UNIVERSITÄTS KLINIKUM FREIBURG

Introduction to Phylogenetics

Sandra Reuter, Medical Center - University of Freiburg, Germany

Disclosure slide

Disclosure of speaker’s interests	
(Potential) conflict of interest	**none**
Potentially relevant company relationships in connection with event [1]	**none**
• Sponsorship or research funding[2] • Fee or other (financial) payment[3] • Shareholder[4] • Other relationship, i.e. …[5]	none

Comparison of bacterial genomes

Chromosomal rearrangements

Pathogenicity islands

Prophages

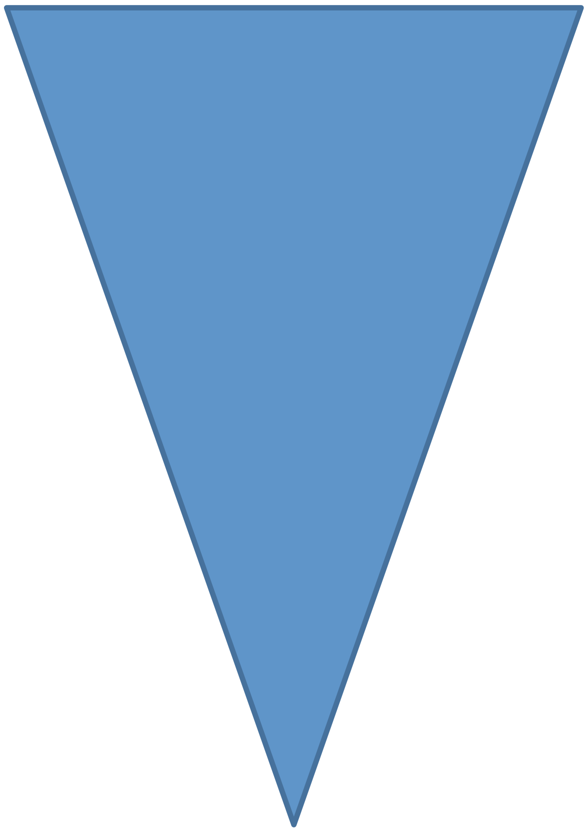
Gene acquisition and loss

Gene rearrangements

Pseudogenes

Phase variation

Whole genomes



Single nucleotides

Comparison of bacterial genomes

Single genomes / nucleotides

- DNA sequence browser (Artemis)
- Investigate the makeup of a single (representative) genome
- First genome projects

Multiple genomes (2-6)

- Artemis Comparison Tool (ACT)
- Direct pairwise comparison
- Detect chromosomal difference between a limited number of fully sequenced genomes

But what about “unlimited” genomes?

Multilocus-sequence typing (MLST)

Uses 7 housekeeping genes

Allele profile > Sequence Type (ST) > Clonal Complex (CC)

Advantages:

Portable, transferrable, unchangeable

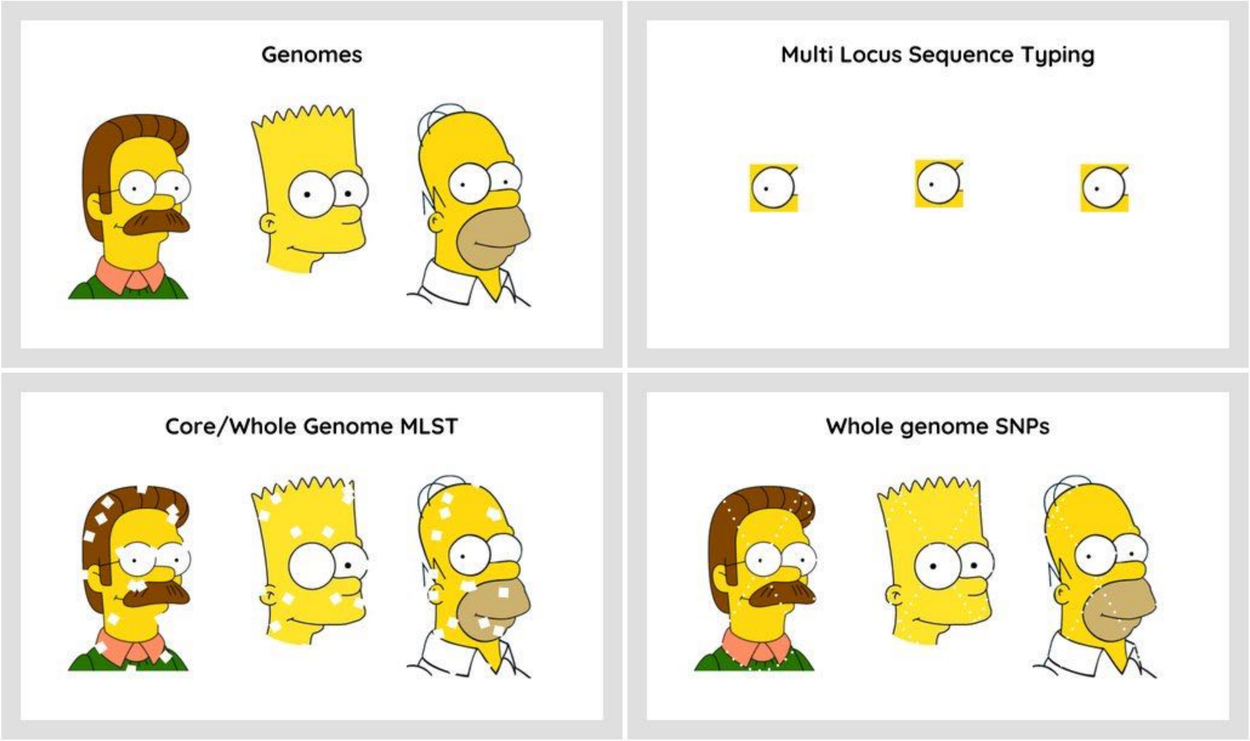
Disadvantages:

Limited resolution in outbreaks and epidemic circulating clones

Multilocus-sequence typing (MLST)

With epidemic clones circulating, typing is at its limit

- EMRSA-15, *K. pneumoniae* ST258, *E. coli* ST131



@torstenseemann

Resequencing

Aims to capture information on

- Single Nucleotide Polymorphisms (SNPs)
- insertions and deletions (indels)
- Copy Number Variants (CNVs)

between variants of the same bacteria

As sequences diverge from the reference, mapping becomes progressively less effective

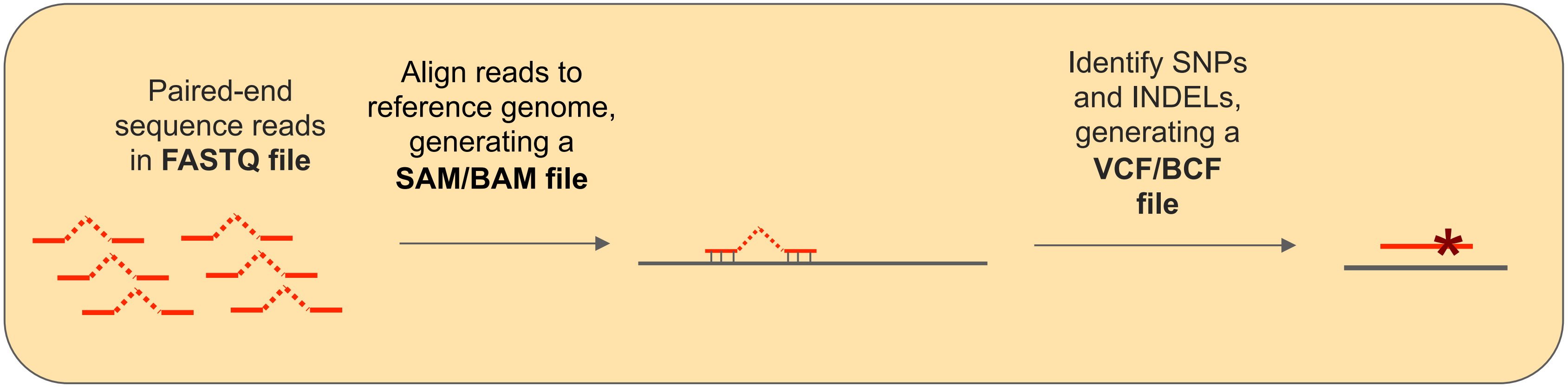
Will give information on the “core genome” – what is shared between isolates (e.g. of a species) – but not on the accessory genome – what is shared only between selected members, or which are unique to a sample

Steps in mapping

Choose a fully finished reference genome

Take fastq reads from machine

Use alignment software (BWA, smalt, tophat,…) to find matches in the reference genome



Paired-end sequence reads in **FASTQ file**
Align reads to reference genome, generating a **SAM/BAM file**
Identify SNPs and INDELs, generating a **VCF/BCF file**
*

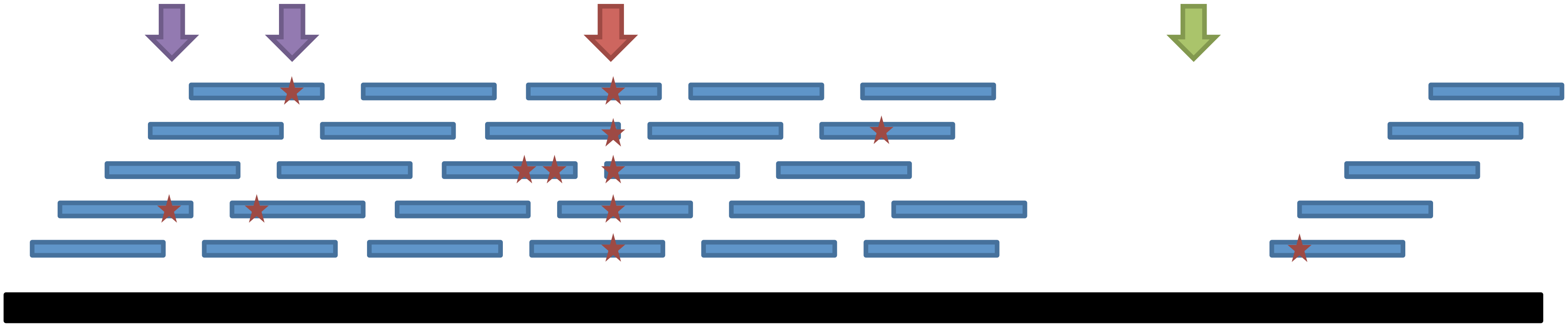
Steps in mapping

Imagine sequencing a zebra…

Sequencing errors

True SNP

Absent region



Reference genome

Steps in mapping

After raw read mapping: filtering

Low quality reads

Low quality mapping

Consider indels (short insertions/deletions)?

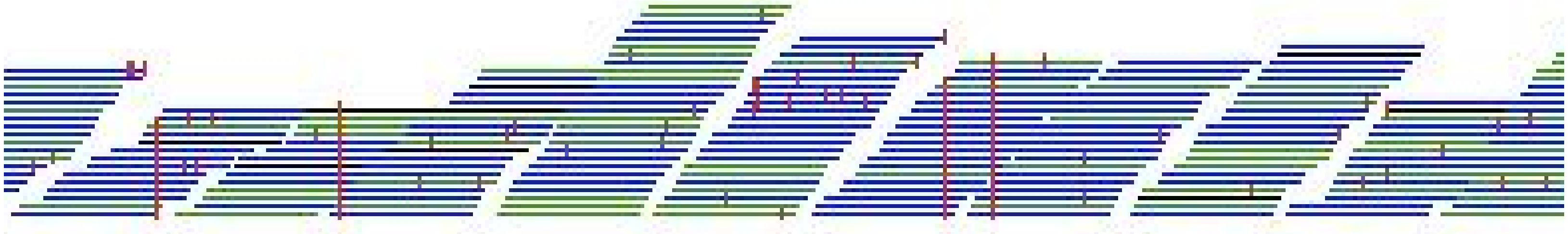
Filter for read depth (e.g. only accept SNPs if in at least 4 reads)

Filter SNPs: presence in at least 75% of reads

What are we looking for / what can mapping do for you?

Phylogenetics

SNP calling



Looking at copy number

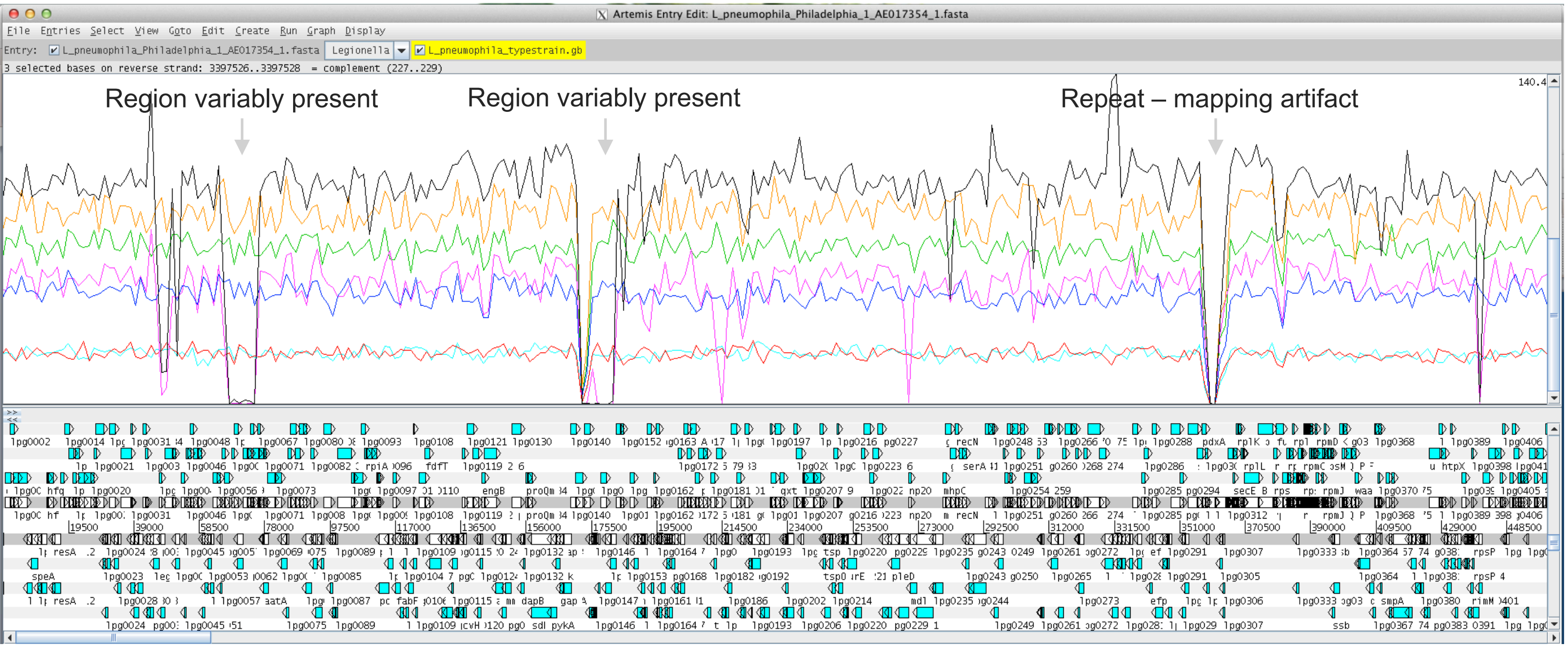
Looking at presence/absence

Checking for errors

Sequencing quality control

Suitability of chosen reference

Assess sequencing quality and coverage



Artemis Entry Edit: L_pneumophila_Philadelphia_1_AE017354_1.fasta

Entry: ☑ L_pneumophila_Philadelphia_1_AE017354_1.fasta Legionella ☑ L_pneumophila_typestrain.gb

3 selected bases on reverse strand: 3397526..3397528 = complement (227..229)

Region variably present

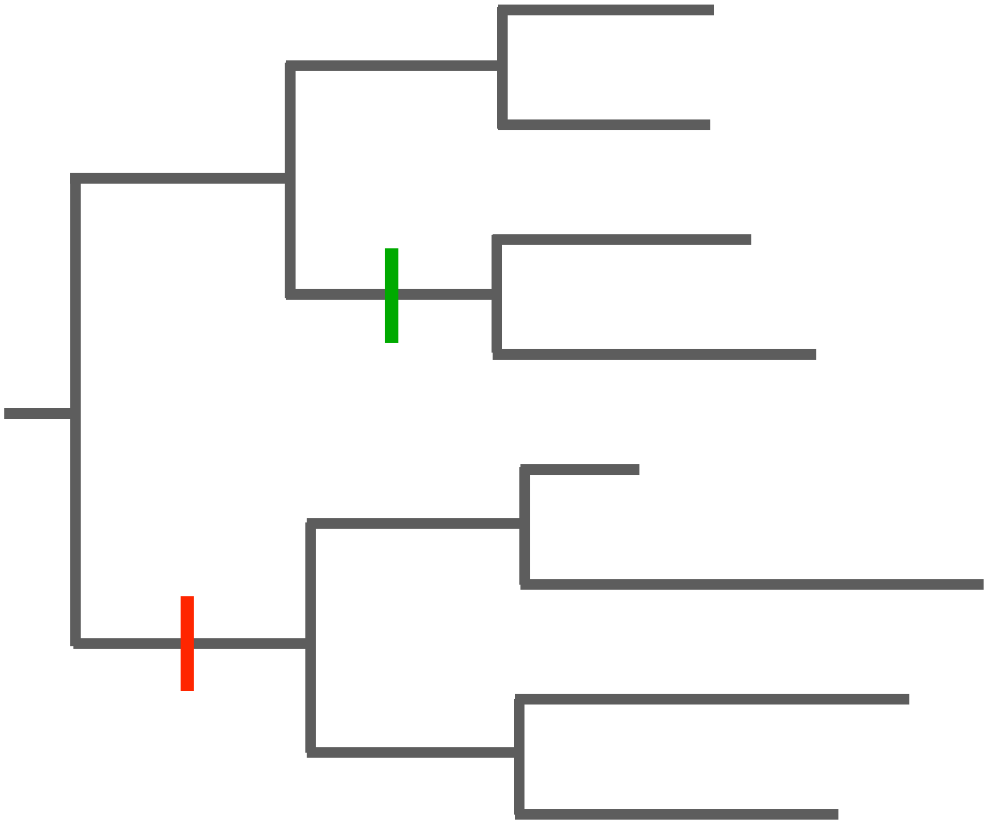
Region variably present

Repeat – mapping artifact

Black: LP617 ~100X
Orange: LP467 ~90X
Green: LP035 ~65X
Pink: LP423 ~60X
Blue: LP056 ~60X
Red: LP033 ~20X
Sky: LP427 ~20X

SNPs can be used to draw a phylogenetic tree

If a SNP is shared by a number of isolates, it is evidence that they may be related and form a group on the tree



ACTTGTAC

ACTTGTAC

ACTTGTTC

ACTTGTTC

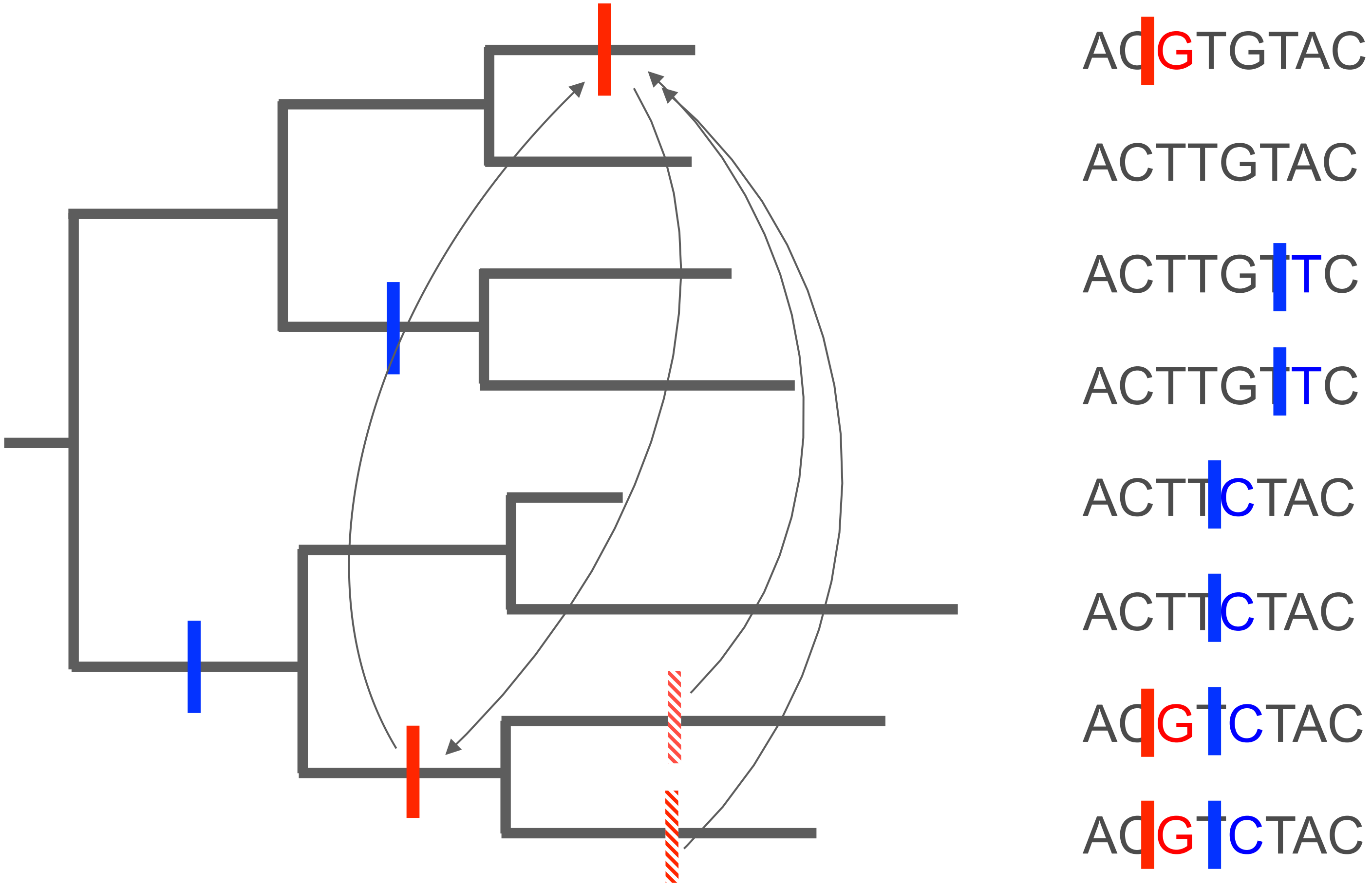
ACTTCTAC

ACTTCTAC

ACTTCTAC

ACTTCTAC

Homoplasies do not “fit” the tree

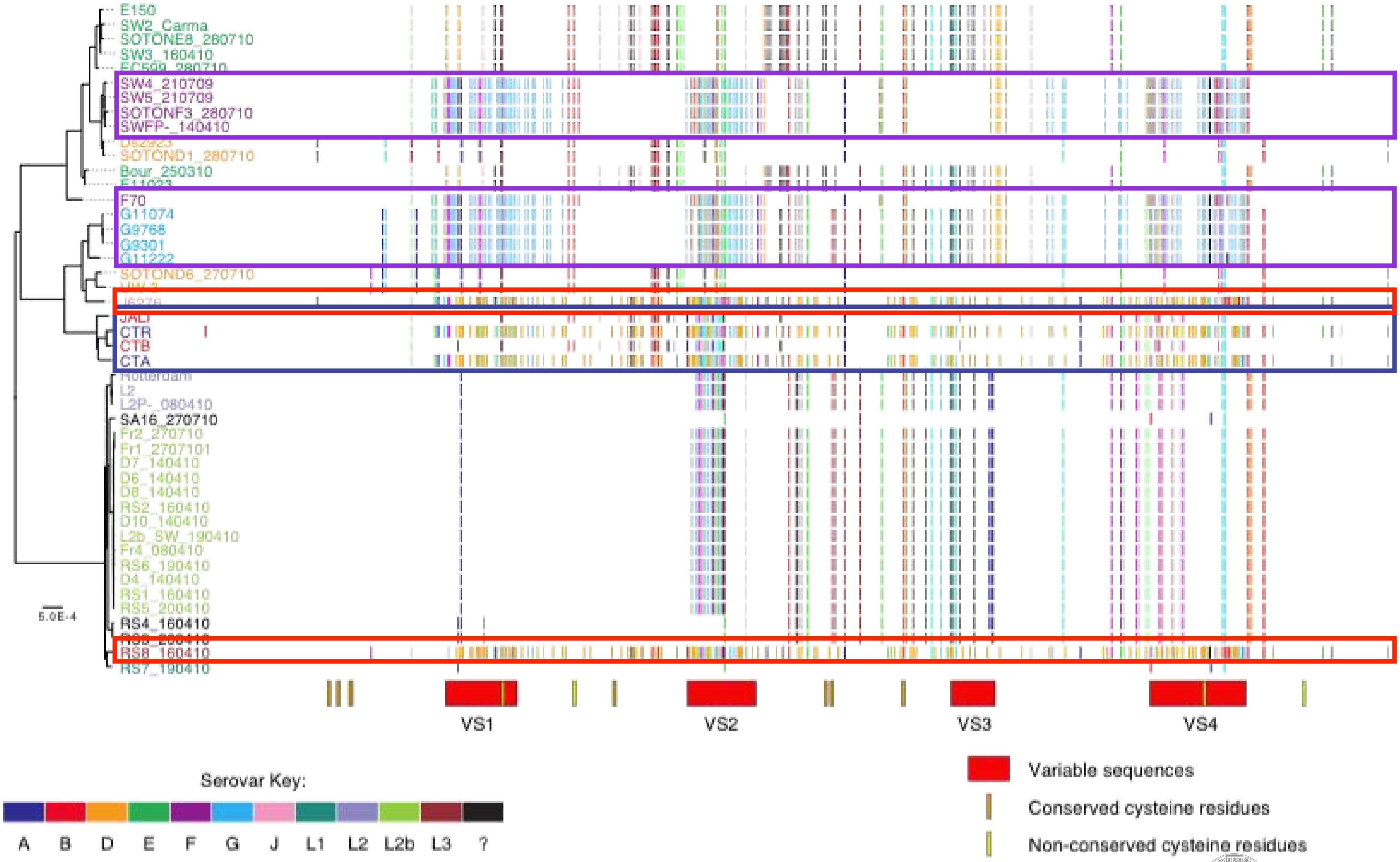


Blue = SNP that fits tree. Red = Homoplasy
Single SNPs may arise independently.
However, if multiple SNPs/patterns are consistent, they may be a sign of recombination!

42000 06200

SNP Barcode

Recombination of *ompA* in *Chlamydia trachomatis*

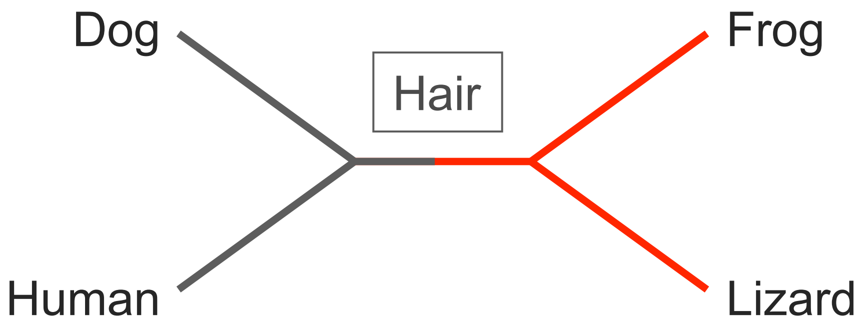


Thomson, Seth-Smith, Harris personal communication

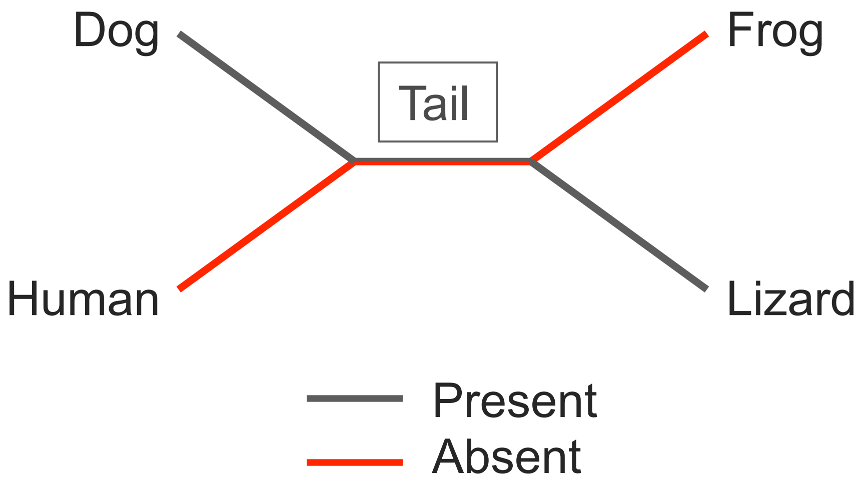
Background information on phylogenetic trees

Homology vs Homoplasy:

- **Homology** describes similarity due to common inheritance from an ancestor. Homologous characters are useful similarity.

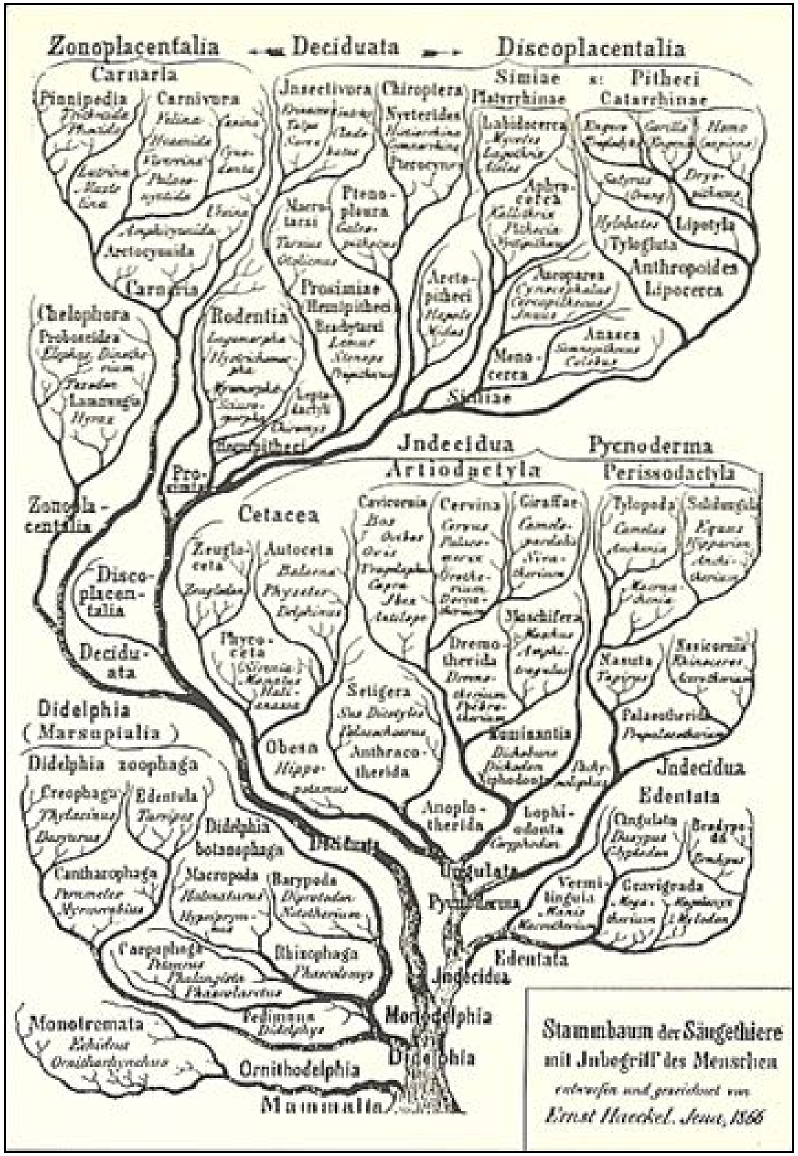


- **Homoplasy** describes similarity due to independent acquisitions of the same or superficially similar character states. Homoplasic characters provide a misleading picture of phylogeny.



Phylogenetic Systematics

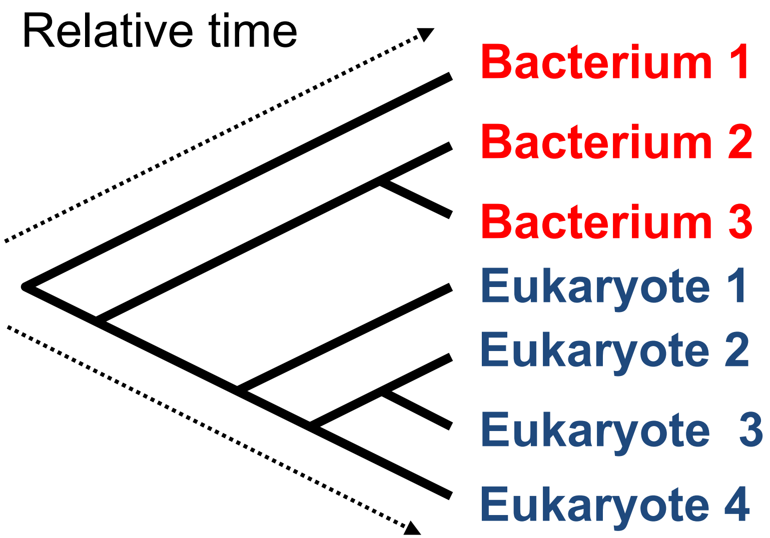
- Phylogenetics aims to reconstruct the ancestry of biological lineages
- It regards homology as evidence of common ancestry
- Relationships are usually portrayed on tree diagrams
- Monophyletic groups (clades) contain taxa that are more closely related to each other than to any outside the group
- Distance between taxa reflects a decreasing number of shared, homologous characters



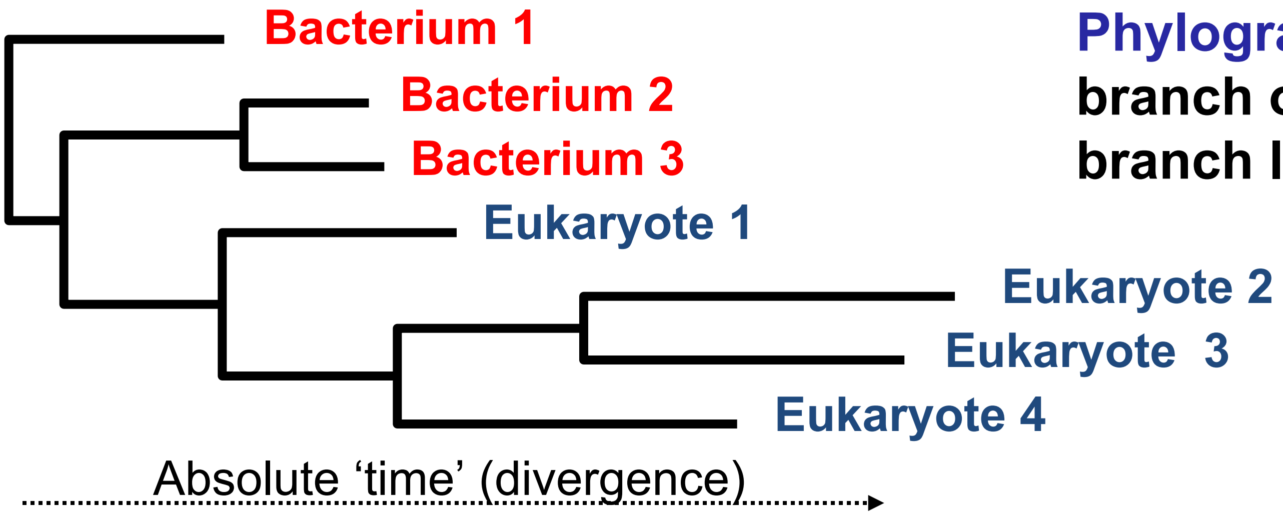
Stammbaum der Säugethiere mit Inbegriff des Menschen

entworfen und gezeichnet von Ernst Haeckel. Jena, 1866

Cladograms and Phylograms

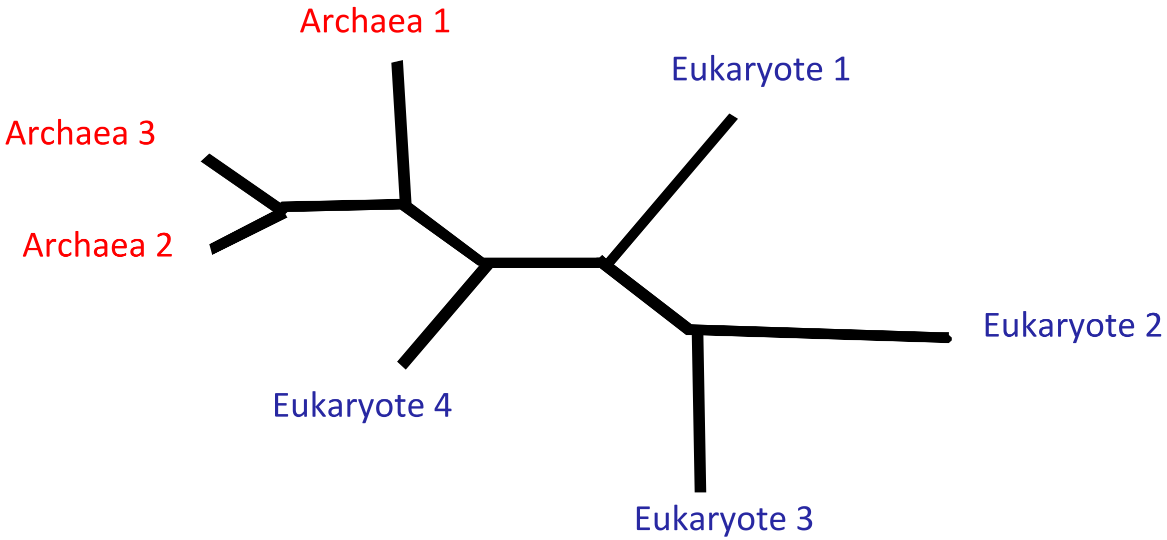


Cladograms show branching order - branch lengths are meaningless



Phylograms show branch order and branch lengths

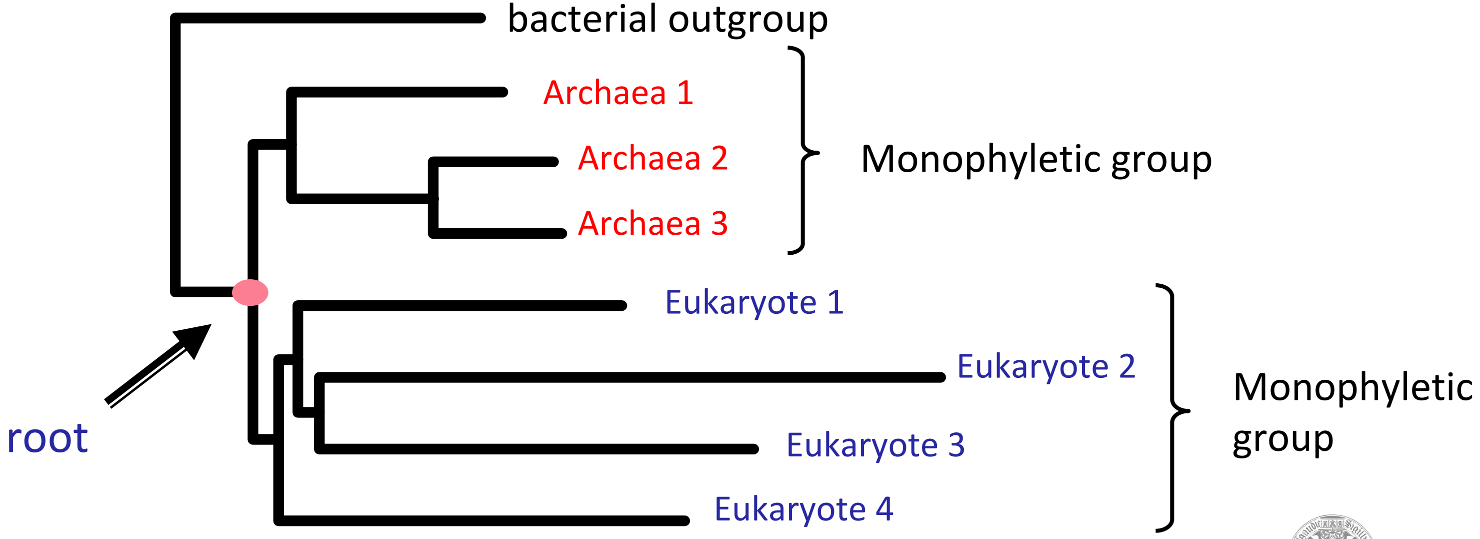
Rooted and unrooted trees



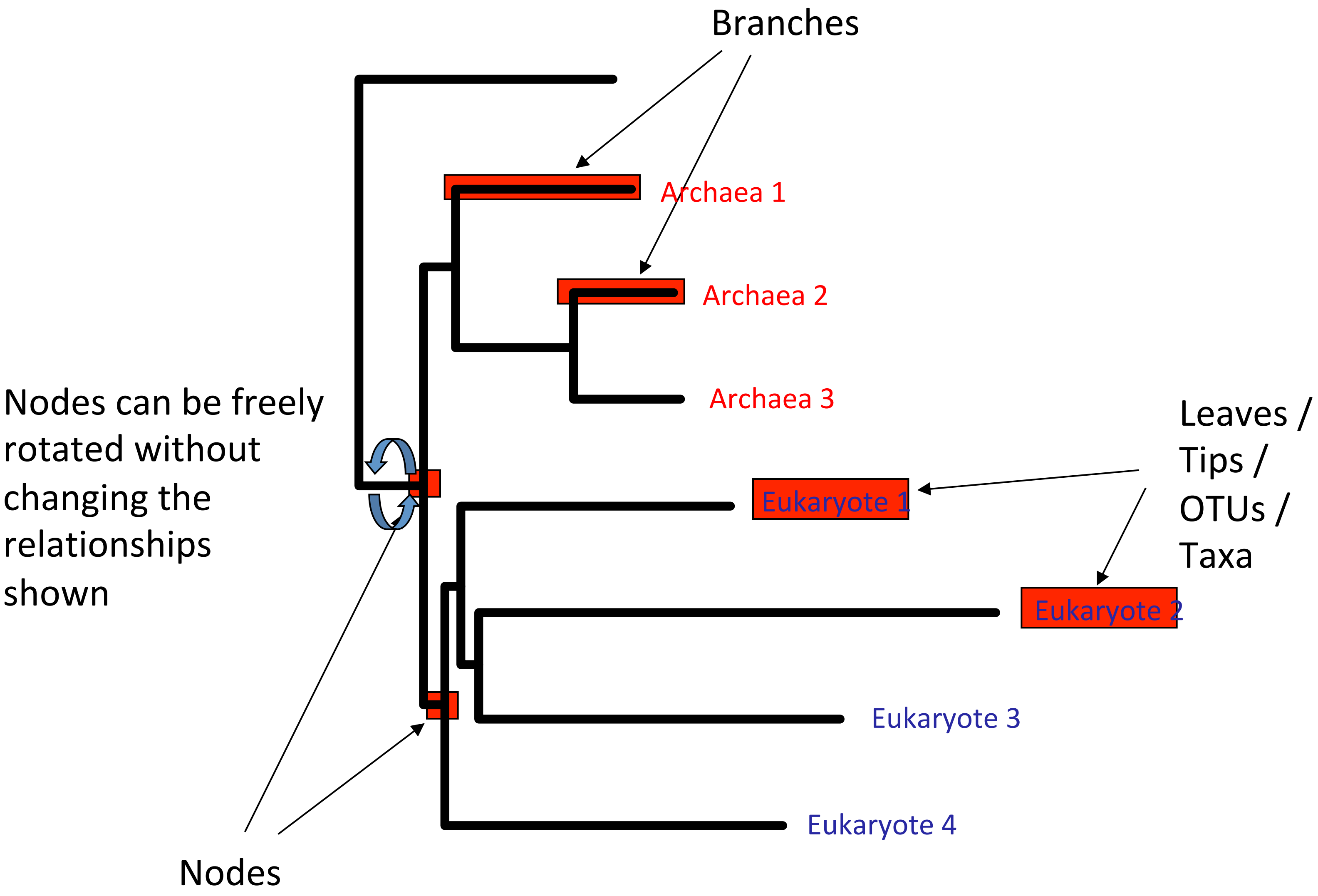
Unrooted tree

The root defines common ancestry

Tree rooted by outgroup



Some tree terms and facts



Branches

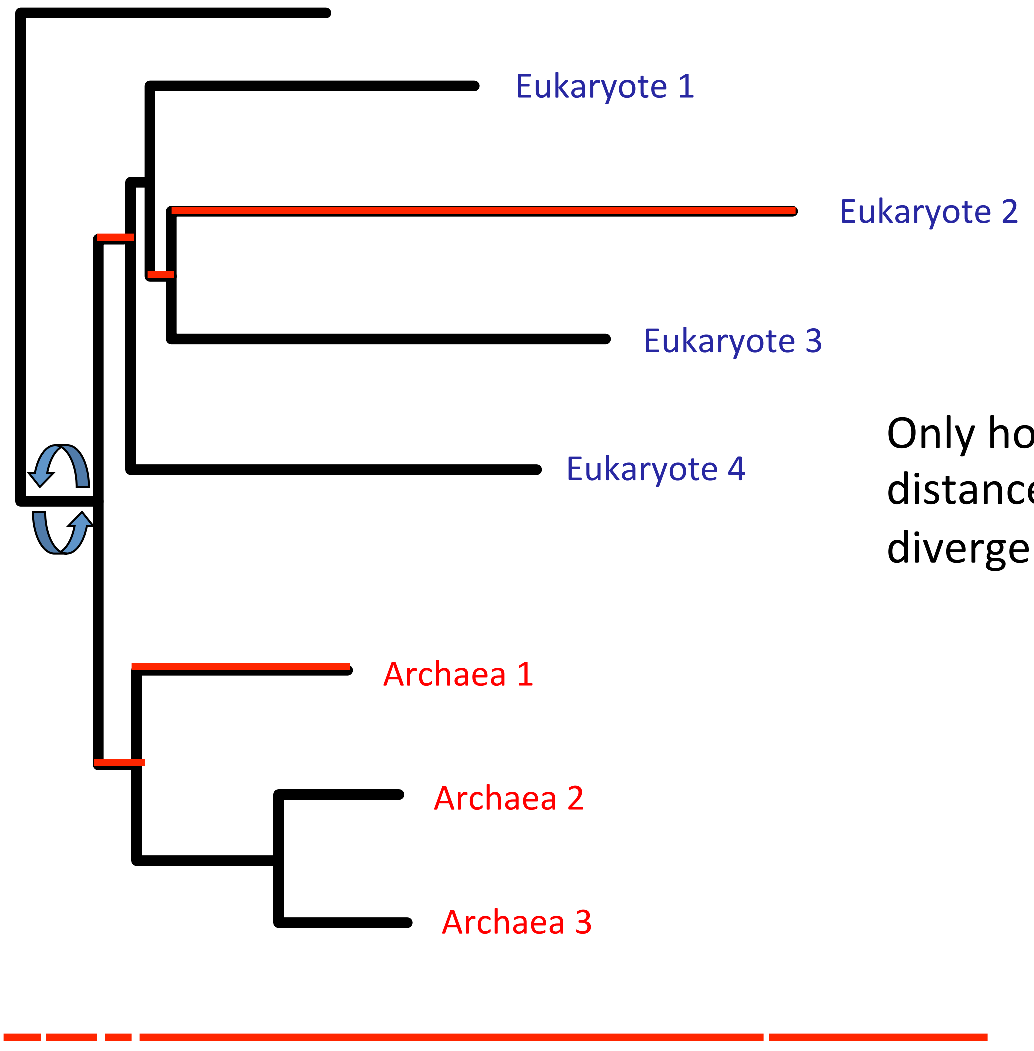
Nodes can be freely rotated without changing the relationships shown

Leaves / Tips / OTUs / Taxa

Nodes

Some tree terms and facts

Nodes can be freely rotated without changing the relationships shown

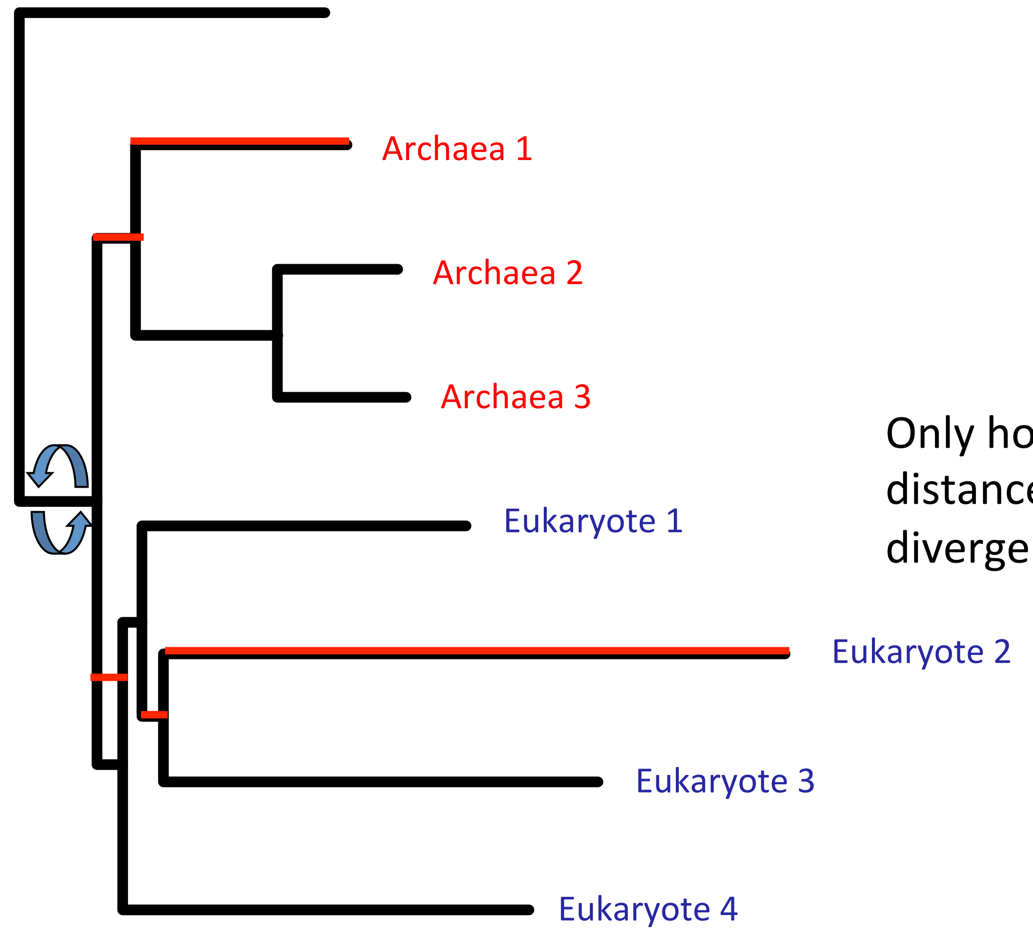


Only horizontal distances indicate divergence

Total distance =

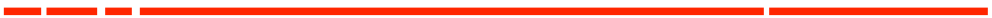
Some tree terms and facts

Nodes can be freely rotated without changing the relationships shown

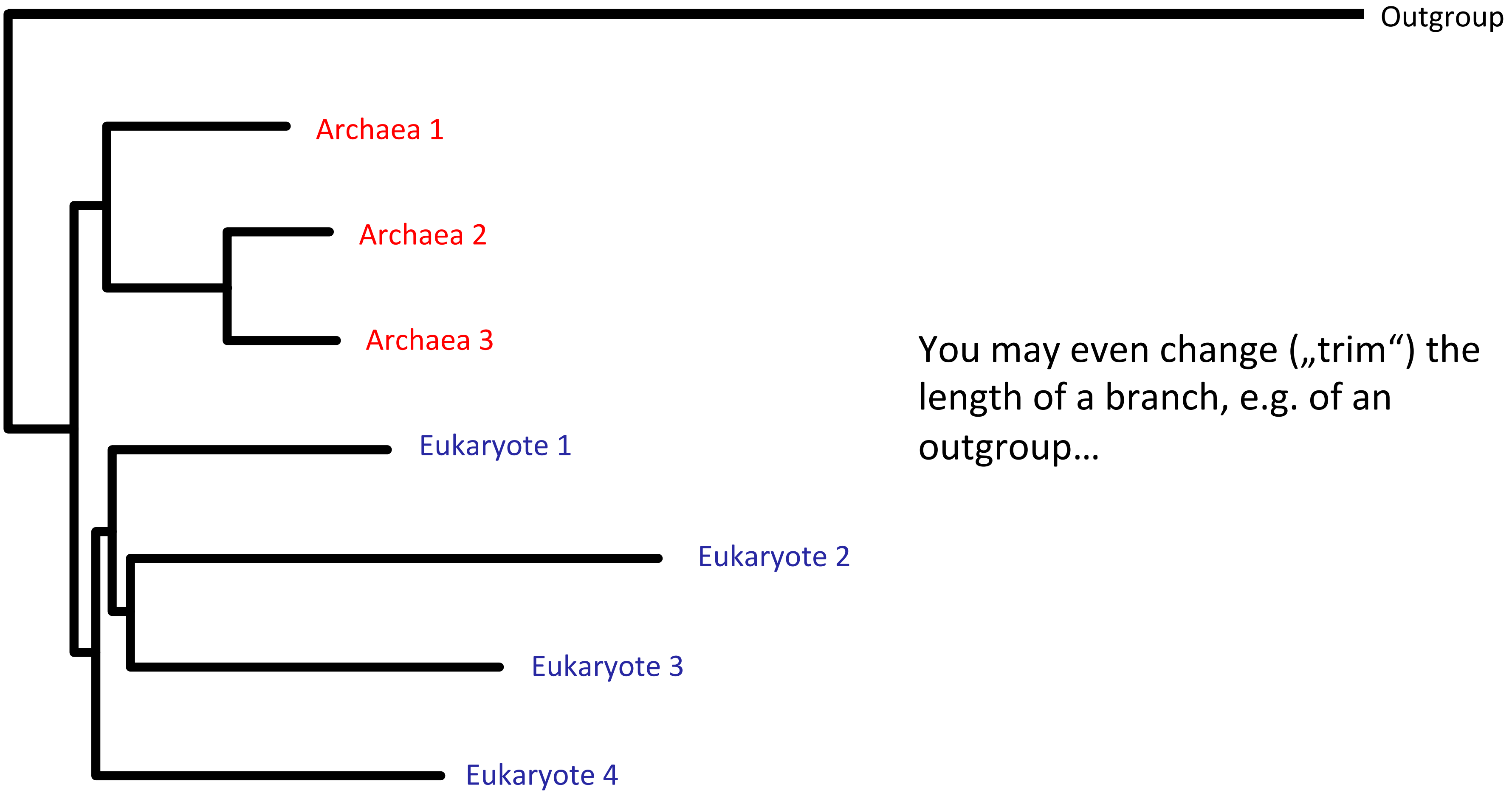


Only horizontal distances indicate divergence

Total distance =

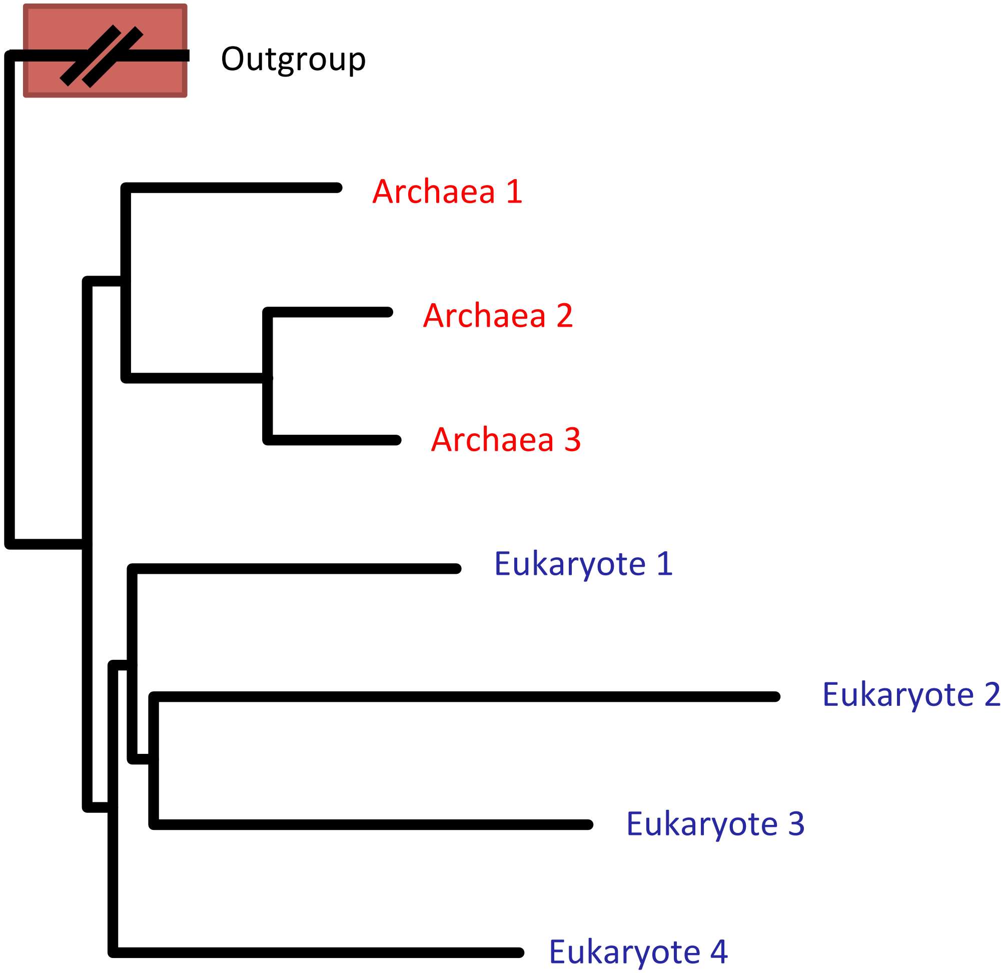


Some tree terms and facts



You may even change („trim“) the length of a branch, e.g. of an outgroup…

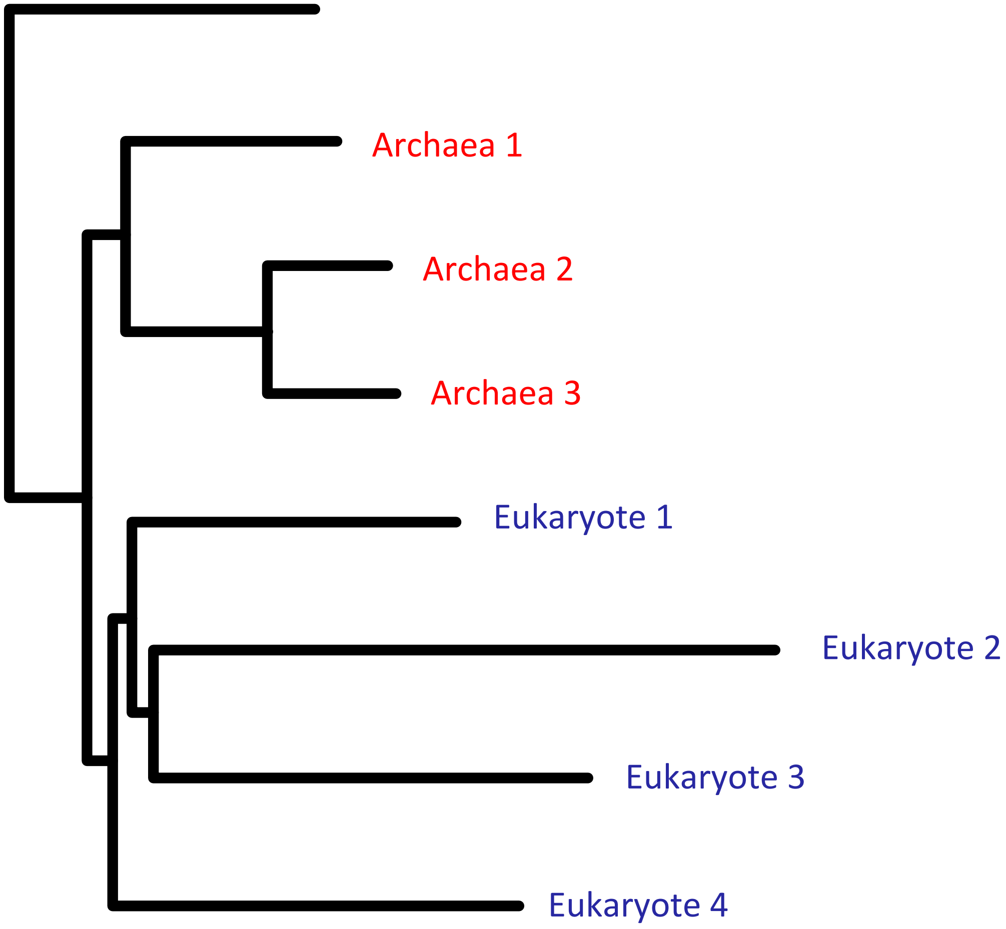
Some tree terms and facts



You may even change („trim“) the length of a branch, e.g. of an outgroup…

…but you need to indicate that you‘ve done so!

Some tree terms and facts



Scale bar

~10 SNPs or 500 SNPs or ~10,000 SNPs…

0.02 – substitutions / nucleotide

> multiply by aln length

Building a phylogenetic tree

- Identify protein, DNA or RNA sequences of interest
  - Fasta format file of concatenated sequences
- Multiple sequence alignment – not for mapping-based trees!
  - ClustalX/muscle
- Construct phylogeny
  - PHYML, RAxML
- View and edit tree
  - FigTree, iTOL, microreact

Note: There are many (many) other programs for alignment, tree building and tree viewing

Estimation of a phylogenetic tree

- Phylogenetic Markers (e.g. 16S rDNA)
  - Ubiquitous distribution
  - Functional consistency (homology)
  - Size (proportional to that information content)
  - Conserved as well as highly-variable structural elements
  - No horizontal / lateral gene transfer (recombination)

Constructing phylogenies

- Stages in phylogenetic analysis:

  1. *Data preparation*
     multiple alignment (DNA / protein)
  2. *Data scoring*
     distance methods: pairwise distances between sequences
     discrete methods: each site in the alignment as a character
  3. *Tree sorting*
     processes for searching ‘tree-space’
  4. *Estimation*
     identifying the most acceptable tree topology and model parameters using a variety of methods (‘clustering’ or ‘optimising’ methods).

- Phylogenetic methods:

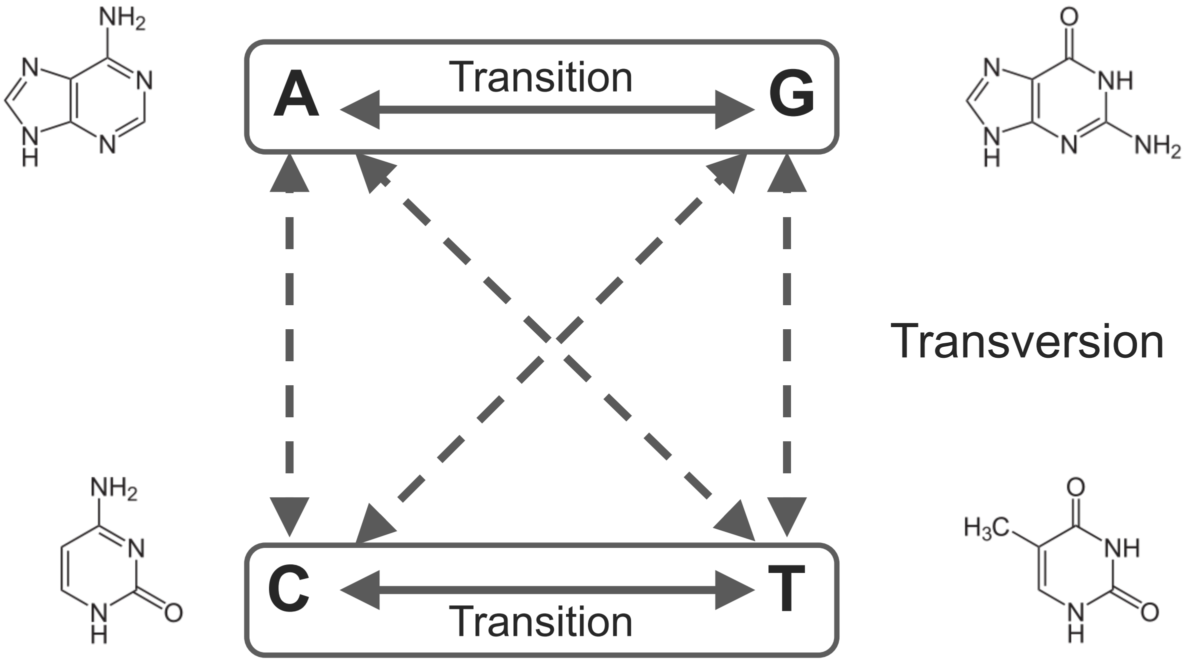
	Clustering	**Optimising**
Distance	Neighbour-joining UPGMA	Minimum evolution
Discrete		Maximum parsimony Maximum likelihood Bayesian inference

Tree estimation

- Evolutionary models
  - Jukes Cantor (JC)
    - JC69: all substitutions equally likely, all bases same frequency
  - Kimura 2 Parameter (K2P), Hasegawa/Kishino (HKY85)
    - Specific likelyhoods for transition and transversions, all bases same frequency
  - General Time Reversal (GTR)
    - GTR: each substitution with their own likelyhood, depending on specific base frequency
  - ➢ Depending on the model, the tree will change

Purine Bases

Pyrimidrine Bases



Tree estimation – distance methods

Method

- Pairwise distances between taxa are calculated (many options)
- Tree topology and branch lengths are estimated from this distance matrix.
- E.g. Neighbour-joining, UPGMA, Minimum Evolution

```
ACGGACCTATCTGGTCTAATTAAA
|X|||||X|||X||||||||||||
ATGGACCAATCCGGTCTAATTAAA
```

```
P distance          010000010001000000000000 = 3
```

With an evolutionary model, e.g. transversions with a higher score than transision:

```
010000020002000000000000 = 5
```

👍 a single tree is estimated, in short time, minimal computational expense

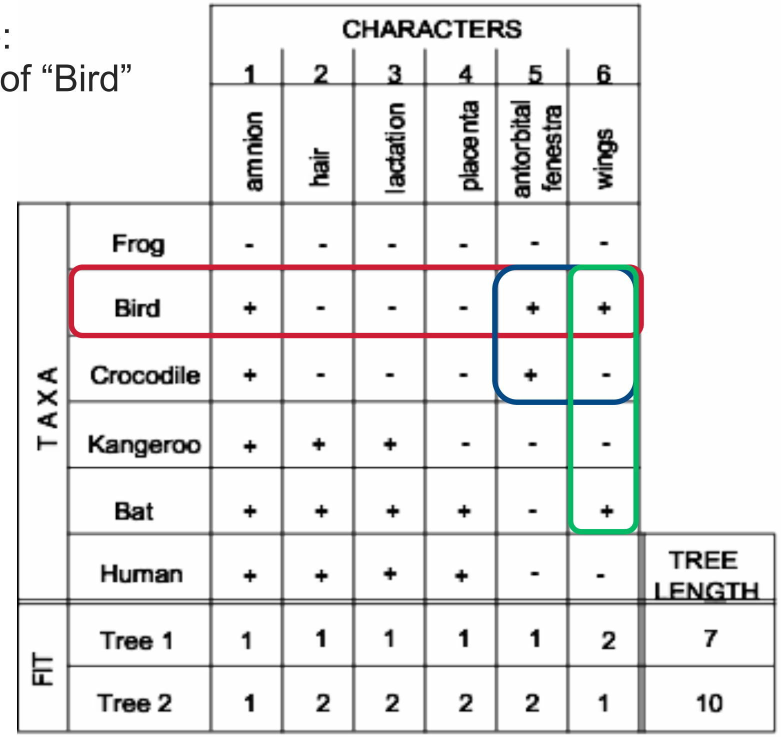
👎 method lacks accuracy (no correction for potential biases), precision, and there is no optimising criterion

Tree estimation – maximum parsimony

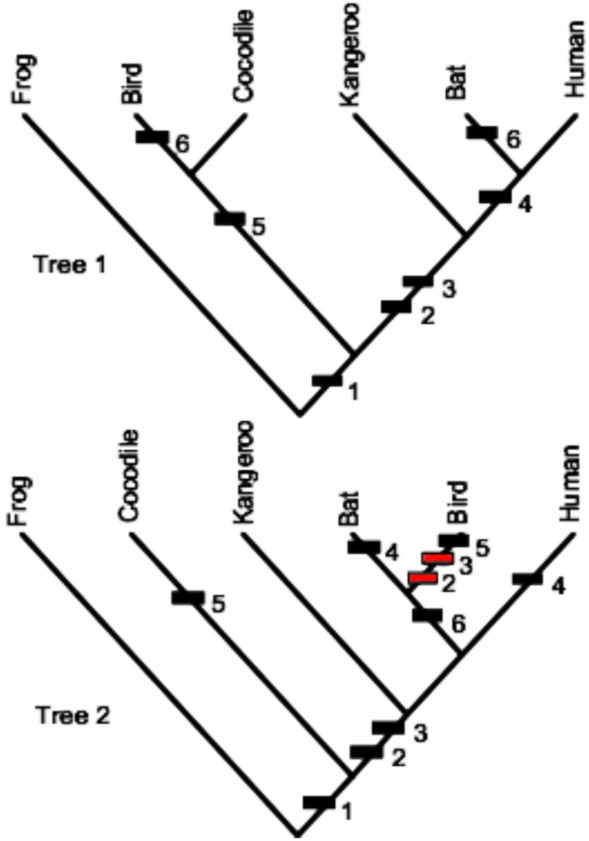
Method

- Evolution is the path of least resistance
- Every topology is valid, the quality is tested
  - Nearest neighbour interchange (NNI)
  - Also to calculate branch lengths
- The parsimoniest tree contains the least number of mutations

Example:
Position of "Bird"



		CHARACTERS						
		1	2	3	4	5	6	
		amnion	hair	lactation	placenta	antorbital fenestra	wings	
TAXA	Frog	-	-	-	-	-	-	
	Bird	+	-	-	-	+	+	
	Crocodile	+	-	-	-	+	-	
	Kangeroo	+	+	+	-	-	-	
	Bat	+	+	+	+	-	+	
	Human	+	+	+	+	-	-	TREE LENGTH
FIT	Tree 1	1	1	1	1	1	2	7
	Tree 2	1	2	2	2	2	1	10



Tree estimation – maximum likelihood

Method

- Each topology is valid
- Likelihood is the probability of the data given a specific model
- Models
  - Several substitution at the same position
  - Transition occurs more often than transversion (change in class of base)
  - Differences in conservation of particular sites
    - E.g. 3. position in a triplet codon
    - Within a gene for correct function

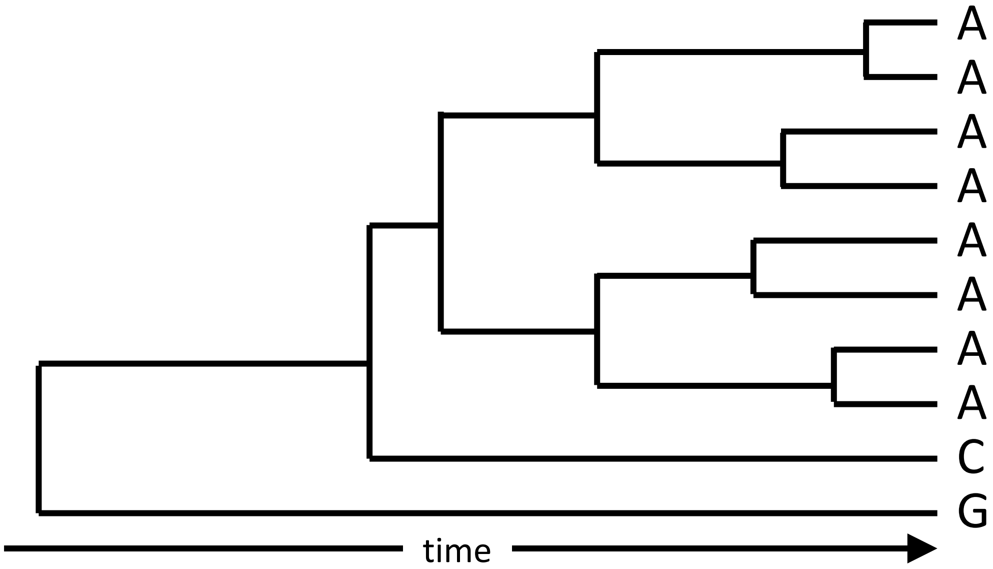
👍 Highly accurate (biological realism via substitution model)

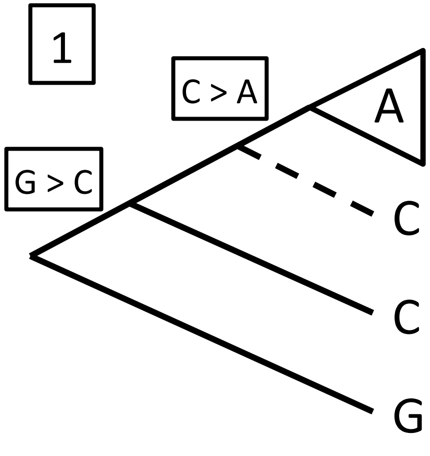
👍 Robust statistical context to evaluate specific hypotheses

👍 Single tree produced that is generally precise

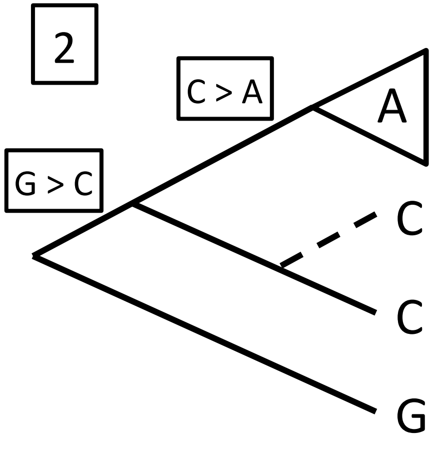
👎 Complexity of estimation process: slow & computationally demanding

Tree estimation – maximum parsimony vs maximum likelihood

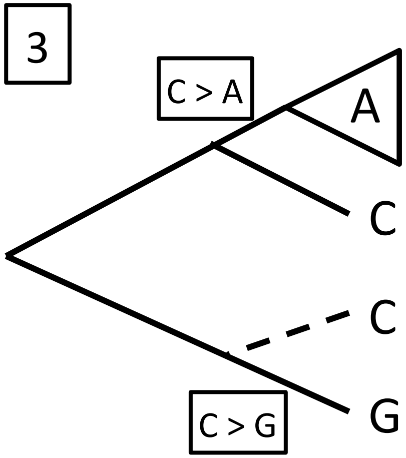




2 Mutations



2 Mutations



2 Mutations

- ➢ Maximum Parsimony will not find a solution
- ➢ Maximum Likelihood excludes 1 & 3 as the timescale is too short!

Bootstrapping

- **Bootstrapping is a way to produce a measure of confidence in the relationships found in a phylogenetic analysis**

- Characters (sites/amino acids) are resampled with replacement to produce a set of replicate data sets

- Each replicate is analysed (e.g. with parsimony/distance/maximum likelihood)

- Frequency of occurrence of groups in the results of these analyses is a measure of support for those groups

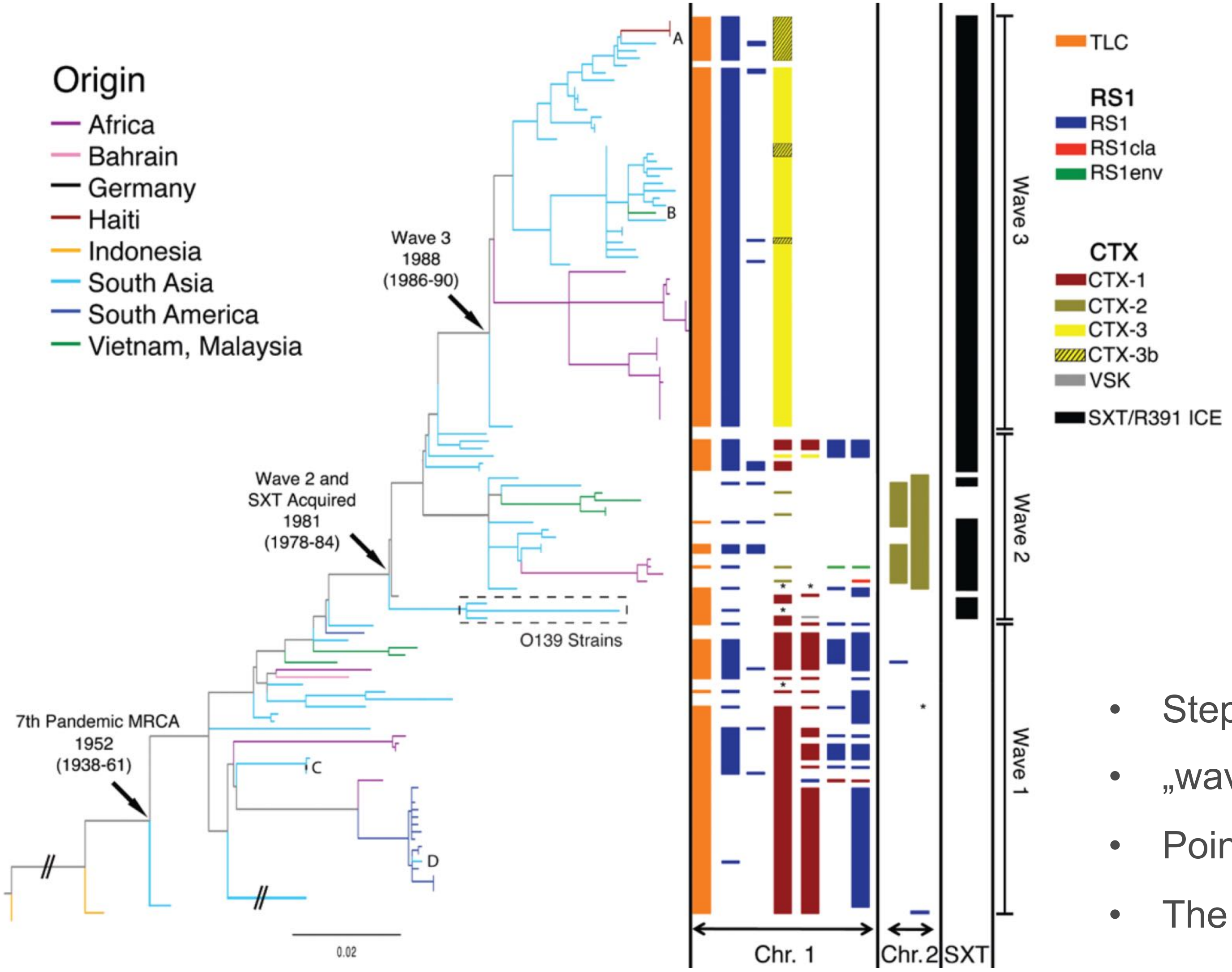
- Bootstrap proportions (BPs) are often represented as a number on each branch of a tree showing how often that relationships occurred in the replicate analyses

	characters								
Taxa	1	2	3	4	5	6	7	8	9
A	A	C	C	T	G	A	T	G	C
B	A	G	C	T	G	G	T	T	C
C	A	G	C	A	G	A	T	G	G
D	T	C	C	T	C	G	T	G	C
E	T	C	T	T	A	A	T	G	C

Random number generator: 9

	characters			
Taxa	2	5	9	2
A	C	G	C	C
B	G	G	C	G
C	G	G	G	G
D	C	C	C	C
E	C	A	C	C

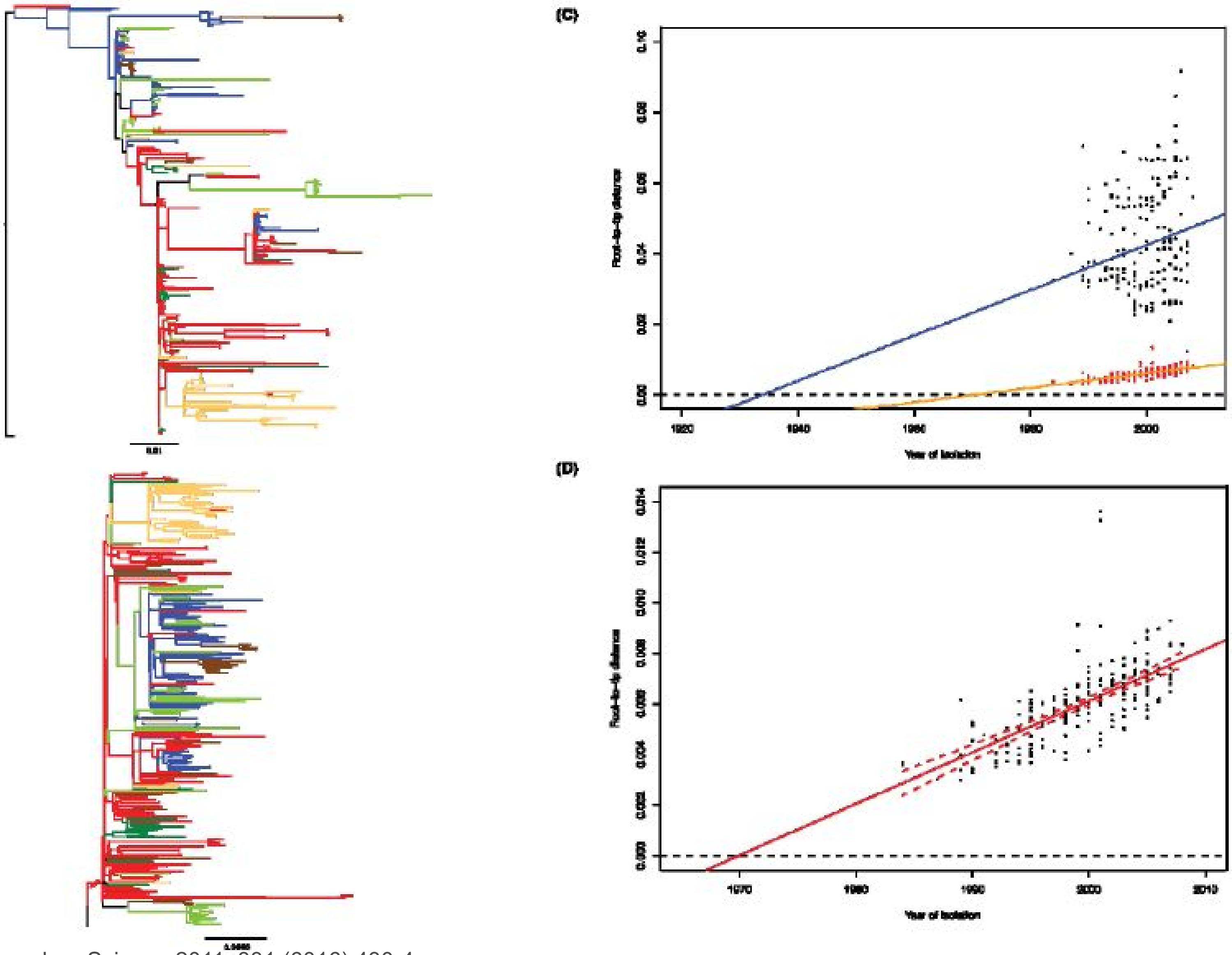
Examples of “tree gazing” – *Vibrio cholerae*



- Step-wise evolution over time
- „waves“
- Point-source trajectory
- The „dinosaur“

Mutreja, Nature 2011, 477(7365):462-5.

Examples of “tree gazing” – effect of recombination: *Streptococcus pneumoniae*



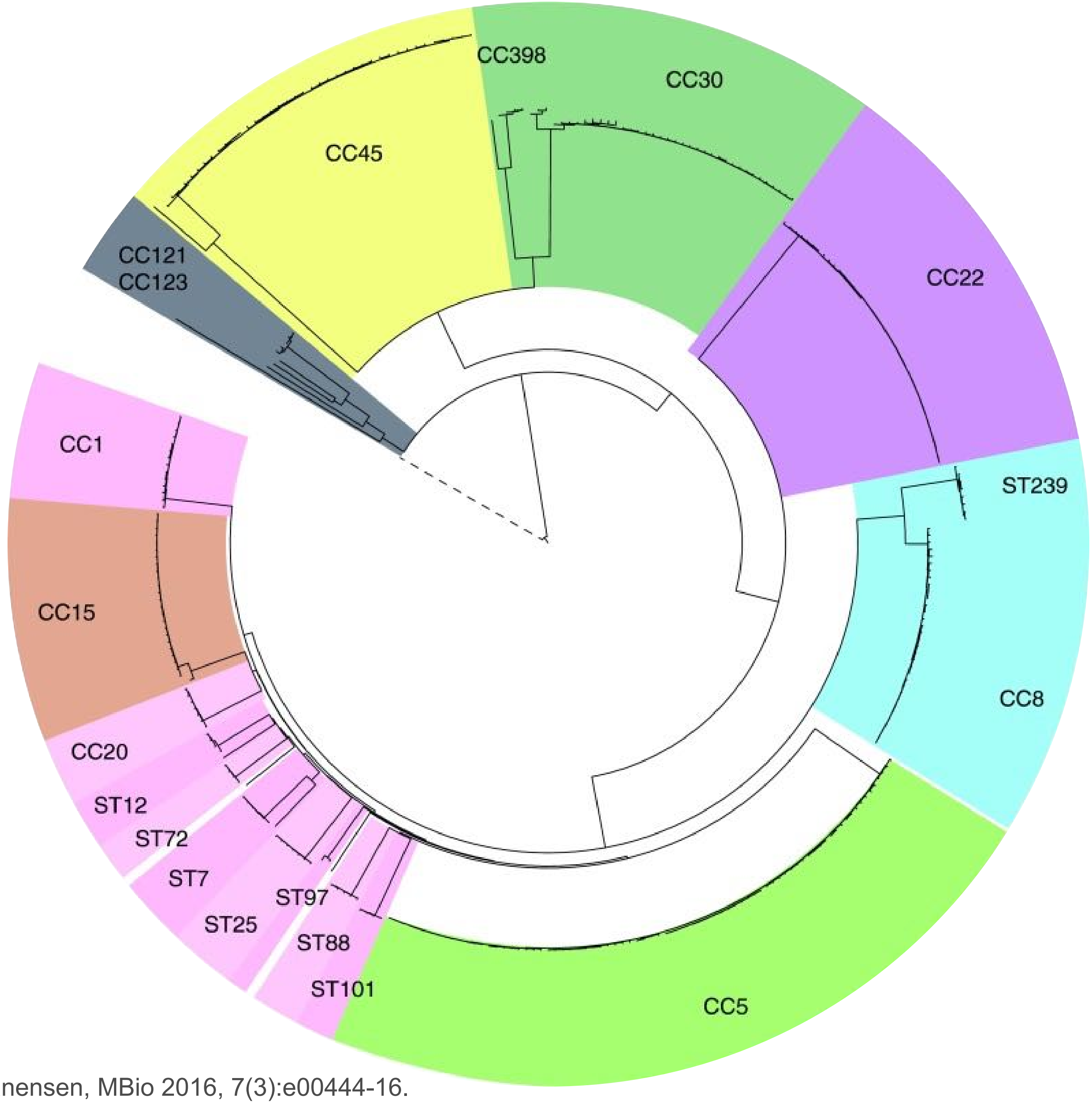
With all SNPs

SNPs not under recombination

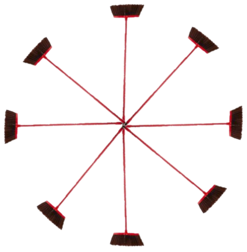
- Branch lengths
- Temporal signal

Croucher, Science 2011, 331 (6016):430-4.

Examples of “tree gazing” – *Staphylococcus aureus*



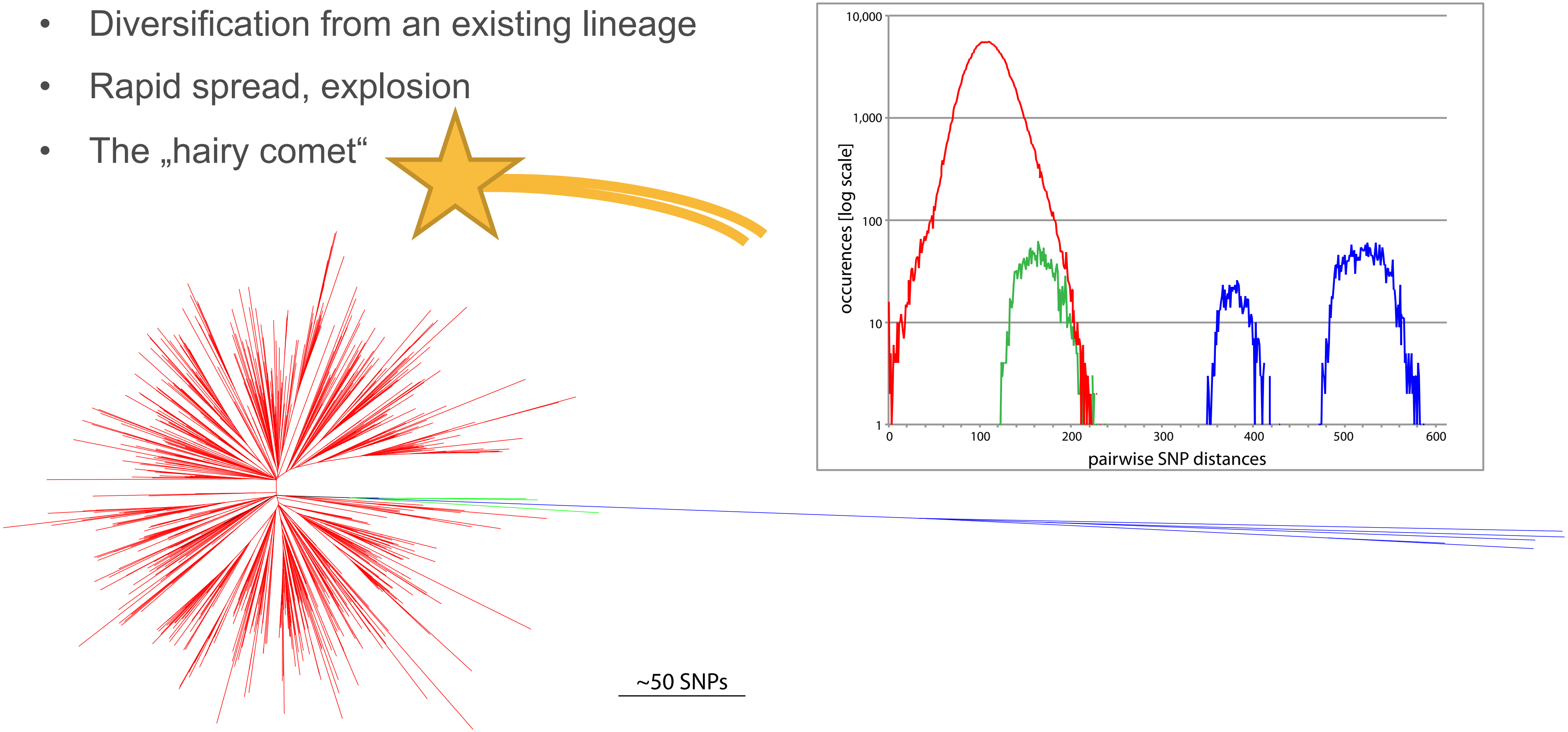
- Evolution over long time periods
- Distinct lineages
- The „broom sticks“



Aanensen, MBio 2016, 7(3):e00444-16.

Examples of “tree gazing” – *S. aureus*

- Diversification from an existing lineage
- Rapid spread, explosion
- The „hairy comet“

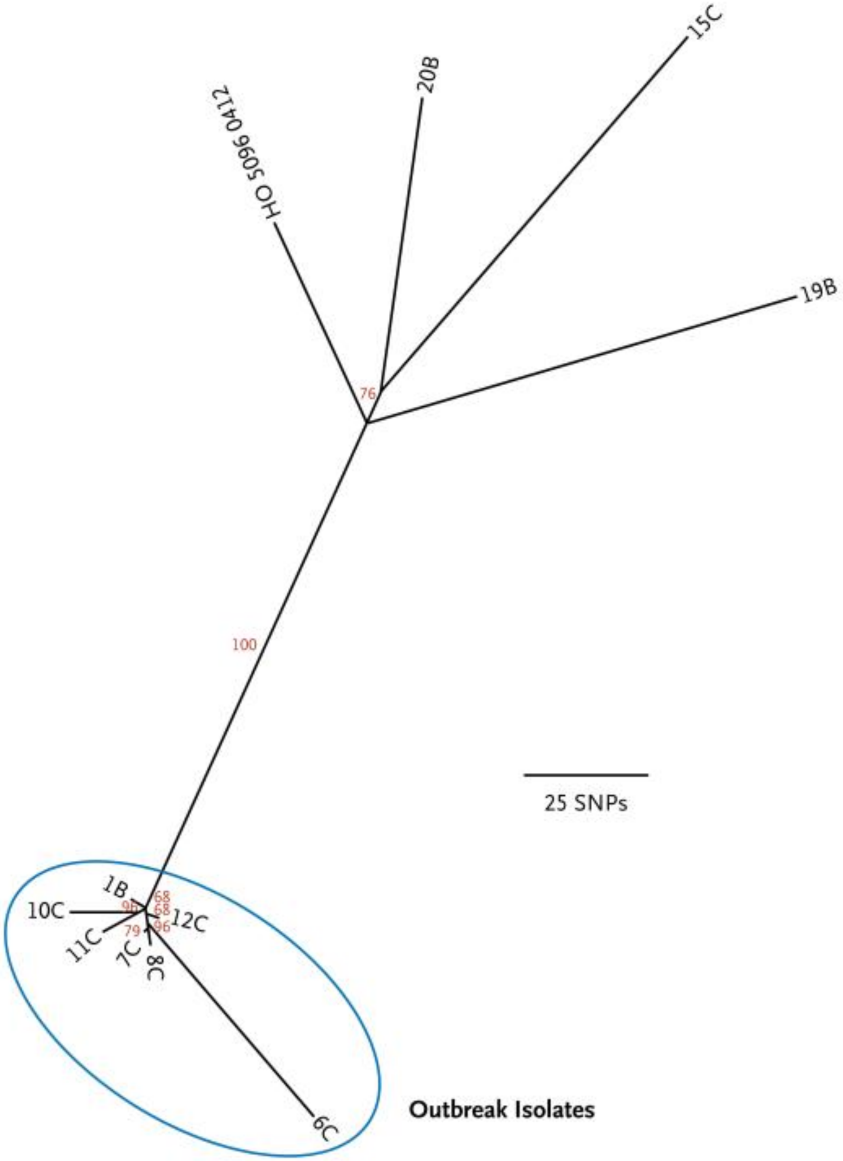


ST22-A2
“head”
“EMRSA-15”
- hospital-adapted

ST22-A1
“pre-head”
non-fluoroquinolone resistant

“the tail”
- community-acquired

Examples of “tree gazing” – *S. aureus*

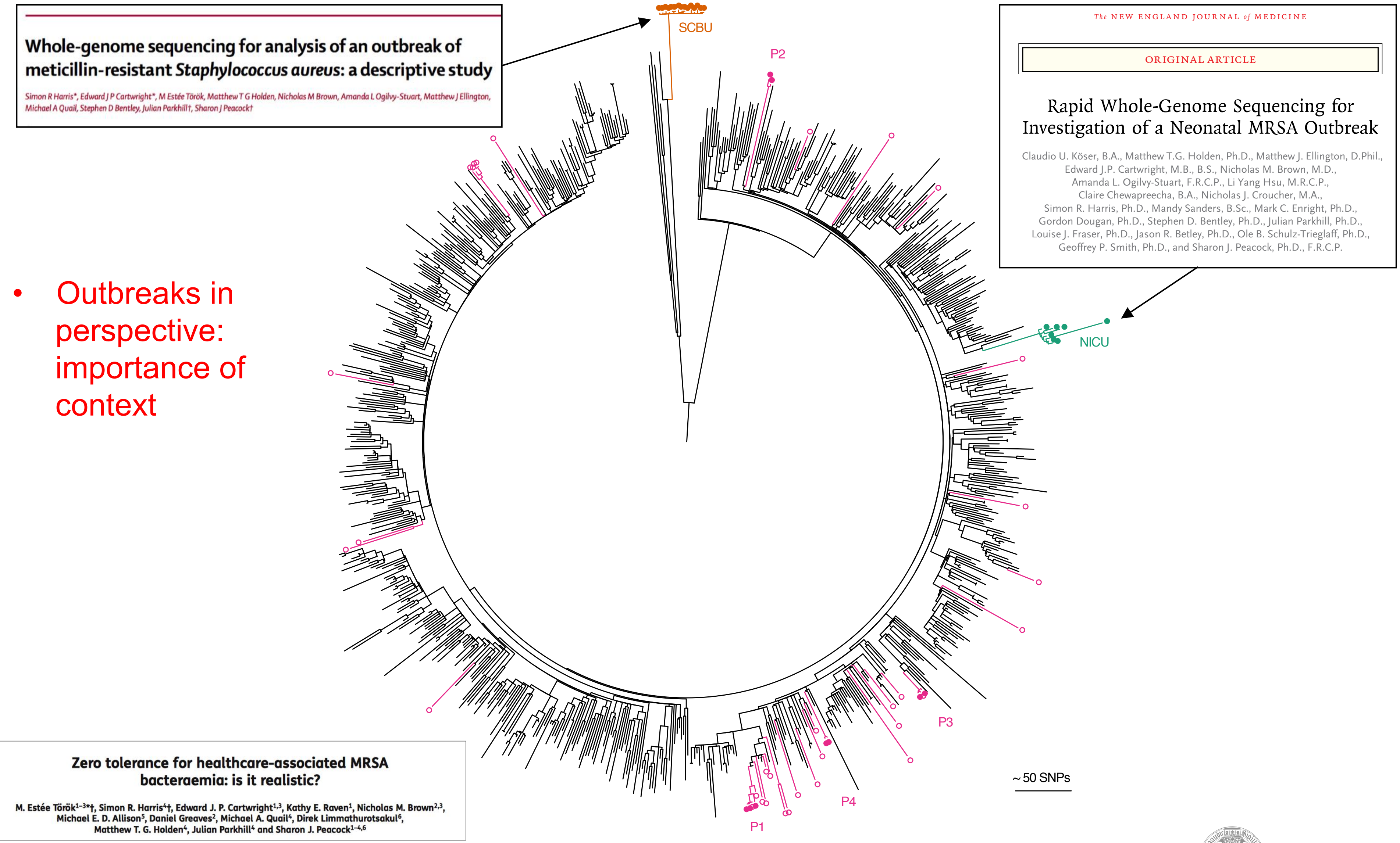


- Outbreak investigation
- Cluster with one isolate sticking out
  - ➢ Hypermutator phenotype
  - ➢ Accumulation of SNPs due to mutS/L mutation (inactivation of error checking)
  - ➢ Beware of absolute numbers of SNPs!

Köser, NEJM 2012, 366(24):2267-75.

Examples of “tree gazing” – *S. aureus*

- Outbreaks in perspective: importance of context



Whole-genome sequencing for analysis of an outbreak of meticillin-resistant *Staphylococcus aureus*: a descriptive study

Simon R Harris*, Edward J P Cartwright*, M Estée Török, Matthew T G Holden, Nicholas M Brown, Amanda L Ogilvy-Stuart, Matthew J Ellington, Michael A Quail, Stephen D Bentley, Julian Parkhill†, Sharon J Peacock†

The NEW ENGLAND JOURNAL *of* MEDICINE

ORIGINAL ARTICLE

Rapid Whole-Genome Sequencing for Investigation of a Neonatal MRSA Outbreak

Claudio U. Köser, B.A., Matthew T.G. Holden, Ph.D., Matthew J. Ellington, D.Phil., Edward J.P. Cartwright, M.B., B.S., Nicholas M. Brown, M.D., Amanda L. Ogilvy-Stuart, F.R.C.P., Li Yang Hsu, M.R.C.P., Claire Chewapreecha, B.A., Nicholas J. Croucher, M.A., Simon R. Harris, Ph.D., Mandy Sanders, B.Sc., Mark C. Enright, Ph.D., Gordon Dougan, Ph.D., Stephen D. Bentley, Ph.D., Julian Parkhill, Ph.D., Louise J. Fraser, Ph.D., Jason R. Betley, Ph.D., Ole B. Schulz-Trieglaff, Ph.D., Geoffrey P. Smith, Ph.D., and Sharon J. Peacock, Ph.D., F.R.C.P.

Zero tolerance for healthcare-associated MRSA bacteraemia: is it realistic?

M. Estée Török[1–3]*†, Simon R. Harris[4]†, Edward J. P. Cartwright[1,3], Kathy E. Raven[1], Nicholas M. Brown[2,3], Michael E. D. Allison[5], Daniel Greaves[2], Michael A. Quail[4], Direk Limmathurotsakul[6], Matthew T. G. Holden[4], Julian Parkhill[4] and Sharon J. Peacock[1–4,6]

Reuter, et al (2016), Genome Res 26(2): 263-270