GoSeqIt

Phylogenetic trees and NDtree

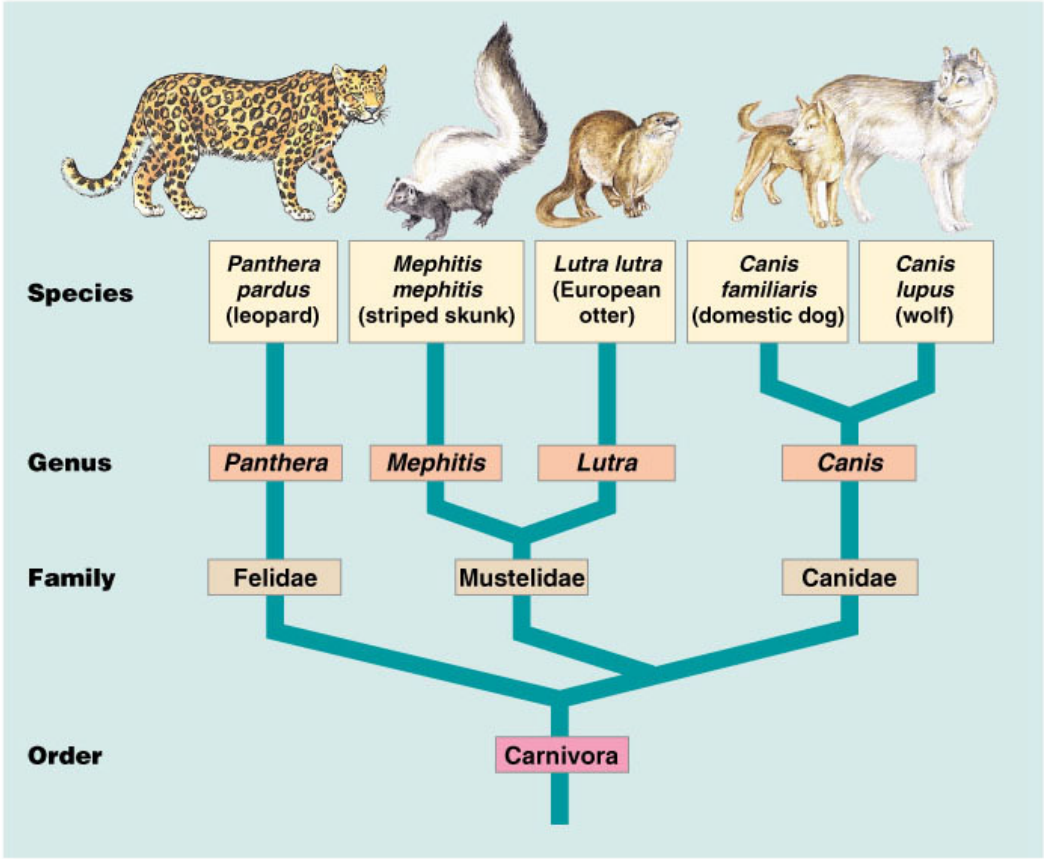
Learning objective:

After this lecture, you should be able to…

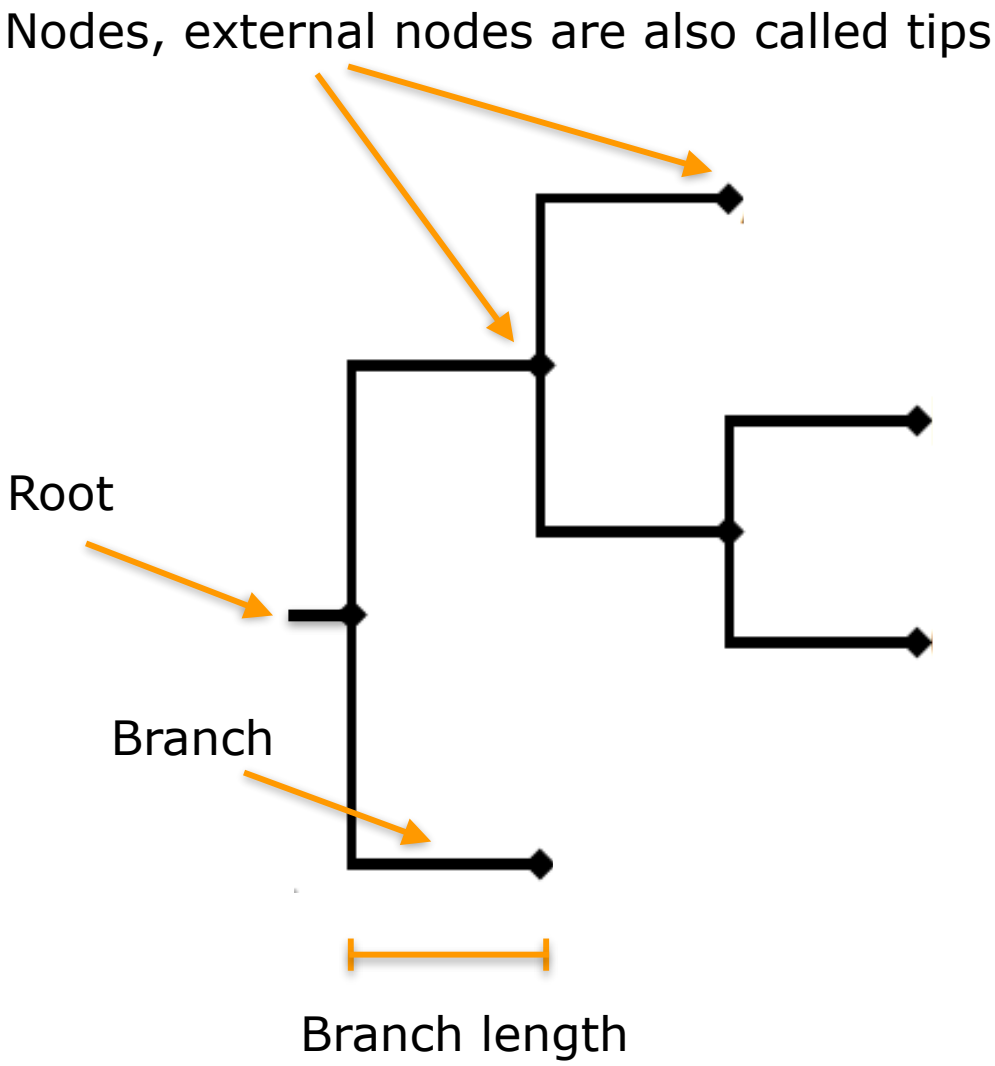
.. account for the general principles behind the CGE methods for creation of phylogenetic trees

Phylogenetic trees

- Phylogenetic trees are a visual representation of the genetic relationship between organisms
- Trees were traditionally made using aligned sequences of single genes or proteins
- Whole genome data can be used to create trees based on
  - SNP calling
  - Kmer overlap
  - Alignment of genomes

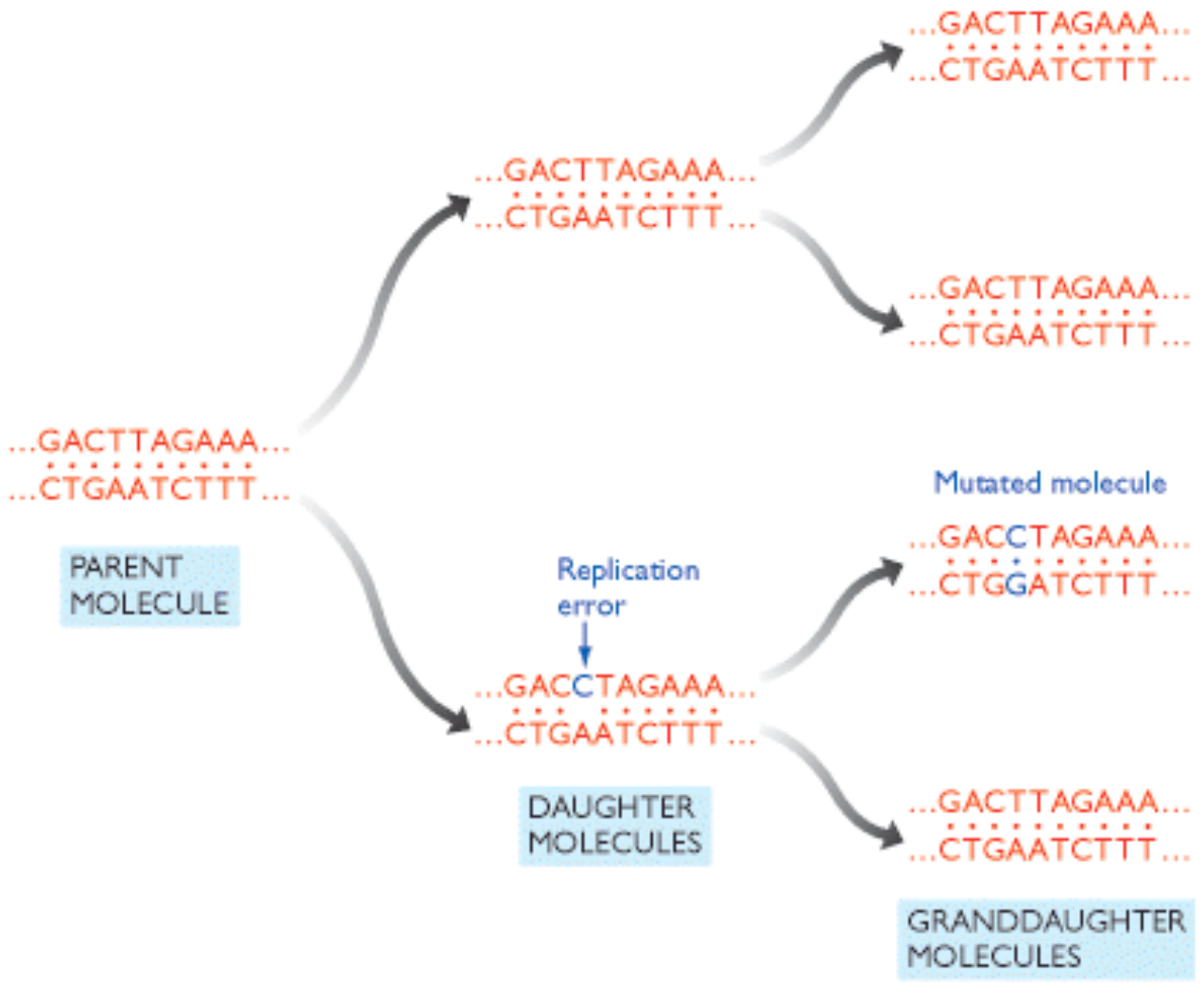


Tree terminology



- The tips represent the taxa in the study
- The "goal" of phylogenetic analysis is to recover "bifurcating" trees, in which each taxon is linked to one other taxon through a node
- Polychromous trees (multiple branches from one node) are less informative because they indicate that multiple taxa are related to each other, but not how

DNA mutations as the basis for evolution



Step 1: Construct distance matrix

Strain A	ATTCAGTAGT
Strain B	ATGCAGTTGA
Strain C	ATGCAATTGT
Strain D	ATCCATTAGC

	A	B	C	D
A				
B				
C				
D				

Create the tree

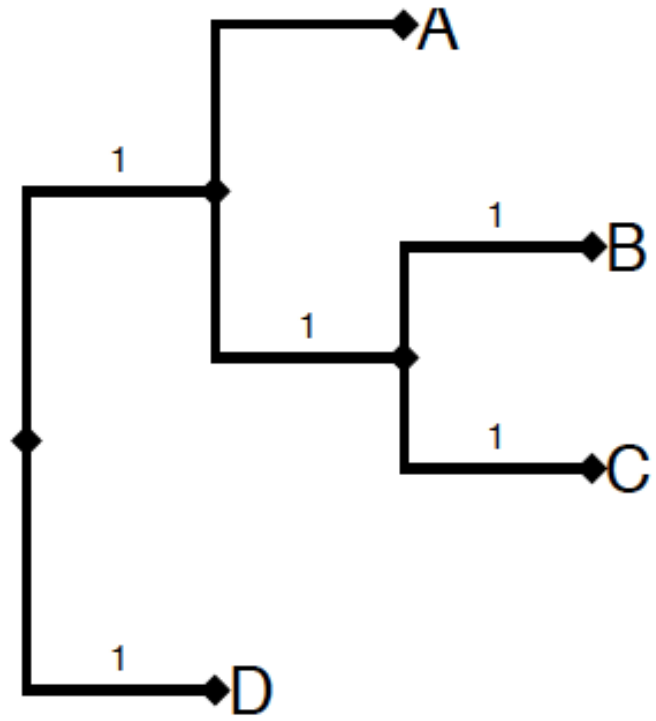
Strain A ATTCAGTAGT

Strain B ATGCAGTTGA

Strain C ATGCAATTGT

Strain D ATCCATTAGC

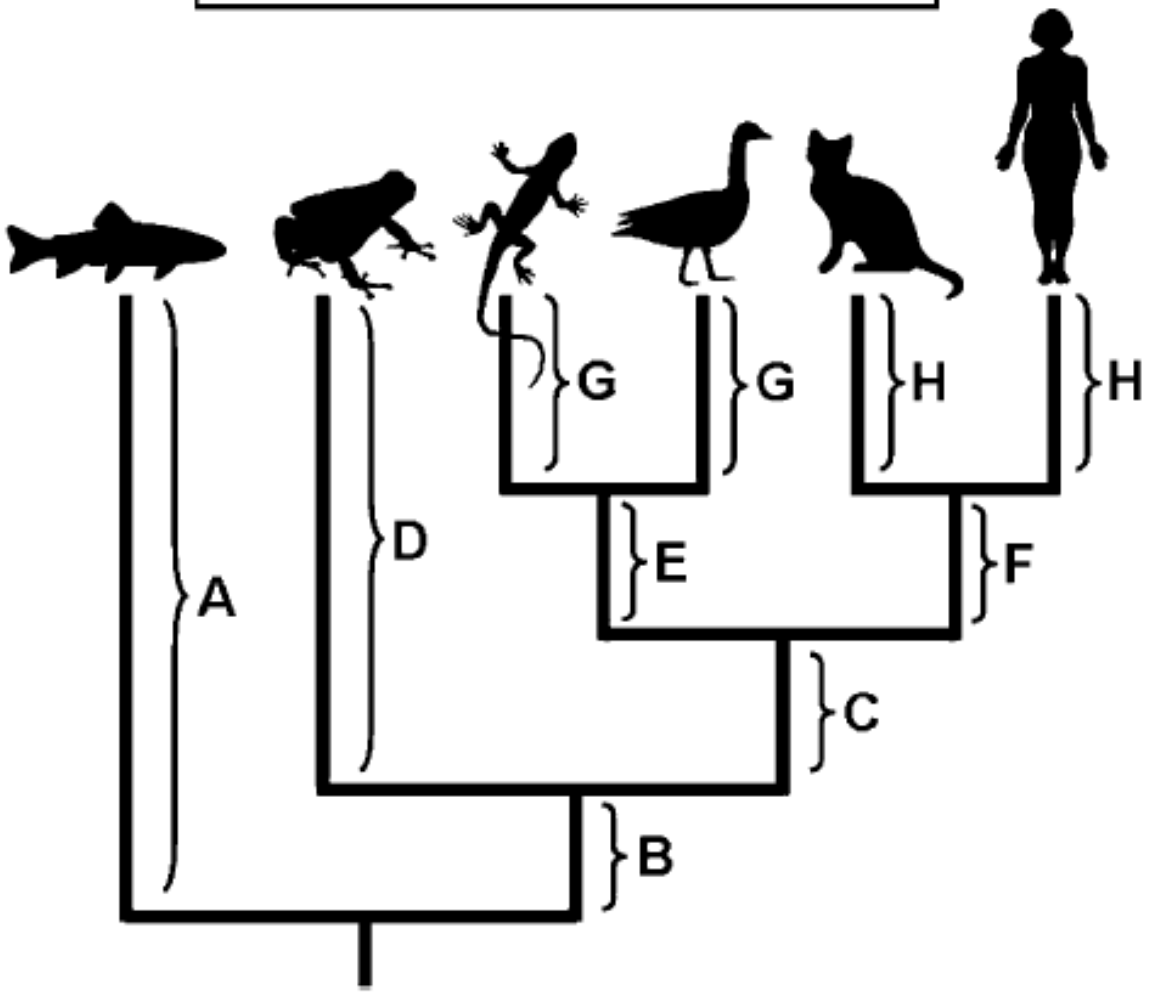
	A	B	C	D
A	0	3	3	3
B	3	0	2	4
C	3	2	0	4
D	3	4	4	0



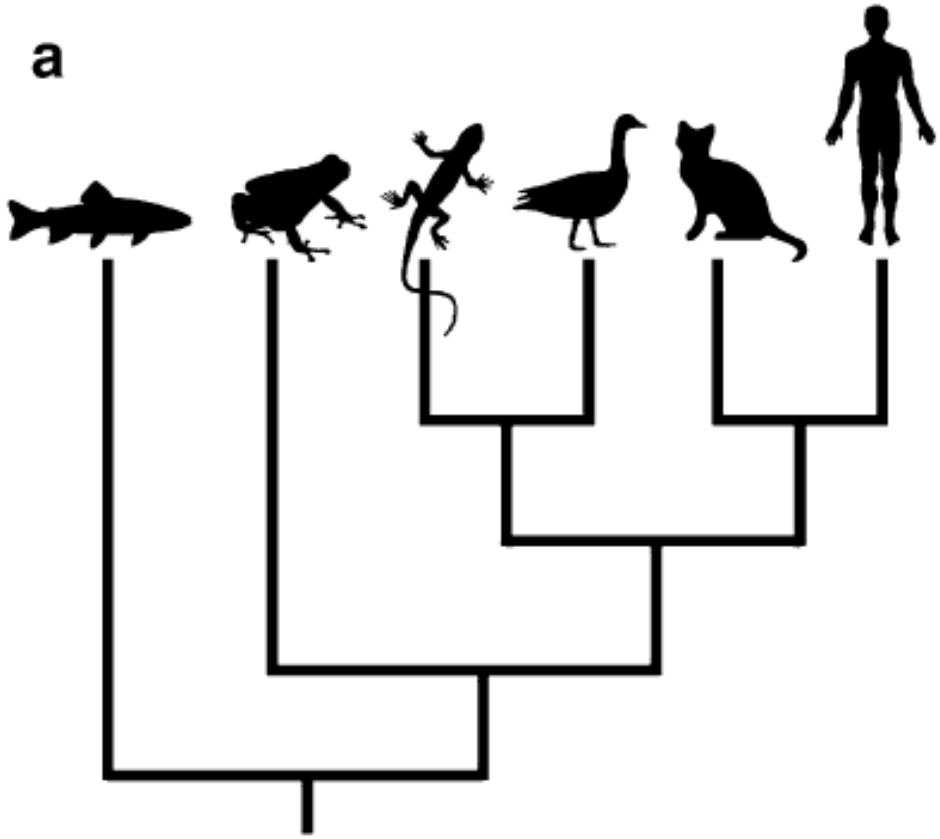
How to read phylogenetic trees

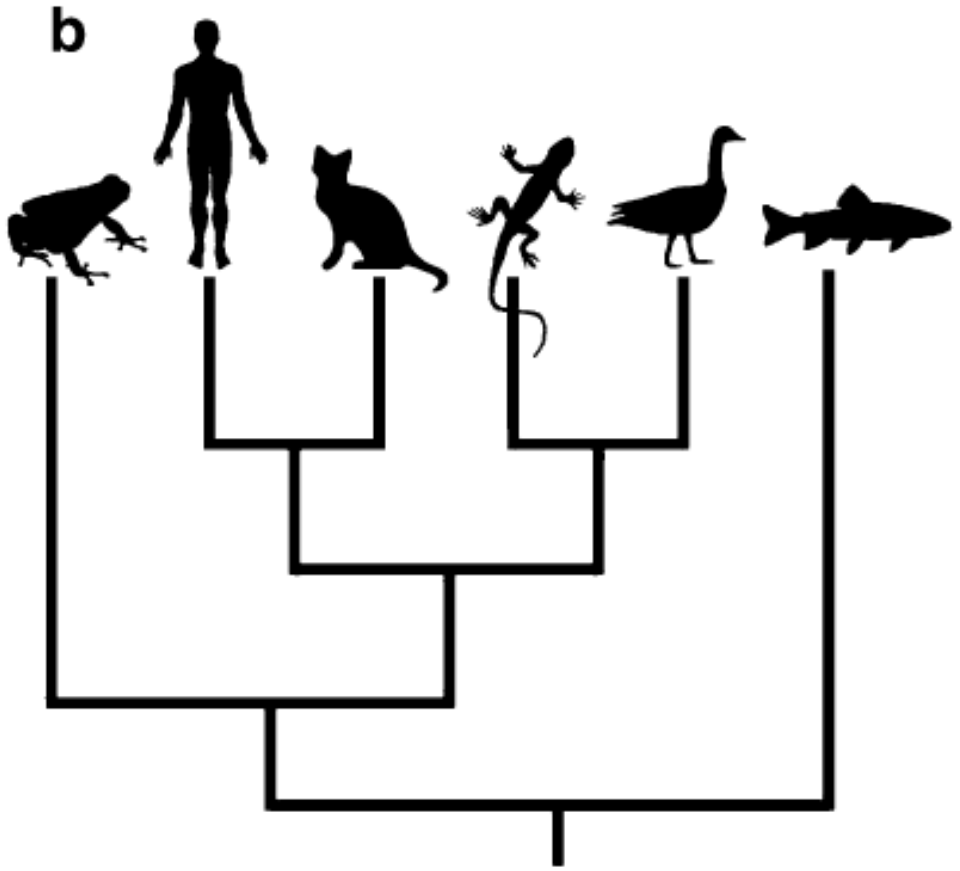
A = (B + D) = (B + C + E + G) = (B + C + F + H)

D = (C + E + G) = (C + F + H)



How to read phylogenetic trees

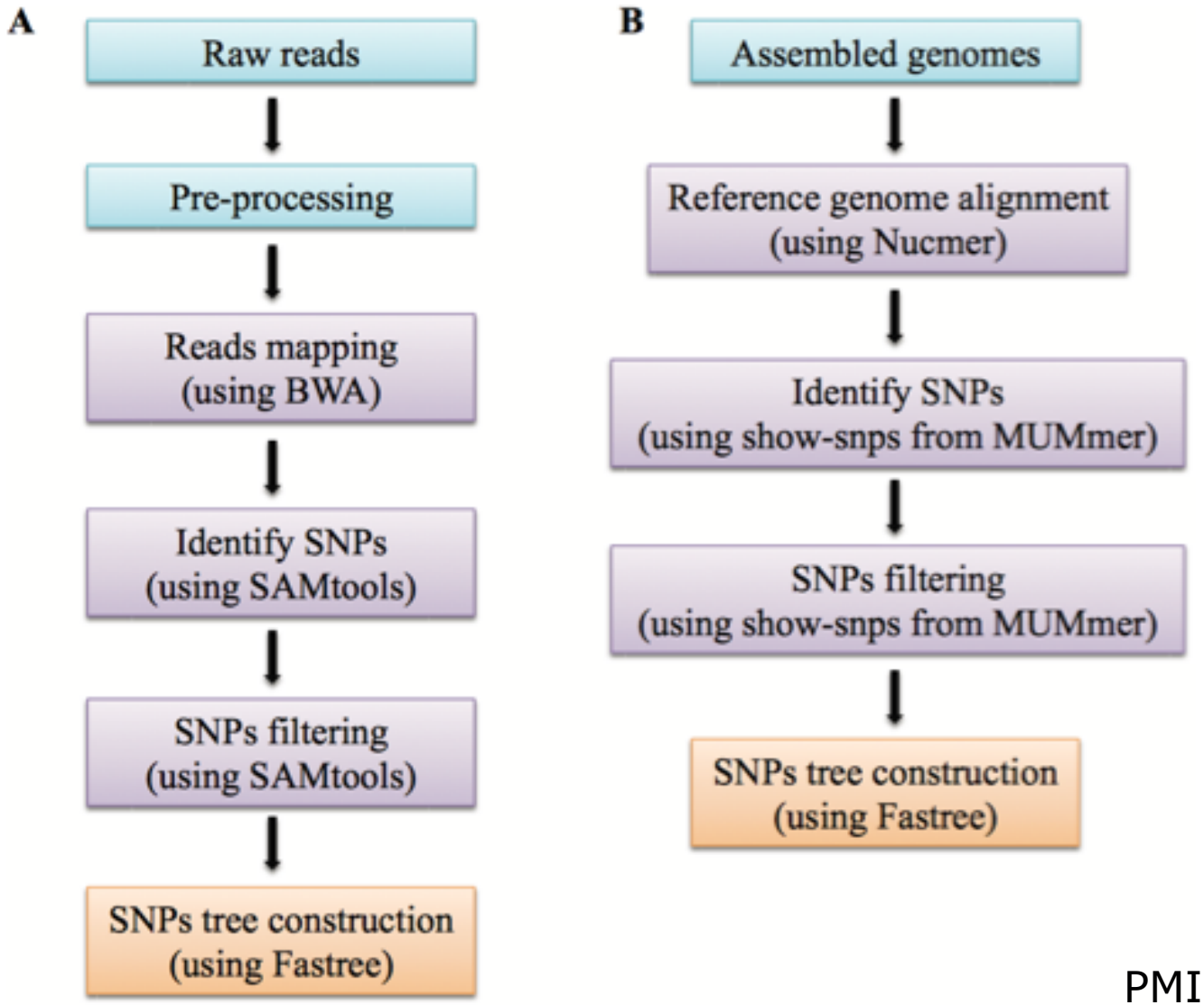




CGE tools for phylogeny

CSIPhylogeny - https://cge.cbs.dtu.dk/services/CSIPhylogeny/

SNP identification



PMID: 25110940

CGE tools for phylogeny

[illegible]t sorting of SNPs

- Rei[illegible]th
- Distance b[illegible]SNPs
- SNP quality
- Read mapping quality

Output

- Matrix of SNP pair counts in text (.tx[illegible]
- Tree build by FastTree algorithm, in Newick[illegible]

CSIPhylogeny is not available on BitBucket

NDtree

https://cge.cbs.dtu.dk/services/NDtree/

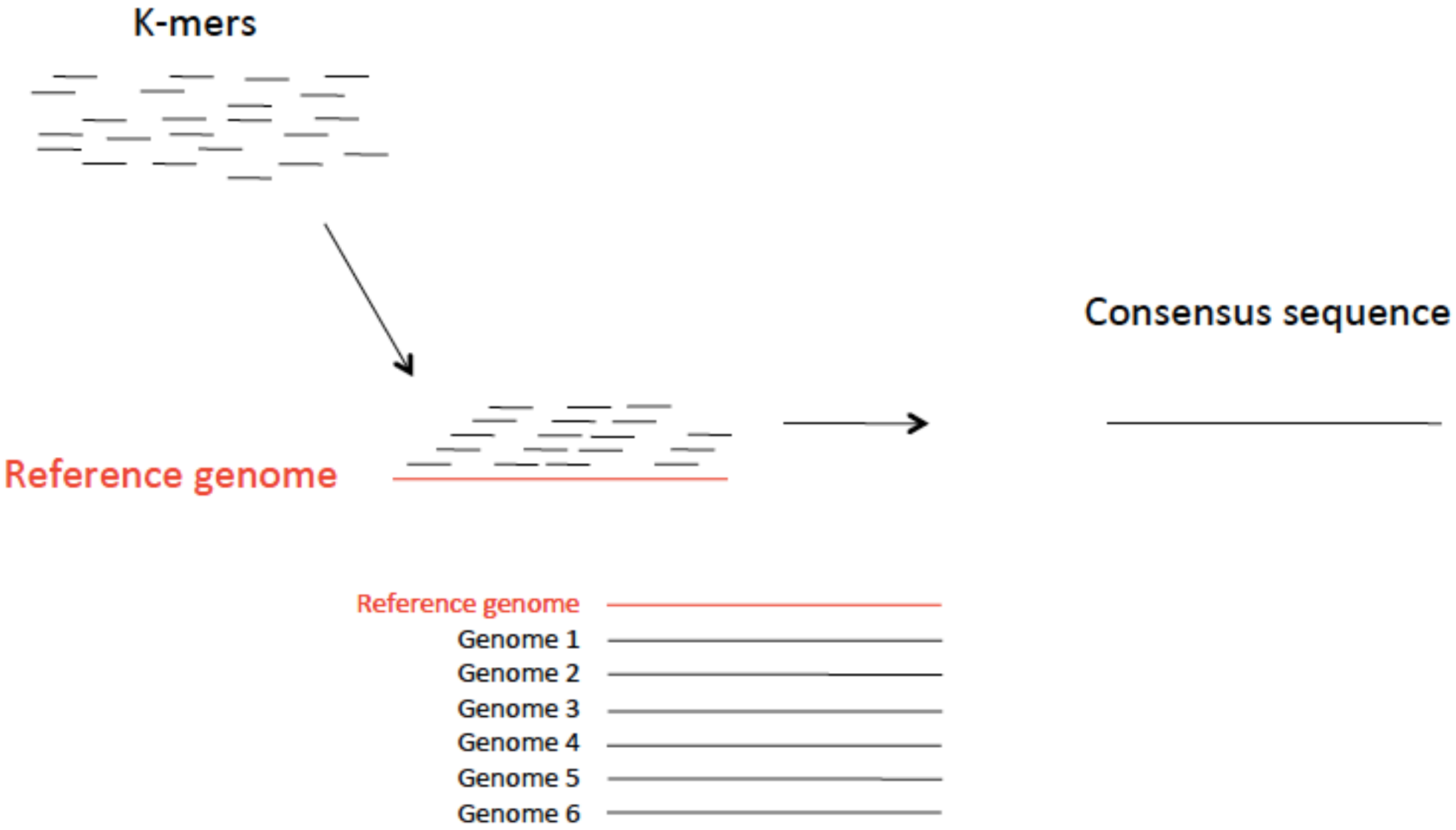
Nucleotide calling

- A different approach where the main distinction is not between if a SNP should be called or not, but between whether or not there is solid evidence for the nucleotide at the given position

Simple mapping approach

- Cuts all reads into k-mers
- Maps all k-mers to reference genome
- Makes ungapped consensus sequences of equal lengths

NDtree - mapping



NDtree

Nucleotide calling for the individual genomes

For each individual genome, all reads are mapped to the consensus sequence and the significance of the base call at each position is evaluated by calculating the number of reads X having the most common nucleotide at that position, and the number of reads Y supporting other nucleotides.

A Z-score is calculated

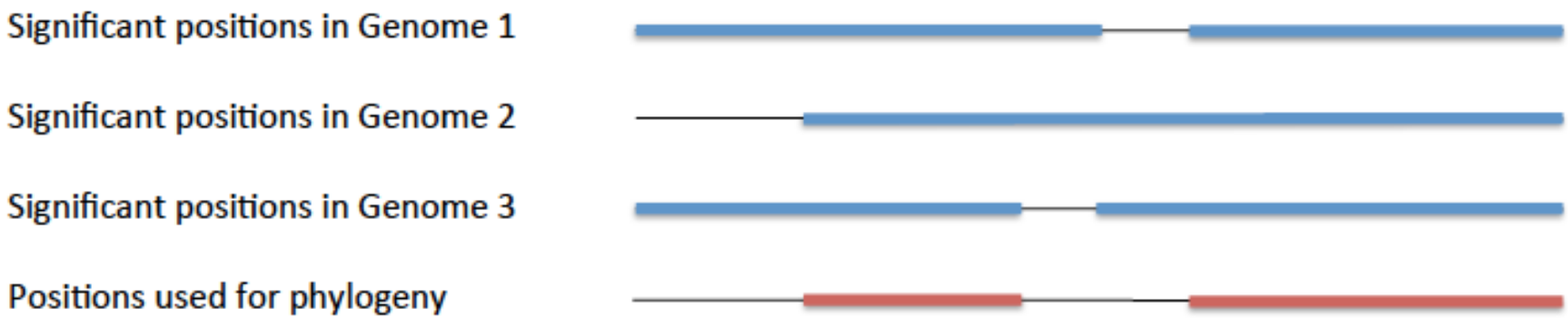
$$Z = \frac{X - Y}{\sqrt{X+Y}} \qquad > 1.96$$

> 90% of reads must support the same base

NDtree

Count nucleotide differences

Each pair of sequences is compared and the number of nucleotide differences in positions called in all sequences is counted



Choosing a reference genome

- For comparison of very closely related isolates, a better level of detail is given by using a closely related reference genome
- To examine an outbreak you could use the draft assembly of one of the outbreak strains as the reference