forests

- For each subtree in each tree in the input tree sequence
- Assign unique IDs to subtrees and represent them as nodes in a DAG



Constructing Genealogical Forests

- For each subtree in each tree in the input tree sequence
- Assign unique IDs to subtrees and represent them as nodes in a DAG



Statistics in Population Genetics

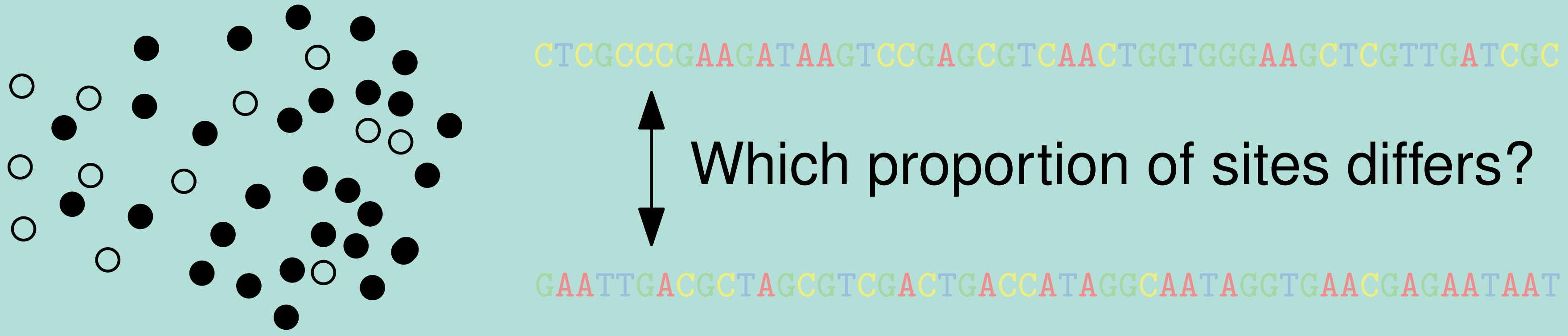
- We consider the genetic states of each genome at each site
- Many common statistics based on these Allele Frequencies
  - Diversity and Divergence
  - Patterson's f_2, f_3, and f_4
  - Fixation Index F_{ST}
  - Tajima's D

Statistics in Population Genetics

- We consider the genetic states of each genome at each site
- Many common statistics based on these Allele Frequencies
  - Diversity and Divergence
  - Patterson's f_2, f_3, and f_4
  - Fixation Index F_{ST}
  - Tajima's D

Diversity

average sequence difference


CTCGCCCGAAGATAAGTCCGAGCGTCAACTGGTGGGAAGCTCGTTGATCGC
GAATTGACGCTAGCGTCGACTGACCATAGGCAATAGGTGAACGAGAATAAT

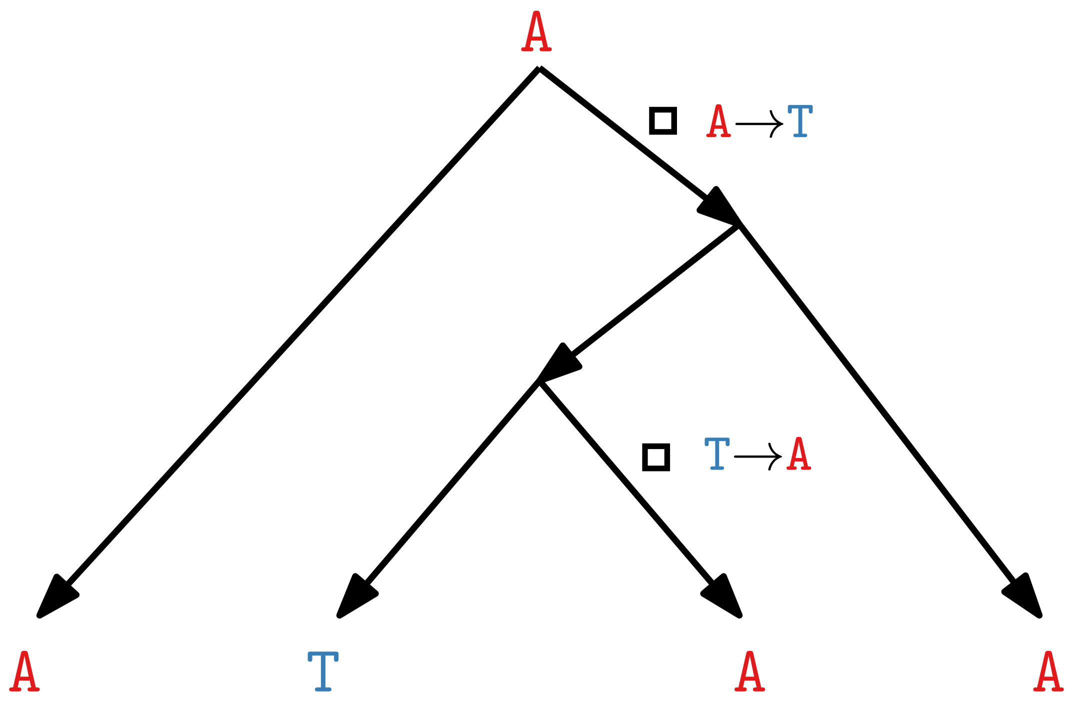
Which proportion of sites differs?

selection of genomes

Storing the Sequence

- These statistics are based on the sequences
- However, storing all sequences base-by-base is not feasible
- Instead, for each site, store the ancestral state and the mutations

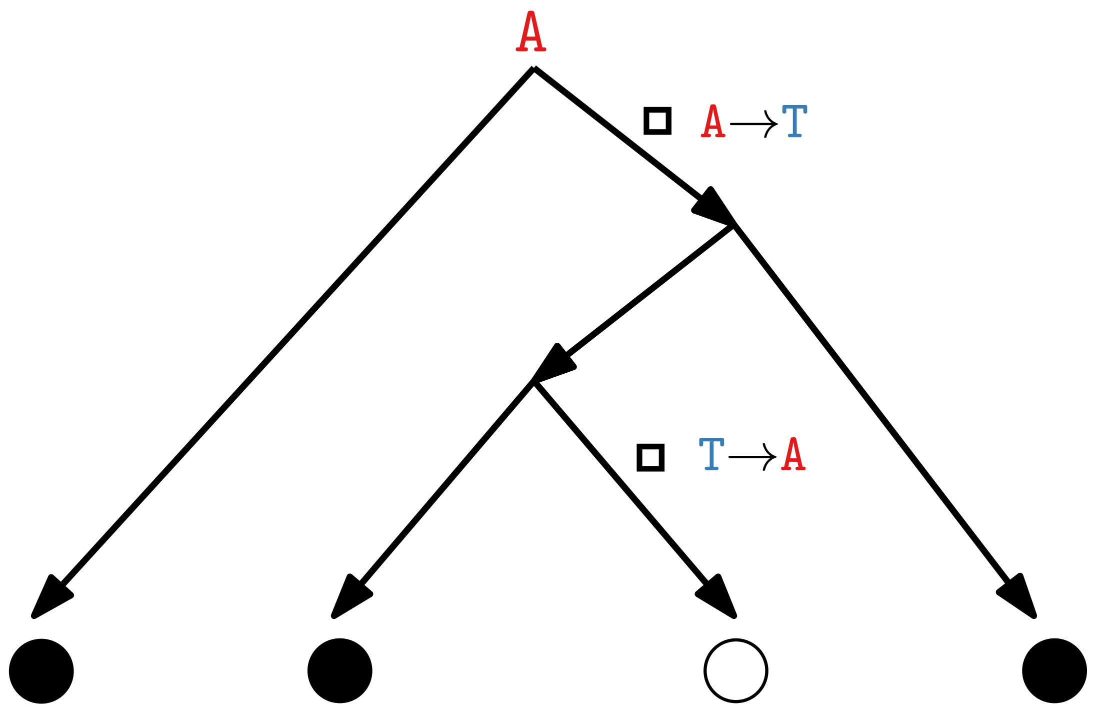
```
TTCGCGCGAAGATAAGTCCGACCGTTAACTGGTGGGAAGCTTGT
CTCGCCCGAAGATAAGTCCGATCGTCAACTGGTAGGGAGCTCGT
CTCGCCCGAAGATAAGTCCGAGCGTACACTGGTGGGAAGCACGT
CTCGCGCGATGTTAAGTCCCACCGTCAACTGGTGGGAAGCTCGT
                    ...
```



Computing Population Genetics Statistics

Diversity: Average sequence difference between two samples

Input: Selection of genomes ●○○●○○● & Tree sequence with mutations □ T→A

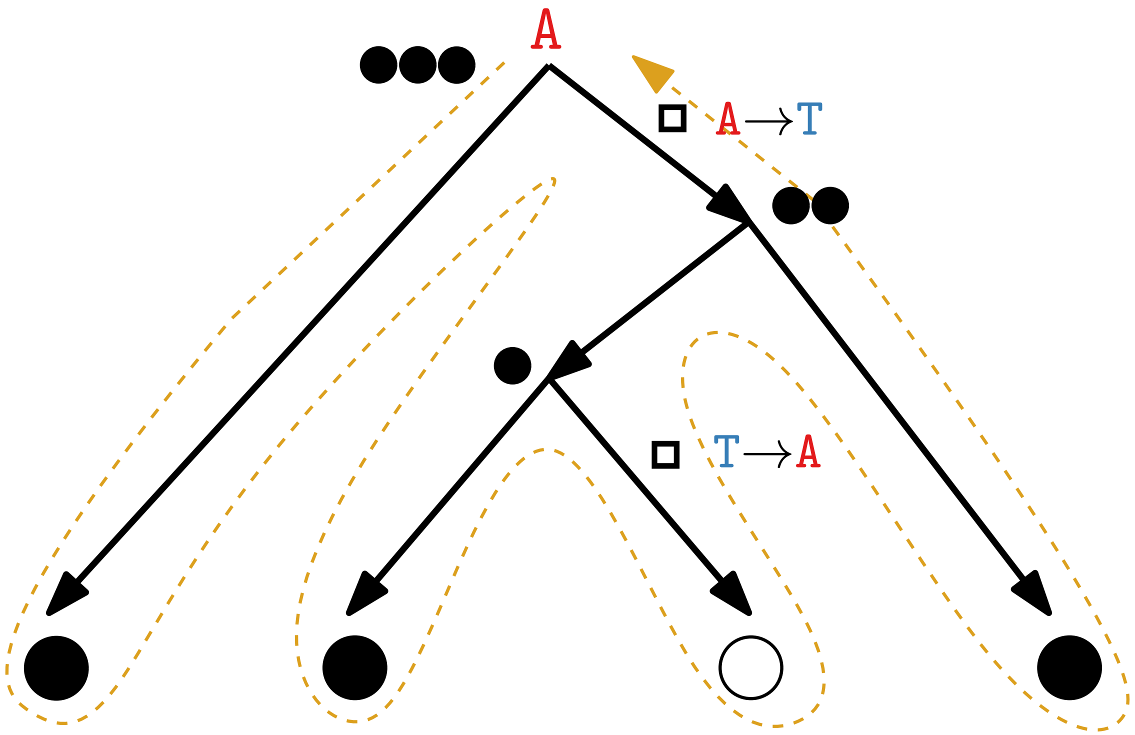


Computing Population Genetics Statistics

Diversity: Average sequence difference between two samples
Input: Selection of genomes ●○○●○● & Tree sequence with mutations □ T→A

(1) Compute number of samples in subtree

- post-order traversal



Computing Population Genetics Statistics

Diversity: Average sequence difference between two samples
Input: Selection of genomes ●○○●○○● & Tree sequence with mutations □ T→A

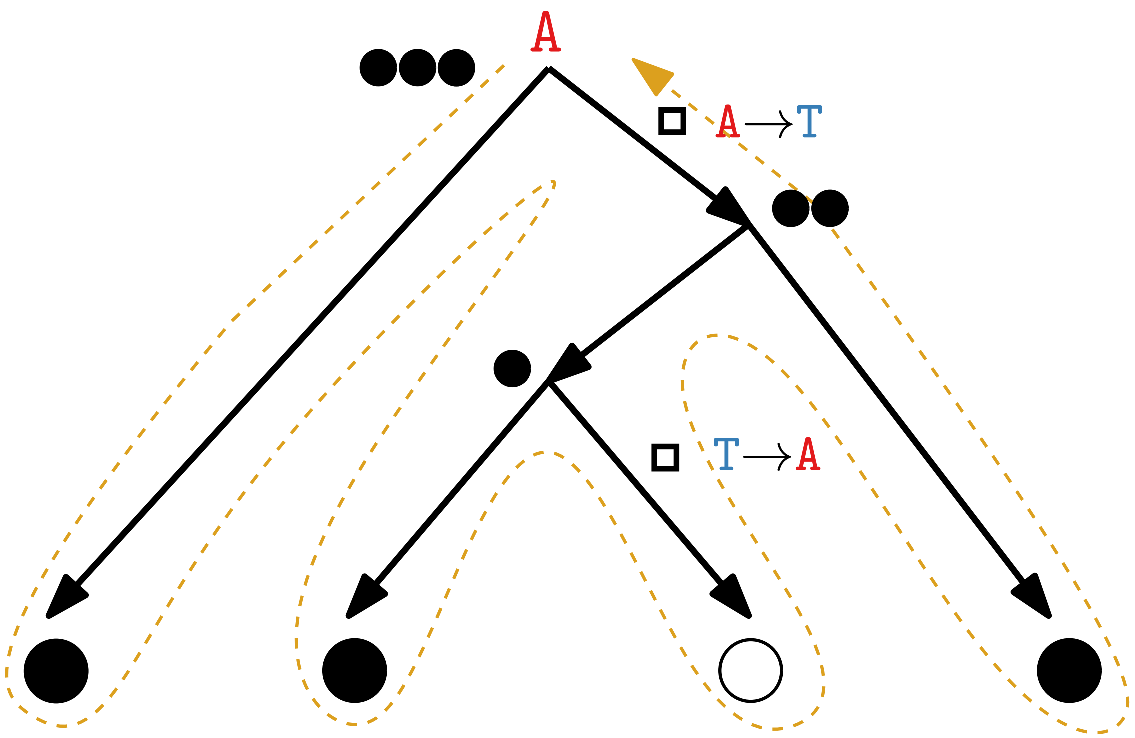
(1) Compute number of samples in subtree

- post-order traversal

(2) Compute allele frequencies

A: 3 C: 0 T: 0 G: 0

- Ancestral State: A
- Mutation A → T at
- Mutation T → A at



Computing Population Genetics Statistics

Diversity: Average sequence difference between two samples

Input: Selection of genomes ●○●○○● & Tree sequence with mutations □ T→A

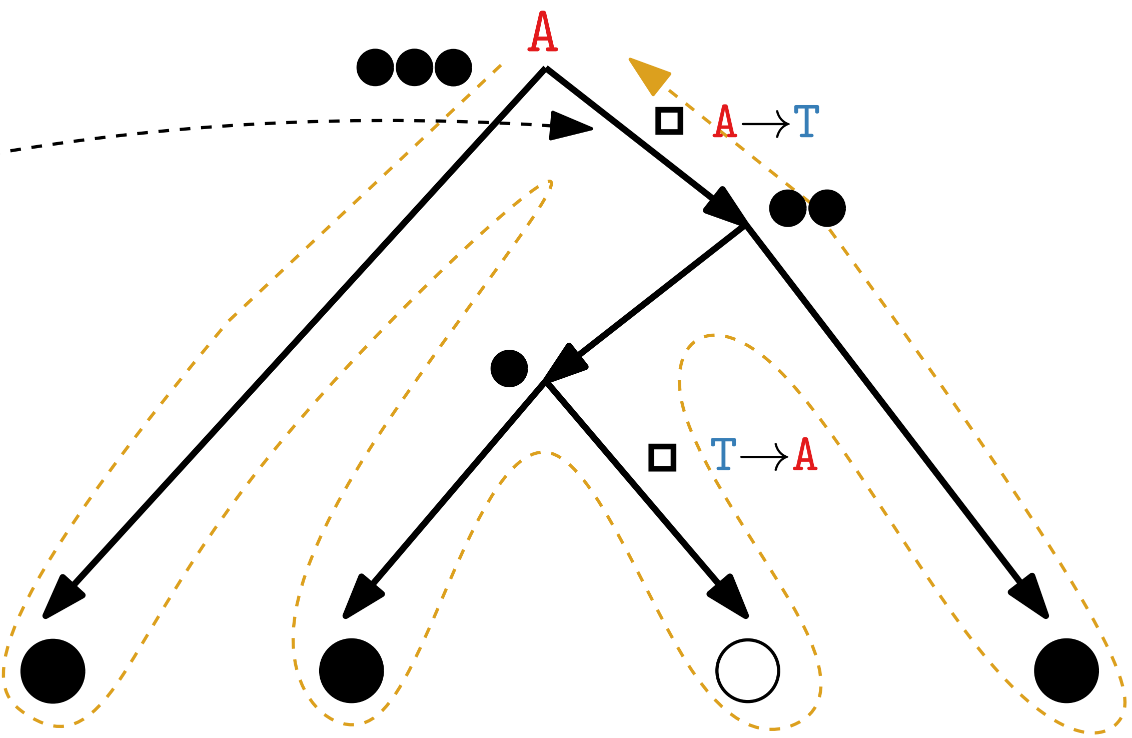
(1) Compute number of samples in subtree

- post-order traversal

(2) Compute allele frequencies

A: 1 C: 0 T: 2 G: 0

- Ancestral State: A
- Mutation A → T at ●
- Mutation T → A at



Computing Population Genetics Statistics

Diversity: Average sequence difference between two samples

Input: Selection of genomes ●○●○○○● & Tree sequence with mutations □ T→A

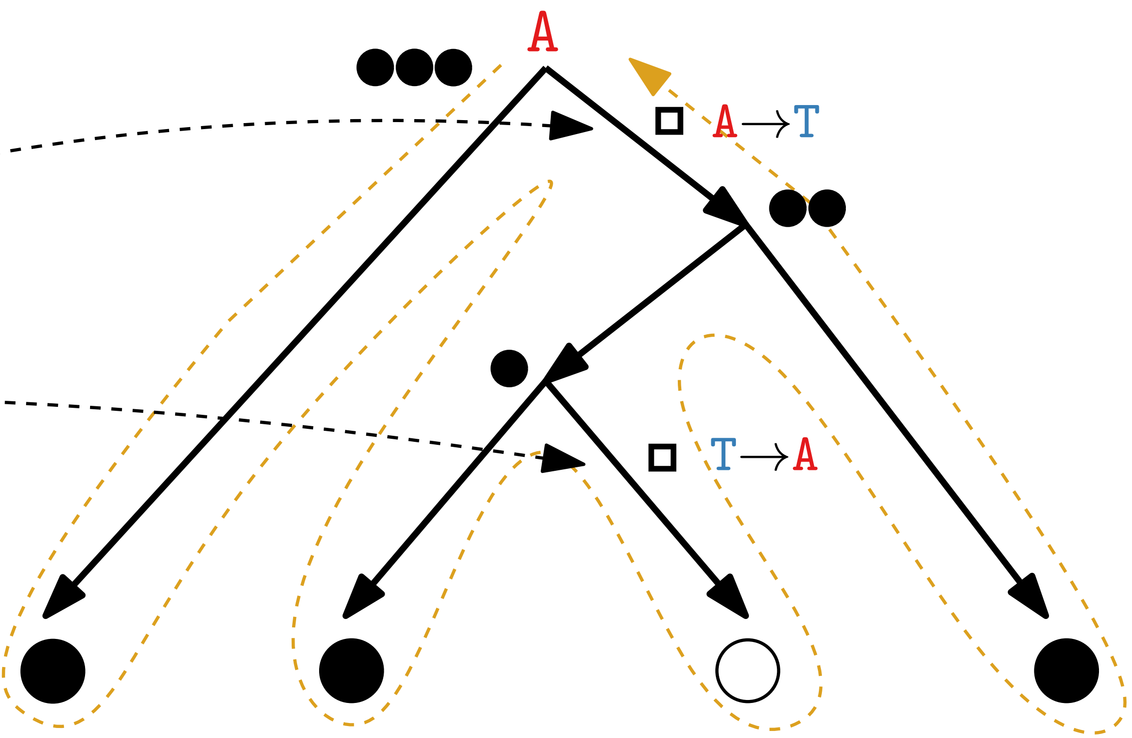
(1) Compute number of samples in subtree

- post-order traversal

(2) Compute allele frequencies

A: 1 C: 0 T: 2 G: 0

- Ancestral State: A
- Mutation A → T at ●
- Mutation T → A at ●



Computing Population Genetics Statistics

Diversity: Average sequence difference between two samples

Input: Selection of genomes ●○●○○● & Tree sequence with mutations □ T→A

(1) Compute number of samples in subtree

- post-order traversal

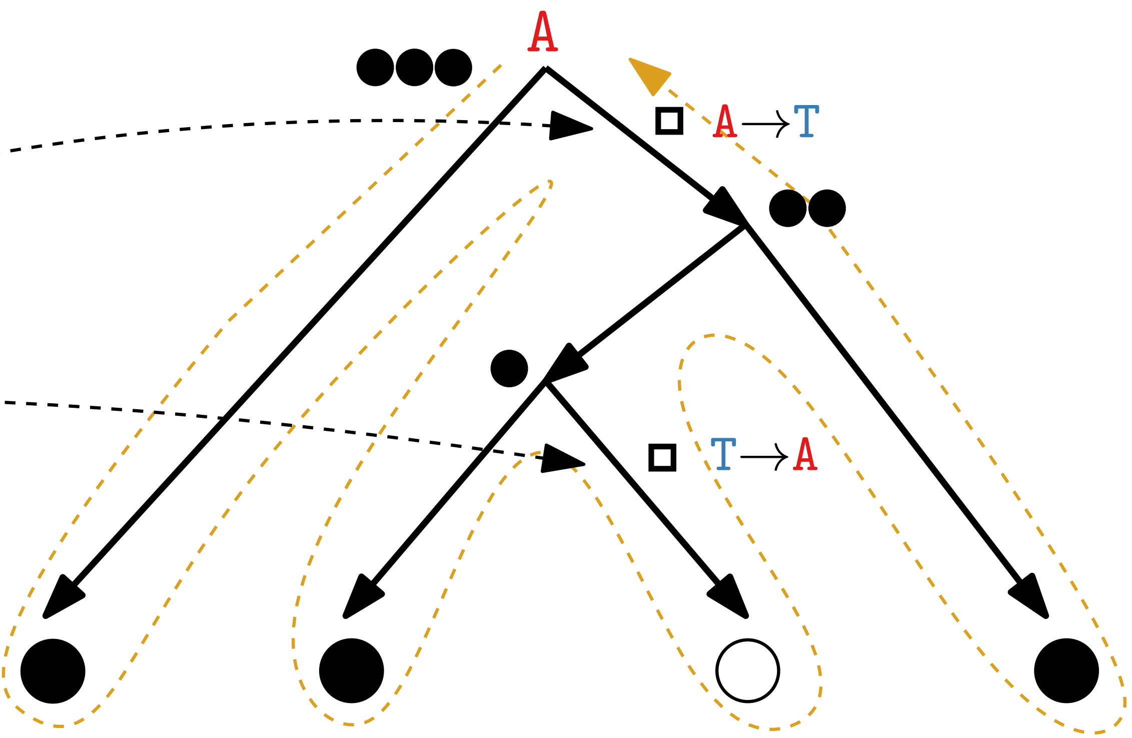
(2) Compute allele frequencies

A: 1 C: 0 T: 2 G: 0

- Ancestral State: A
- Mutation A → T at ●
- Mutation T → A at ●

(3) Compute average difference

- Diversity = $\frac{\text{freq}_A \cdot (1 - \text{freq}_A) + \ldots}{n \cdot (n-1)}$



Computing Population Genetics Statistics

Diversity: Average sequence difference between two samples

Input: Selection of genomes ●○●○○● & Tree sequence with mutations □ T→A

(1) Compute number of samples in subtree

- post-order traversal

90 % of runtime

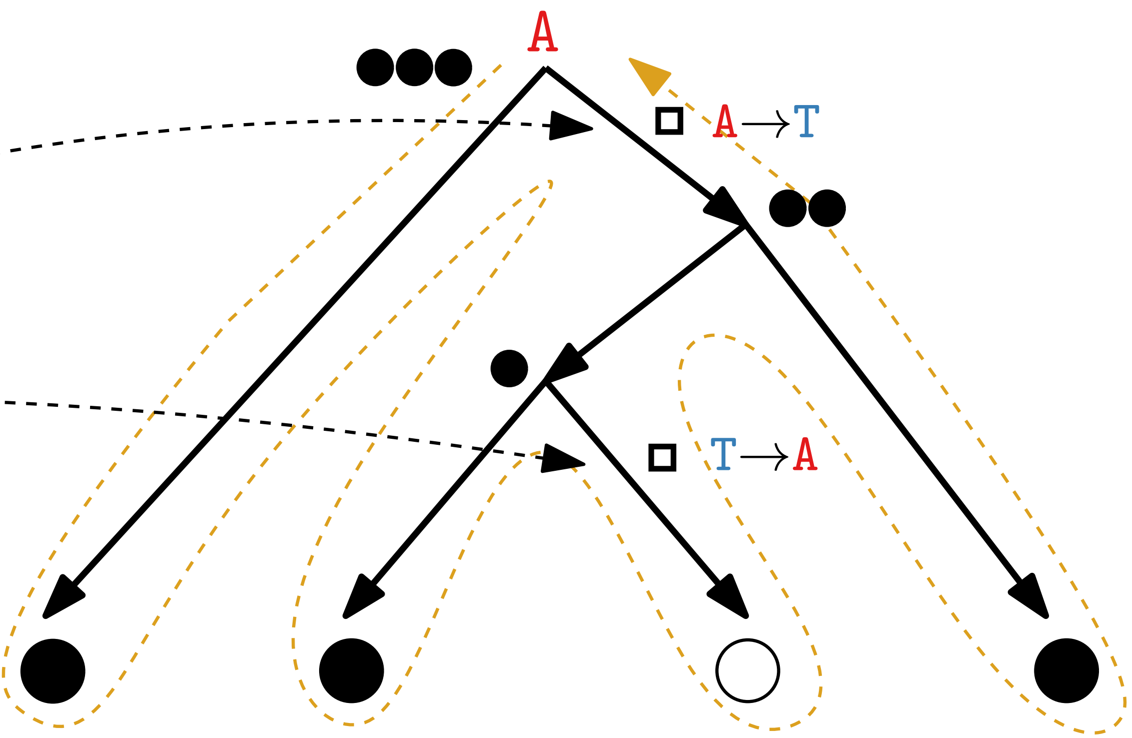
(2) Compute allele frequencies

A: 1 C: 0 T: 2 G: 0

- Ancestral State: A
- Mutation A → T at ●
- Mutation T → A at ●

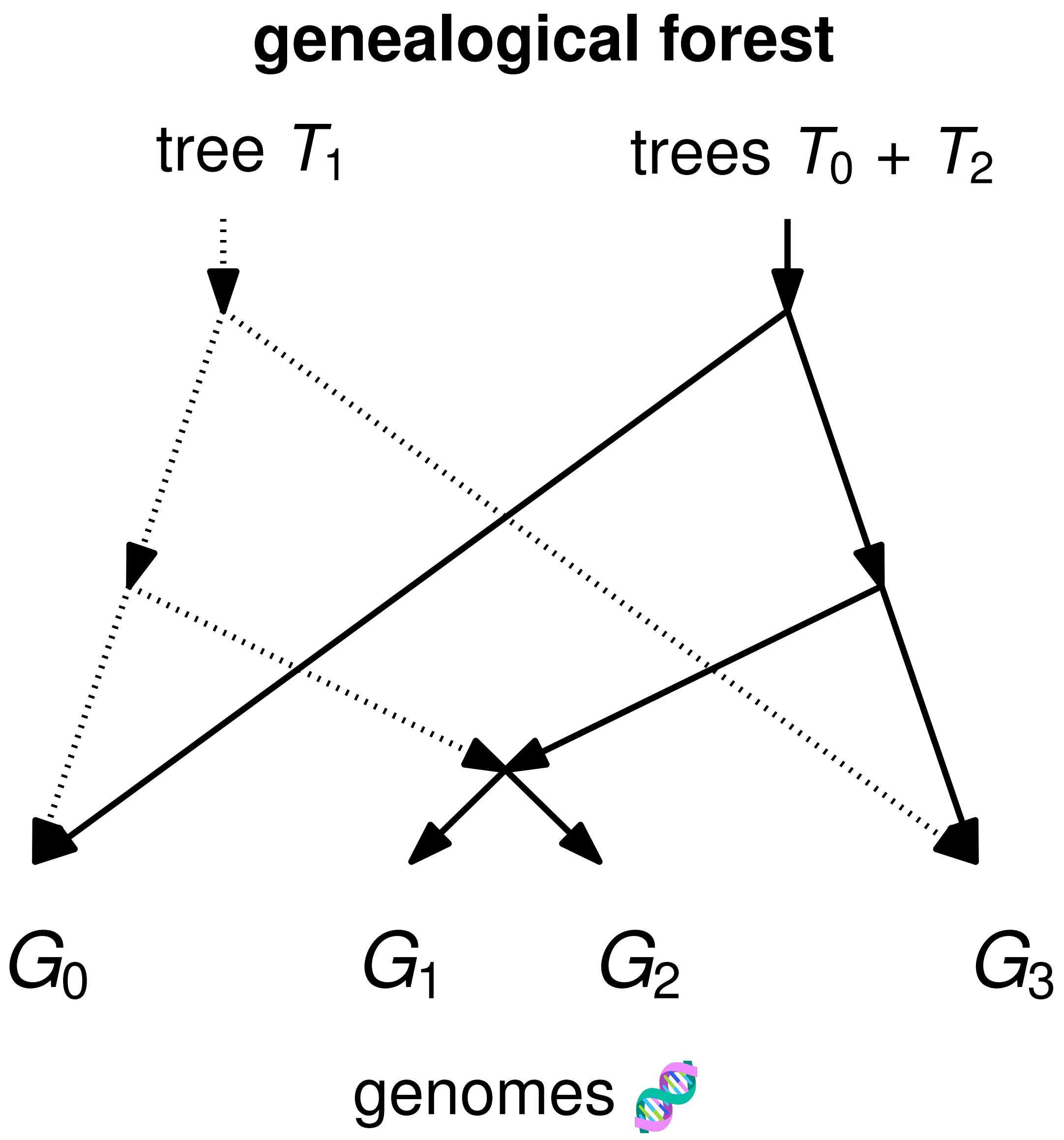
(3) Compute average difference

- $\text{Diversity} = \frac{\text{freq}_A \cdot (1-\text{freq}_A) + \ldots}{n \cdot (n-1)}$



Post-Order Traversal

- Children of a node processed before the node
- Associate a node's result with its DAG node ID
  $\Rightarrow$ Straight-forward memoization

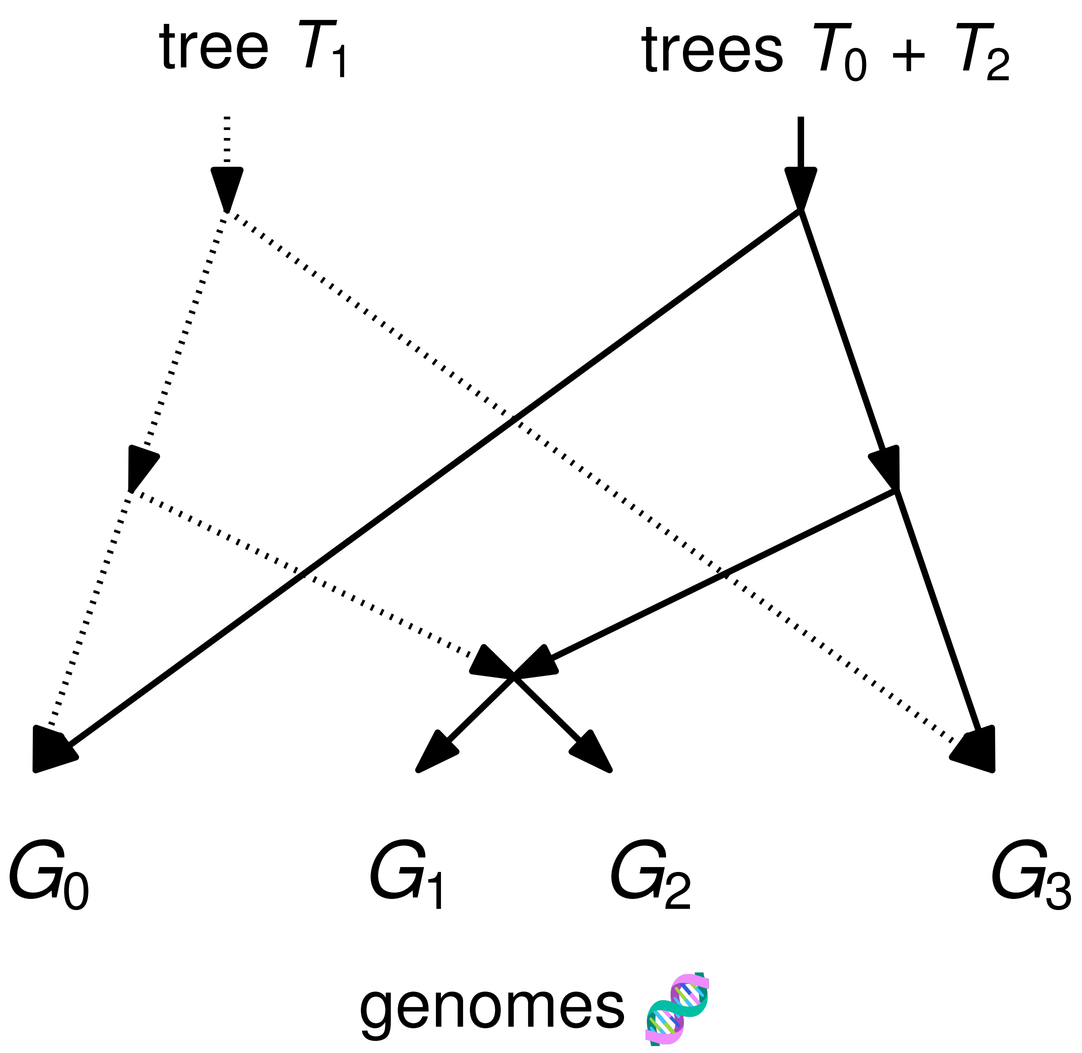


Post-Order Traversal

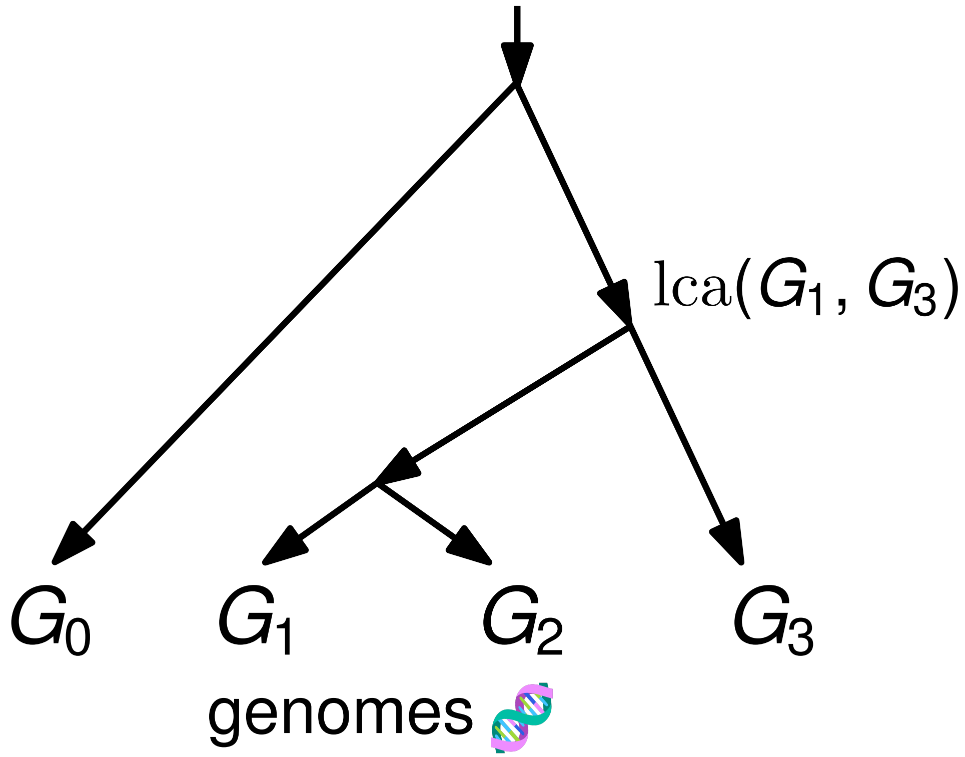
- Children of a node processed before the node
- Associate a node's result with its DAG node ID
  $\Rightarrow$ Straight-forward memoization

Intermediate results reused
5.1 times on average

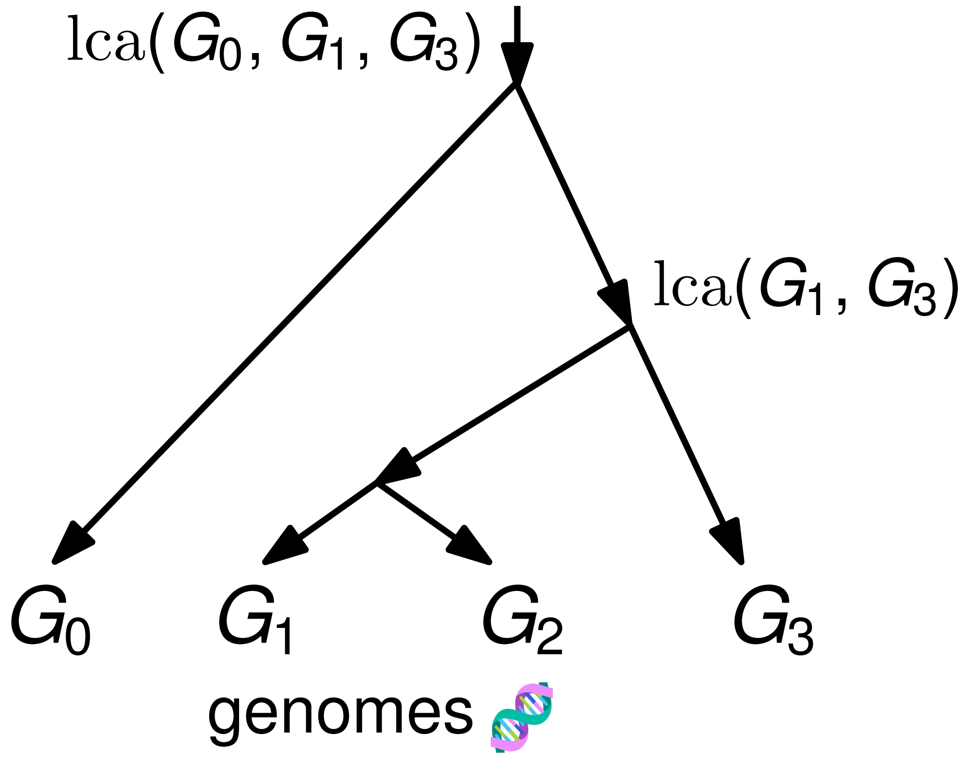
genealogical forest



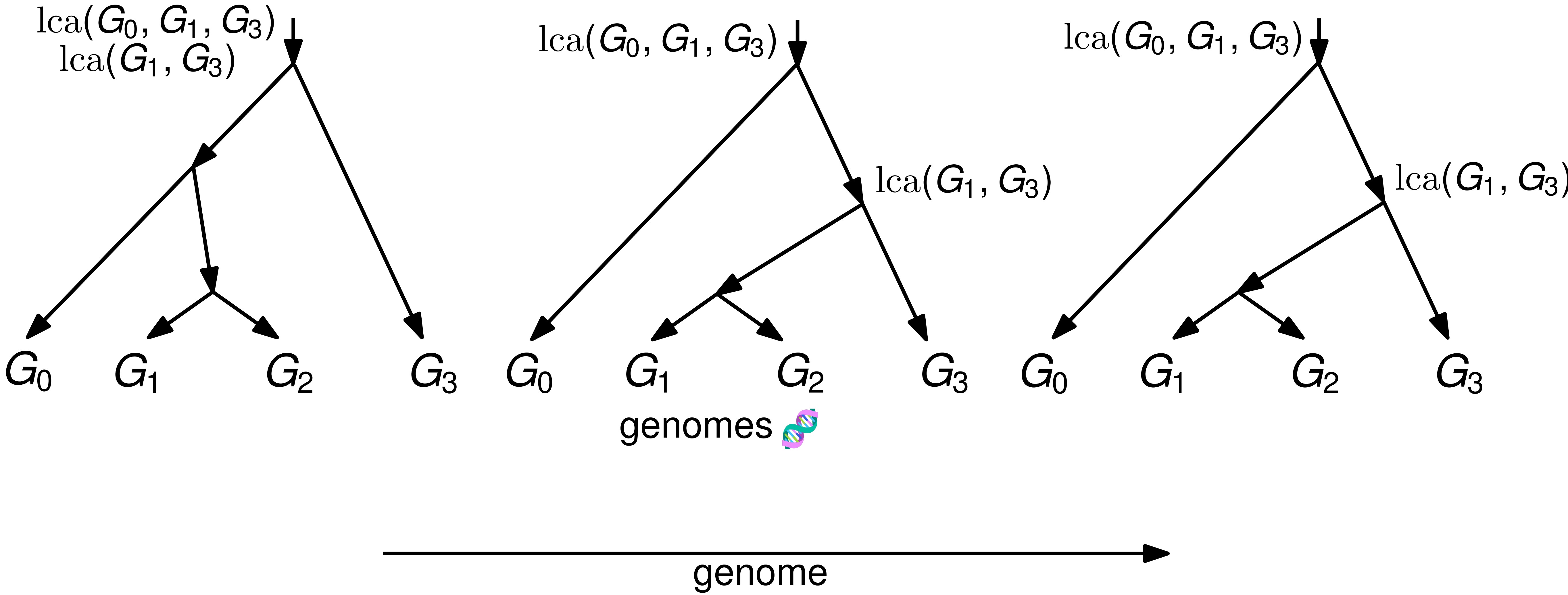
Lowest Common Ancestors



Lowest Common Ancestors



Lowest Common Ancestors



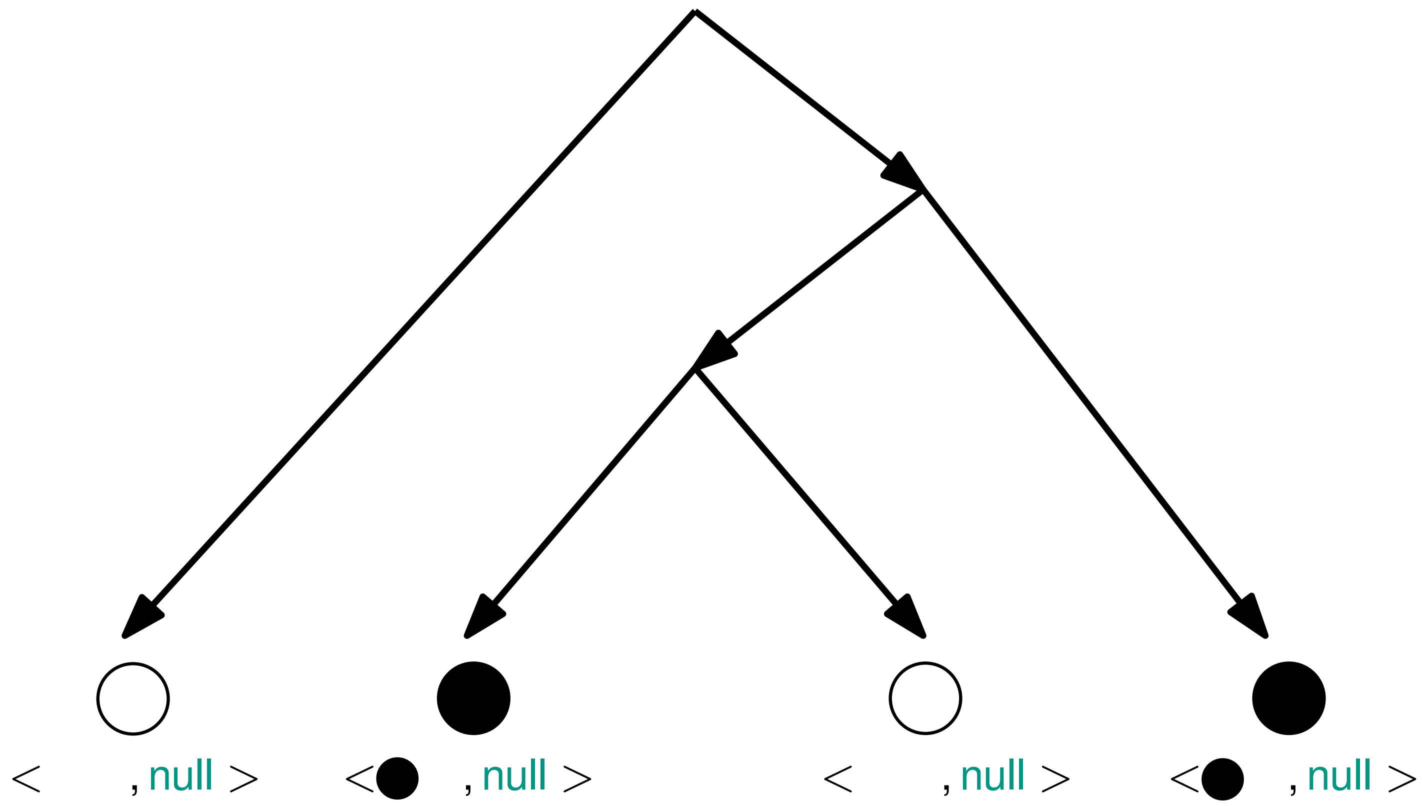
Computing the Lowest Common Ancestor

Lowest Common Ancestor Node farthest from the root where paths to root converge

Input: Selection of samples & Tree sequence

(1) Compute number of samples in subtree

- post-order traversal
- $<$ sample count, LCA $>$

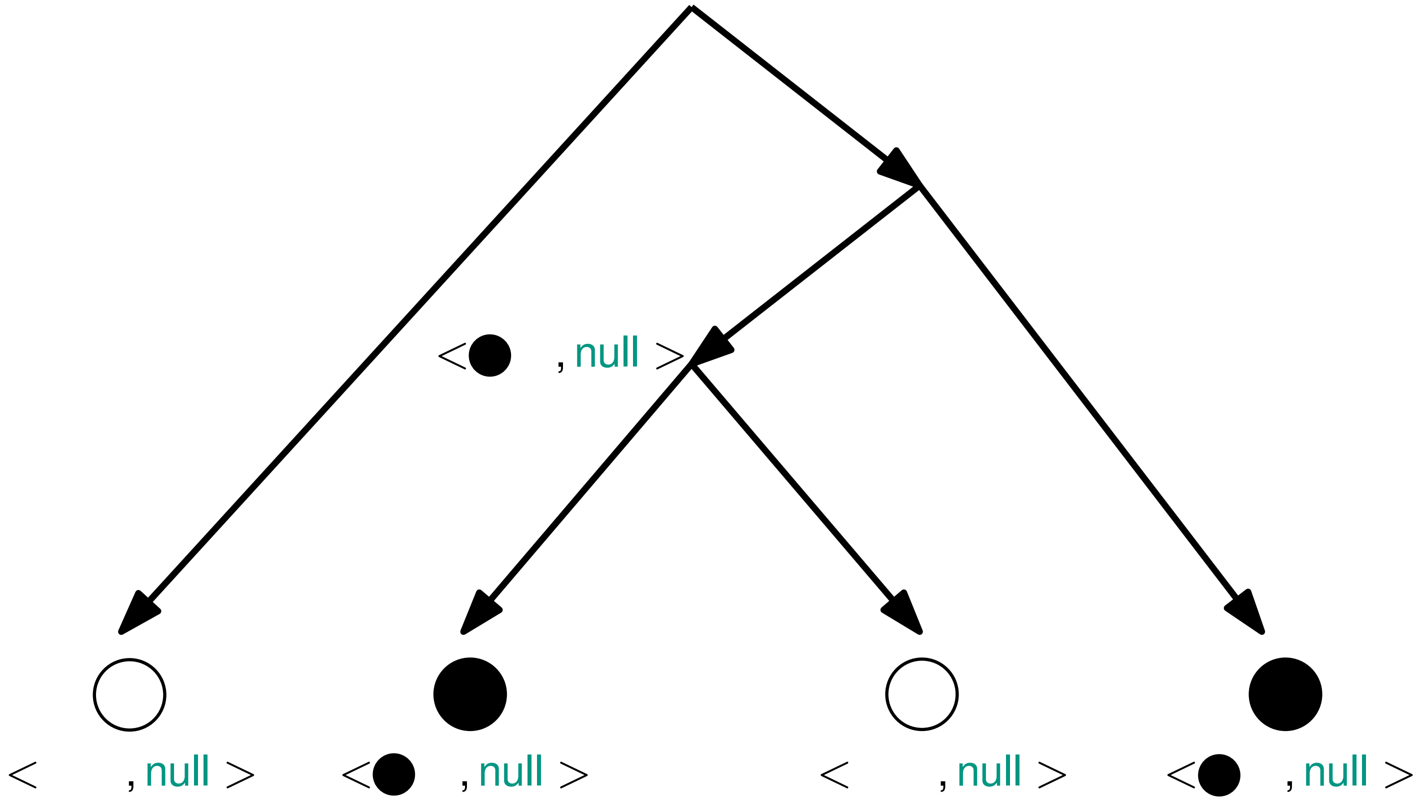


Computing the Lowest Common Ancestor

Lowest Common Ancestor Node farthest from the root where paths to root converge
Input: Selection of samples ●○●○○● & Tree sequence

(1) Compute number of samples in subtree

- post-order traversal
- < sample count, LCA >

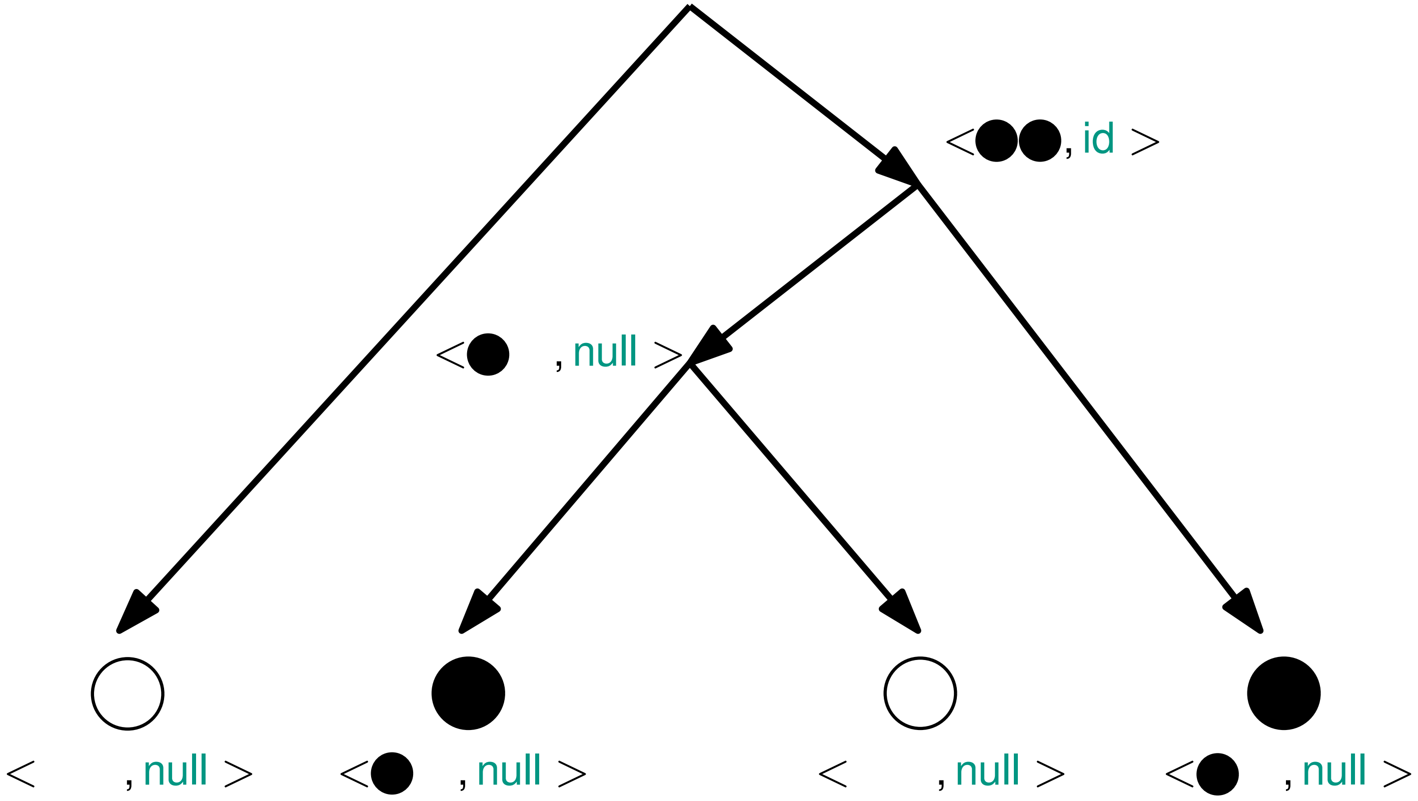


Computing the Lowest Common Ancestor

Lowest Common Ancestor Node farthest from the root where paths to root converge
Input: Selection of samples ●○●○○○● & Tree sequence

(1) Compute number of samples in subtree

- post-order traversal
- $<$ sample count, LCA $>$



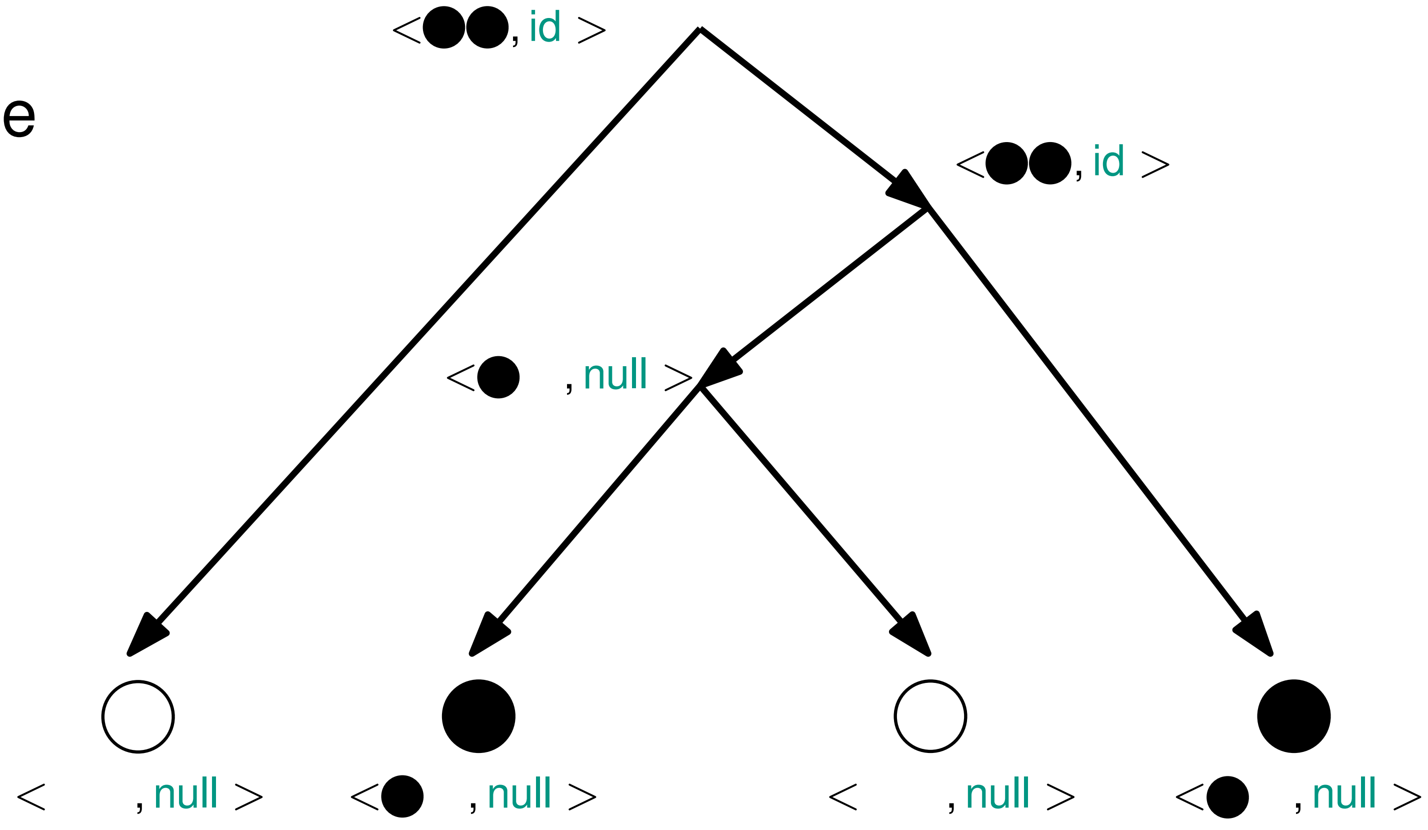
Computing the Lowest Common Ancestor

Lowest Common Ancestor Node farthest from the root where paths to root converge
Input: Selection of samples ●○○●○○● & Tree sequence

(1) Compute number of samples in subtree

- post-order traversal
- $<$ sample count, LCA $>$

(2) Return LCA per tree in tree sequence



Computing the Lowest Common Ancestor

Lowest Common Ancestor Node farthest from the root where paths to root converge

Input: Selection of samples ●○○●○○● & Tree sequence

(1) Compute number of samples in subtree

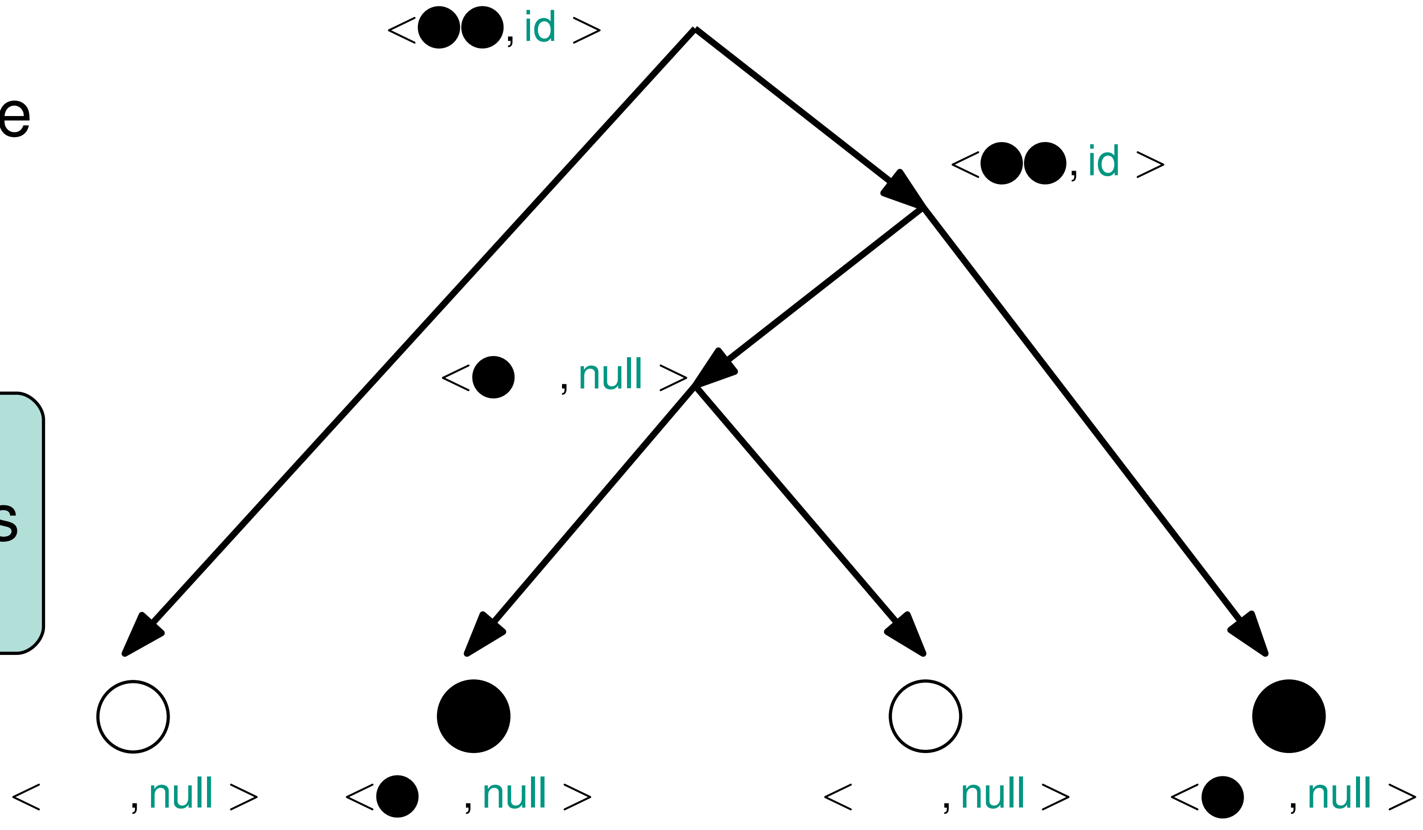
- post-order traversal
- < sample count, LCA >

(2) Return LCA per tree in tree sequence

Runtime

`gfkit:` independent of selected samples

`tskit:` chains pair-wise queries



Evaluation

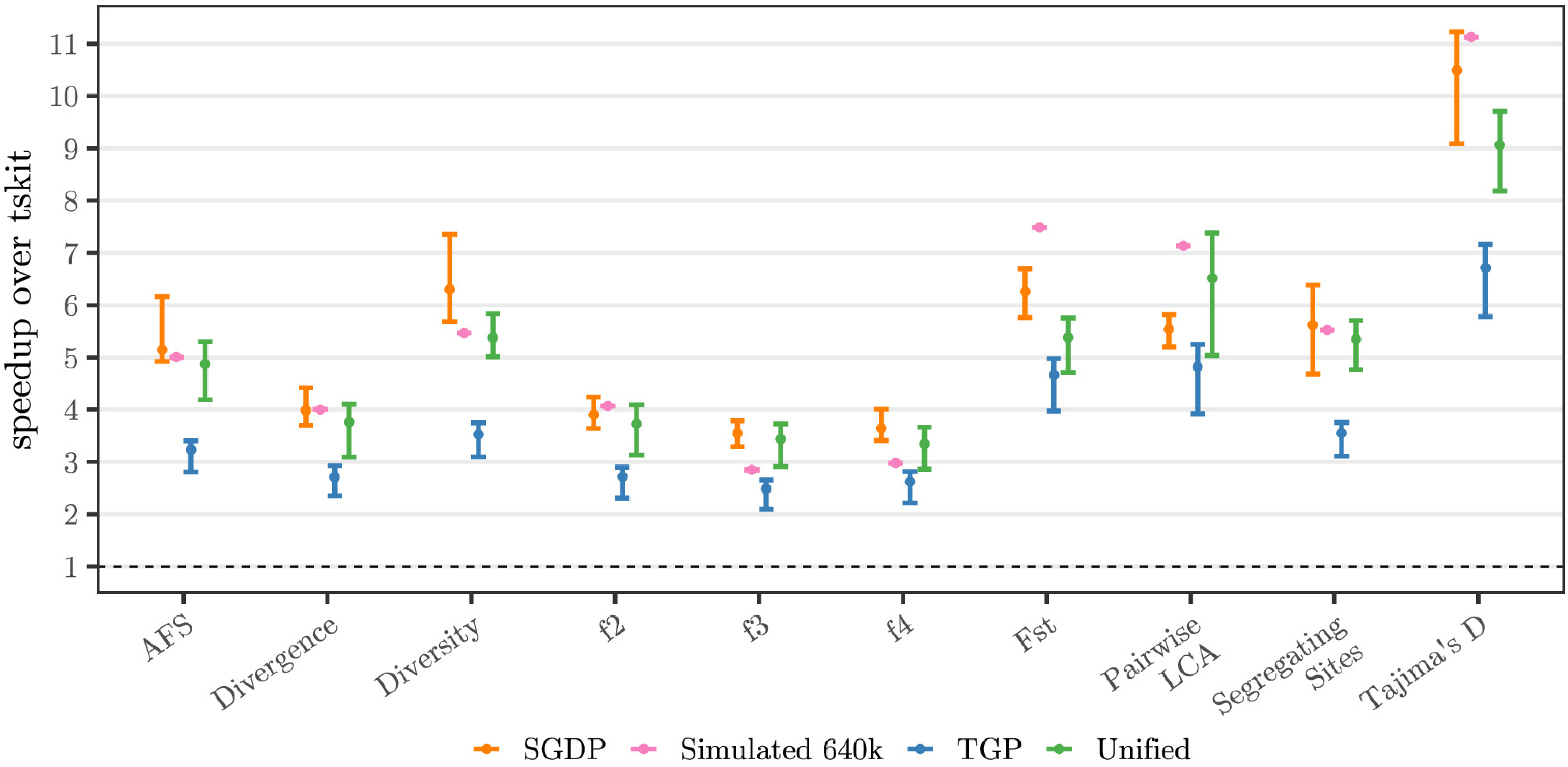
Software and Hardware

- C++20, CMake 3.25.1, gcc 12.1, ld 2.38
- AMD EPYC 7551P CPU
- 8 banks of 32 GiB DDR4 RAM
- All experiments are single-threaded

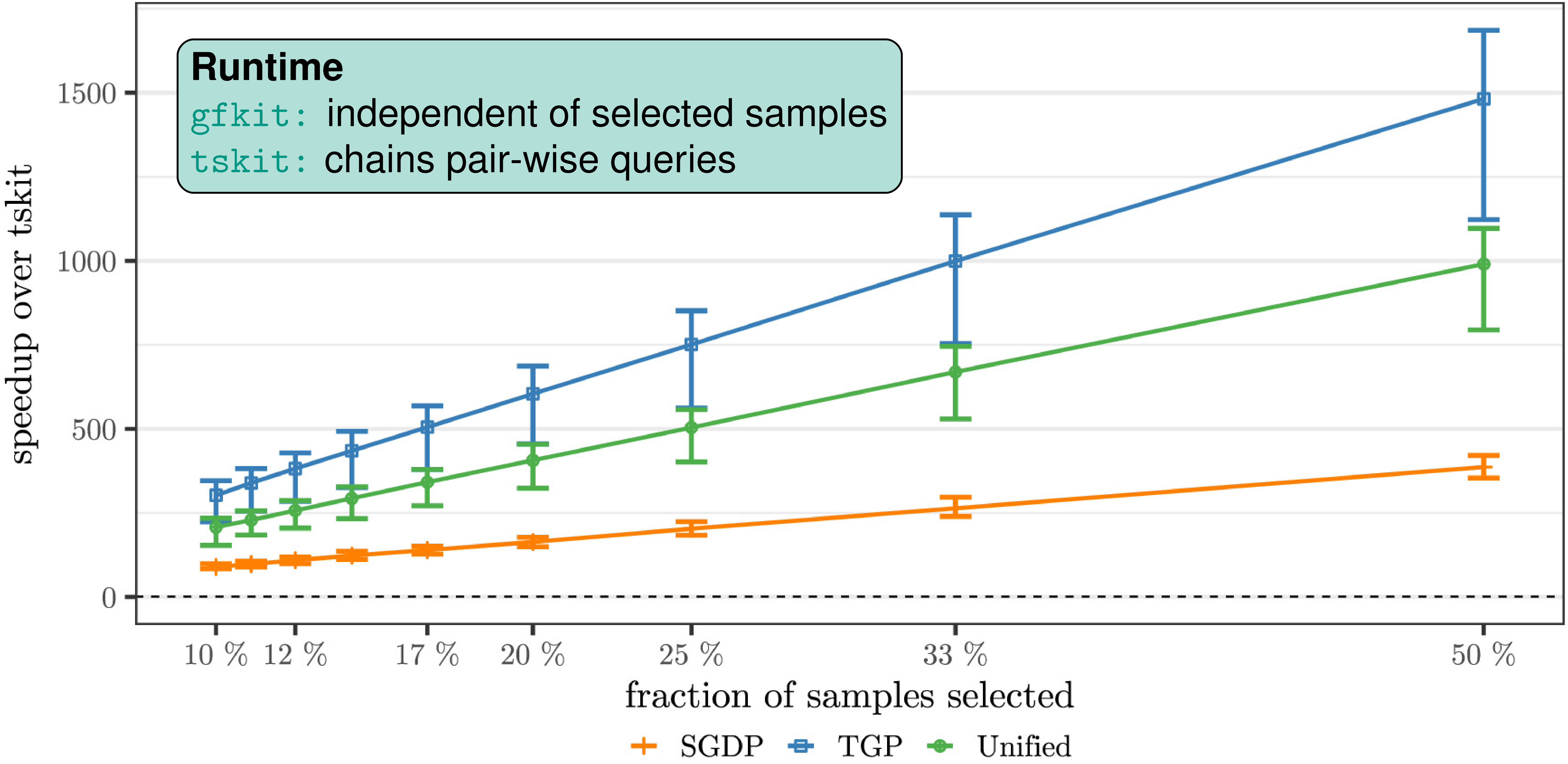
Datasets

- Human (GRCh38)
- Empirical
  - Thousand Genomes Project (Auton et al., 2015)
  - Simons Genome Diversity Project (Mallik et al., 2016)
  - Unified (TGP+SGDP+Ancestral; Wohns et al., 2022)
- Simulated: Chromosome 20, $640\,000$ samples

Evaluation: Computing Statistics



Evaluation: Computing the LCA

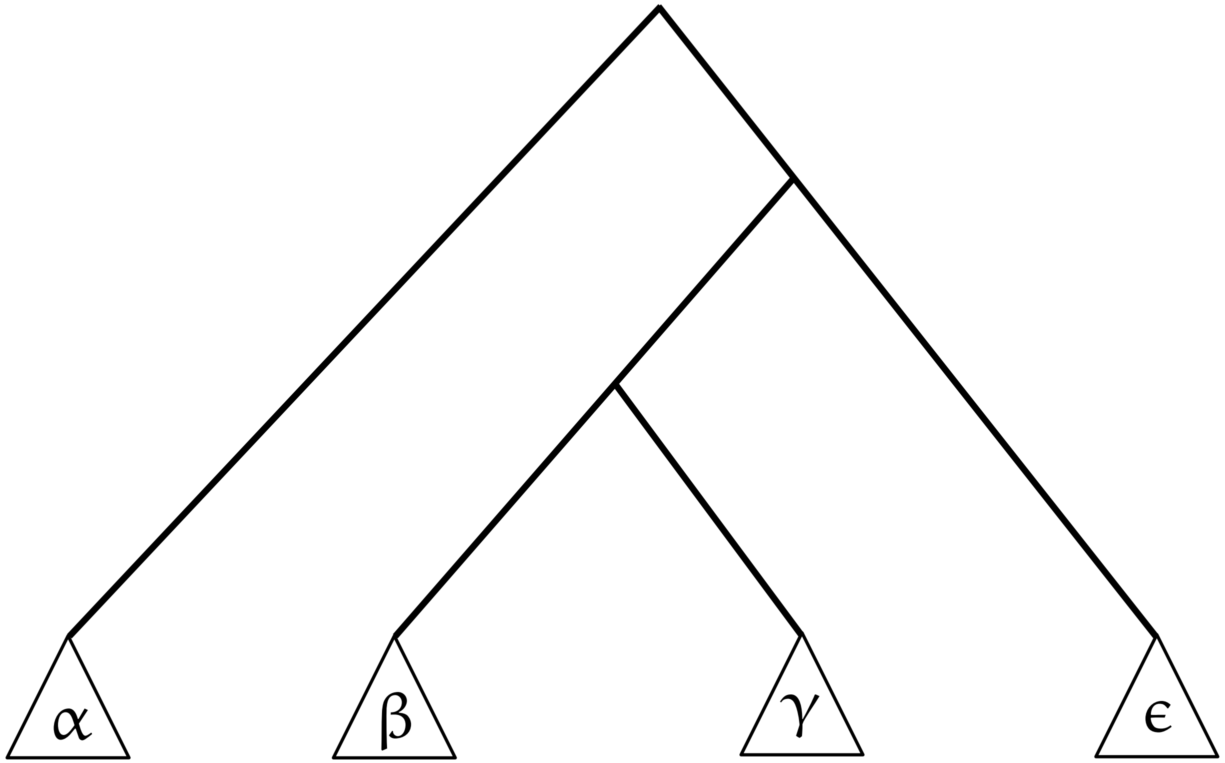


Runtime

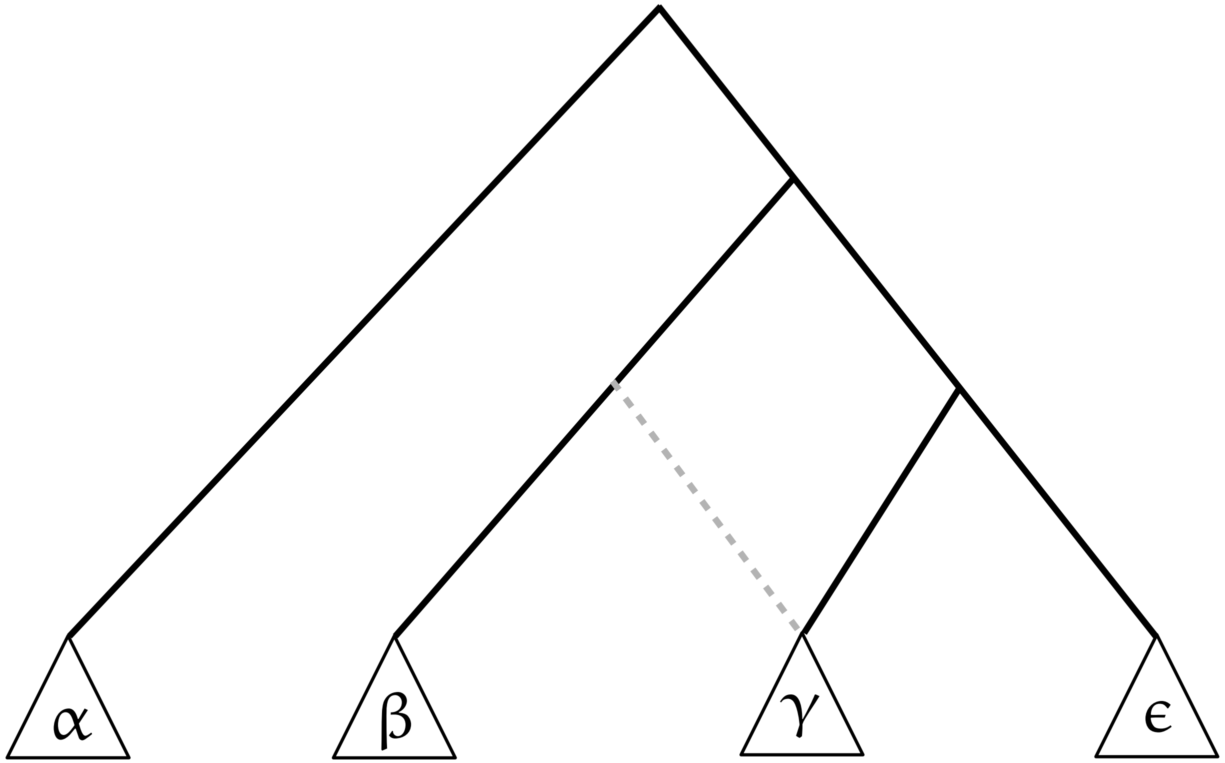
`gfkit:` independent of selected samples

`tskit:` chains pair-wise queries

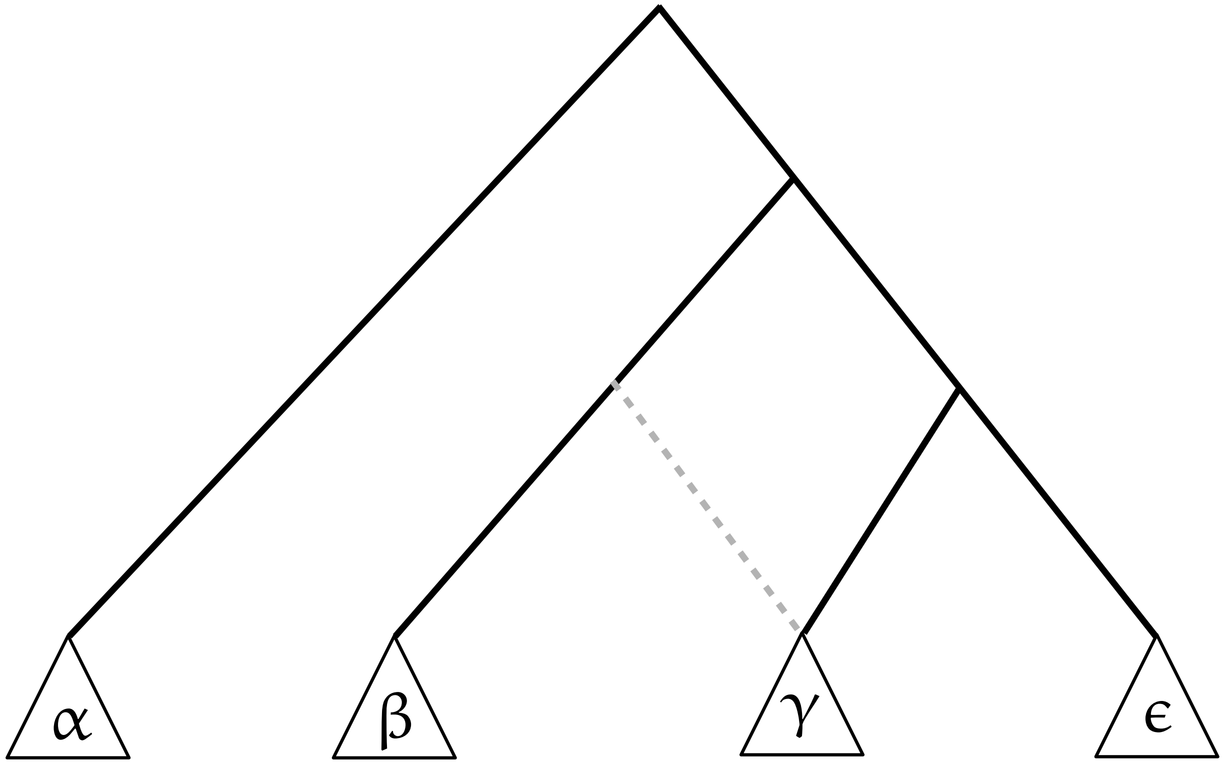
Memory Consumption



Memory Consumption



Memory Consumption



Factoring out all unique subtrees

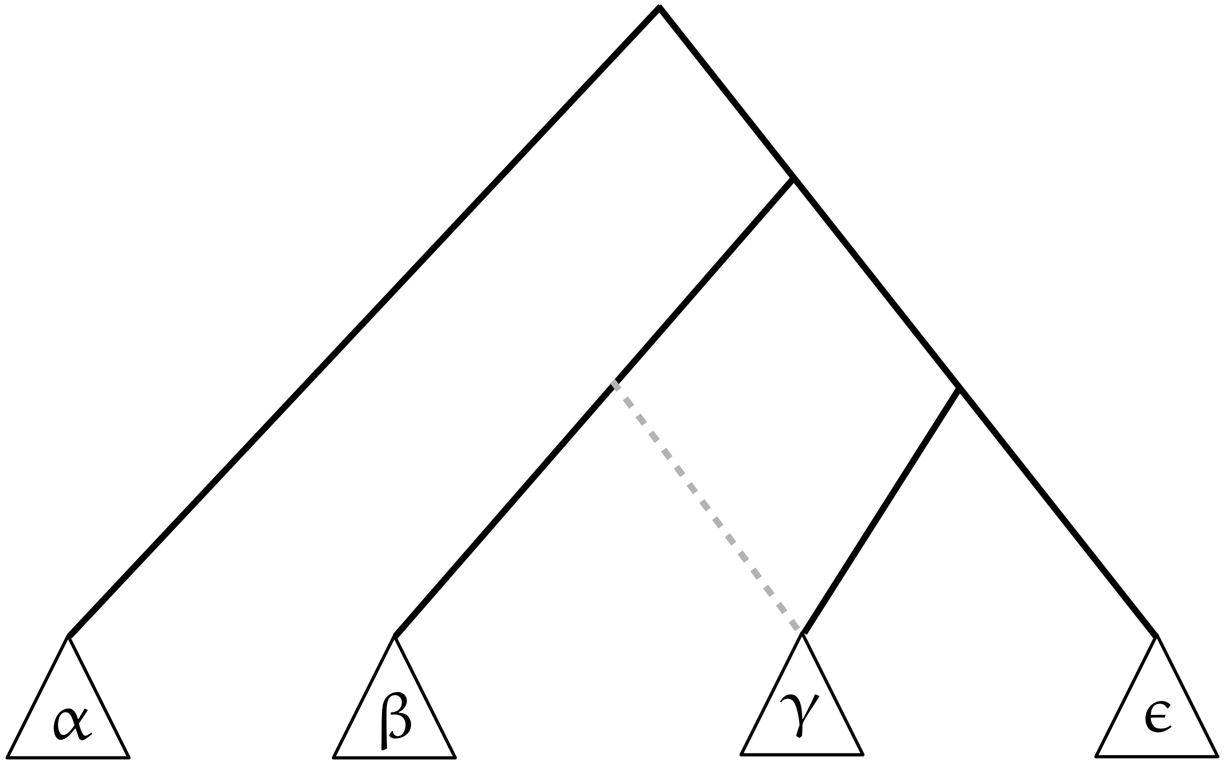
- A single edit (edge out/edge in)
- Possibly many new subtrees

↕ vs.

Reusing subtrees across all trees

- Unique subtrees encoded once, even if far apart along the genome
- Each subtree reused 5.1 times

Memory Consumption



Factoring out all unique subtrees

- A single edit (edge out/edge in)
- Possibly many new subtrees

↕ vs.

Reusing subtrees across all trees

- Unique subtrees encoded once, even if far apart along the genome
- Each subtree reused 5.1 times

`gfkit` needs 2.7 to 7.90 more space to store the trees

Limitations & Future Work

Current Limitations

- Higher memory usage on current empirical data
- Not all of `tskit`'s features supported yet, e.g.
  - branch and node-based statistics
  - partial trees
  - augmentation to full ARGs

Future Work

- Parallelization
- Top-tree based compression possible?
- Balanced-Parenthesis based encoding + string compression
- “Almost all” LCA-queries & All-Pairs LCA
- Automatic subpopulation detection

Conclusion

- Evolutionary history of recombining organisms better modelled with multiple trees
- State-of-the-art: Store edit operations between trees along the genome
- Novel approach: Encode trees as DAG, storing unique subtrees only once
- Advantage: Straight-forward memoization of intermediate results
- Speedup: 2.1 to 11.2 (median 4.0; AFS-based statistics), 100 to 1000 (LCA)
- Main drawback: Higher memory usage