Bacterial phylogenomics & phylodynamics

Module 3460: Pathogen Genomics

Dr Zoe Anne Dyson
Assistant Professor
Department of Infection Biology
London School of Hygiene & Tropical Medicine
zoe.dyson@lshtm.ac.uk

Intended learning outcomes

1. Recognise the basic principles of phylogenetics
2. Interpret data on a phylogenetic tree
3. Explain the methods used to infer a phylogenetic tree from bacterial pathogen whole genome sequencing data
4. Explain core concepts related to phylodynamics and how these can provide insights into pathogen evolution and epidemiology

Intended learning outcomes

1. Recognise the basic principles of phylogenetics
2. Interpret data on a phylogenetic tree
3. Explain the methods used to infer a phylogenetic tree from bacterial pathogen whole genome sequencing data
4. Explain core concepts related to phylodynamics and how these can provide insights into pathogen evolution and epidemiology

Class schedule

Time	Activity
14:00-14:50	Lecture - how to read and interpret a phylogenetic tree - recombination filtering in phylogenetic analysis
14:50-15:00	Break
15:00-16:00	Lecture - methods of phylogenetic tree inference - phylodynamic methods and applications

Join at menti.com | use code 8692 7095

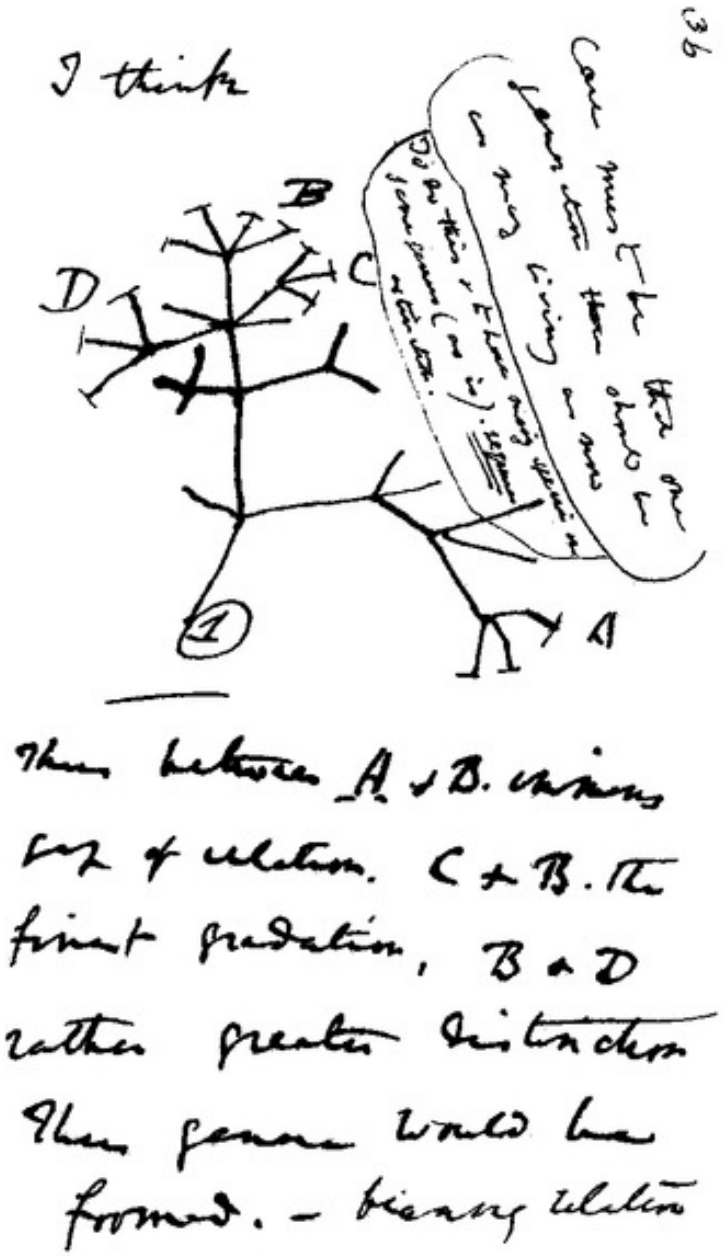
Mentimeter

What do you think of when you hear the terms phylogenomics?

What is phylogenomics?

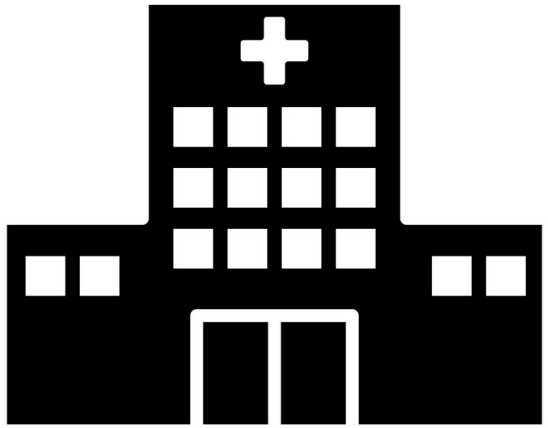
"The study of the genetic relationships between different biological entities"

- In **molecular phylogenetics** we use molecular data from biological entities such as **nucleotide** or **amino acid** sequences to infer evolutionary relationships
- This lecture will focus on using nucleotide sequences derived from **whole genome sequence (WGS)** data of **bacterial pathogens**



Charles Darwin, 1837

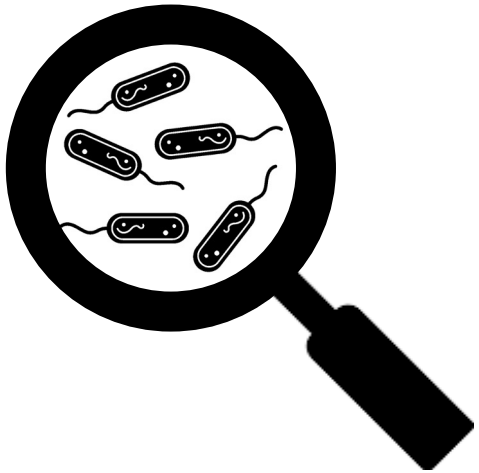
Applications of bacterial phylogenomics



Transmission & outbreak investigation

Pathogen surveillance

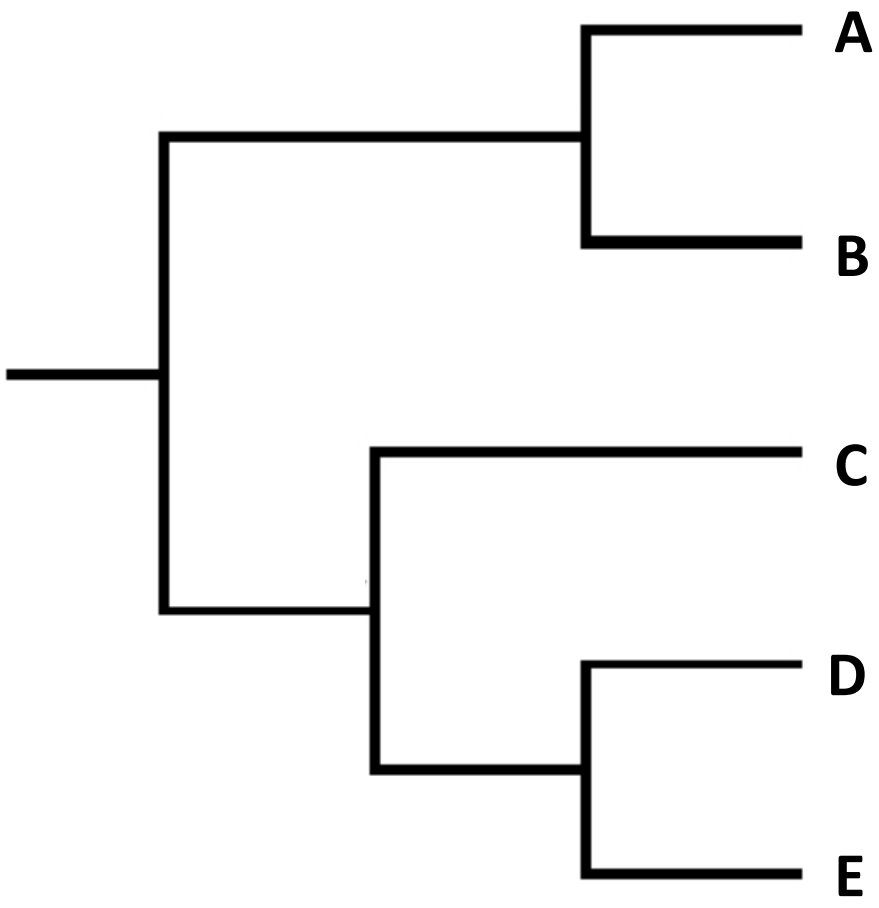
Intervention targeting & drug development



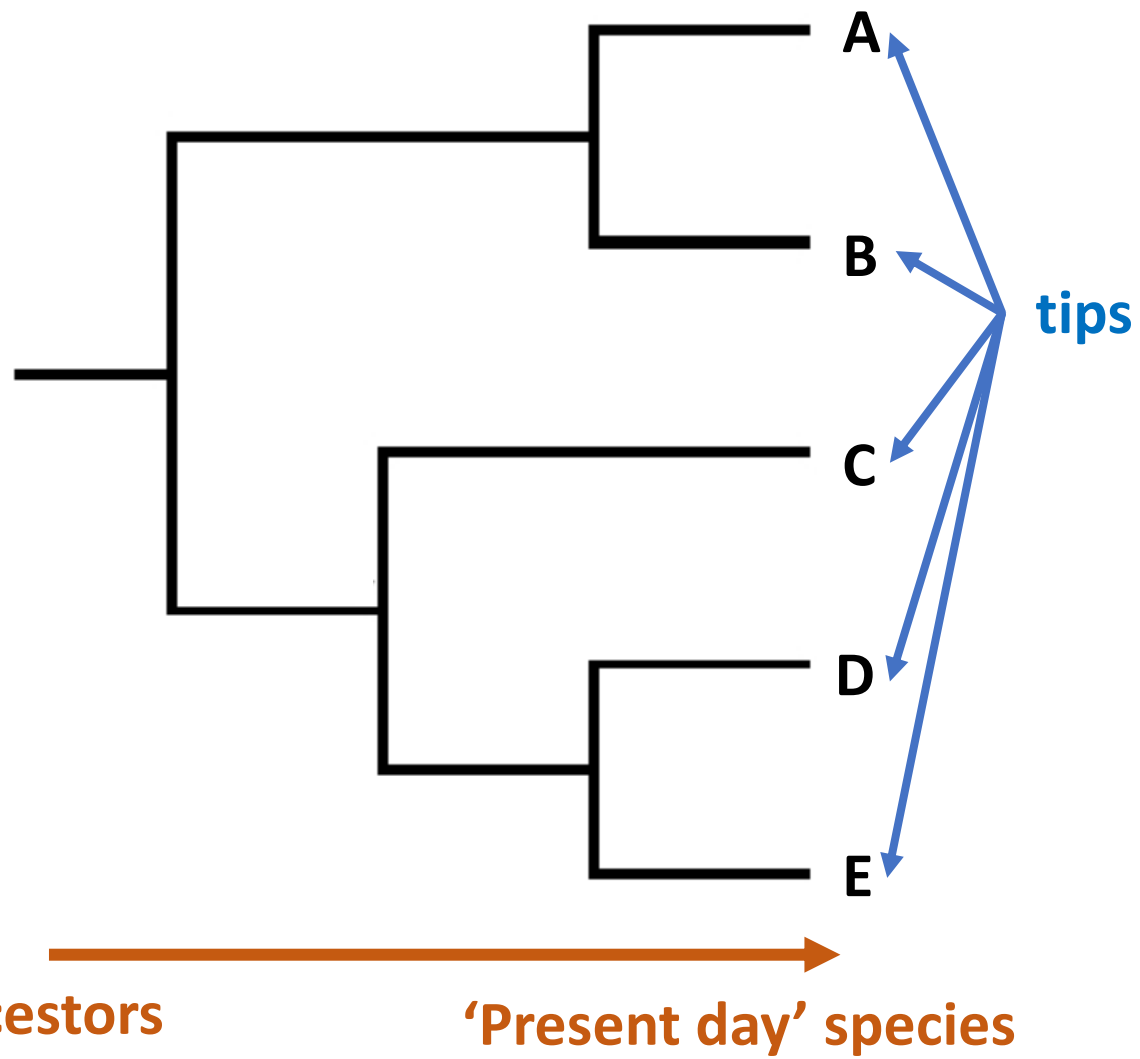
Characterization

And many more!!!

Central to phylogenomics is the phylogenetic tree



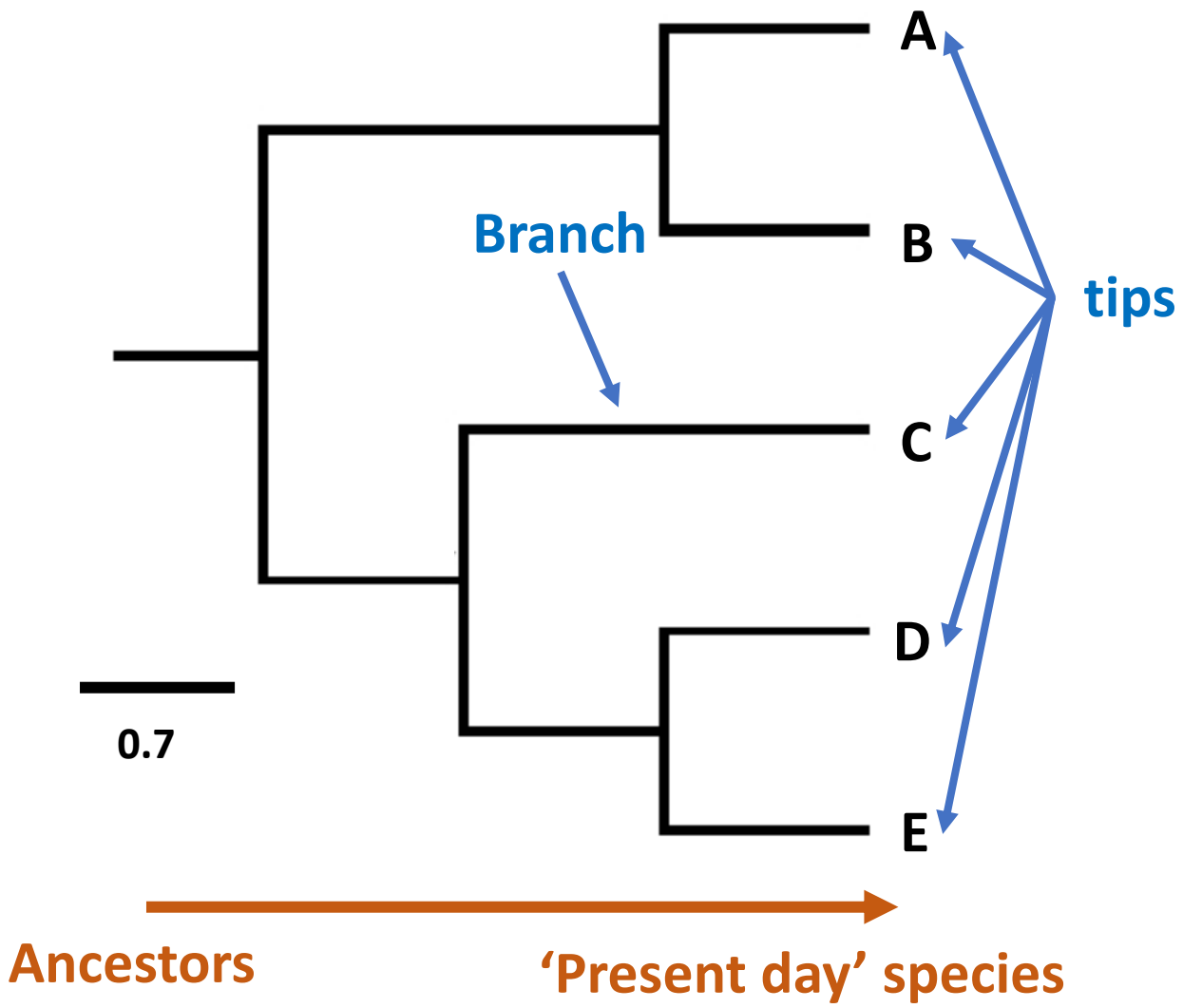
How to read a phylogenetic tree



- A graph showing **evolutionary relationships** between taxa/leaves/tips/external nodes (in this case between known sequences **A,B,C,D,E**)

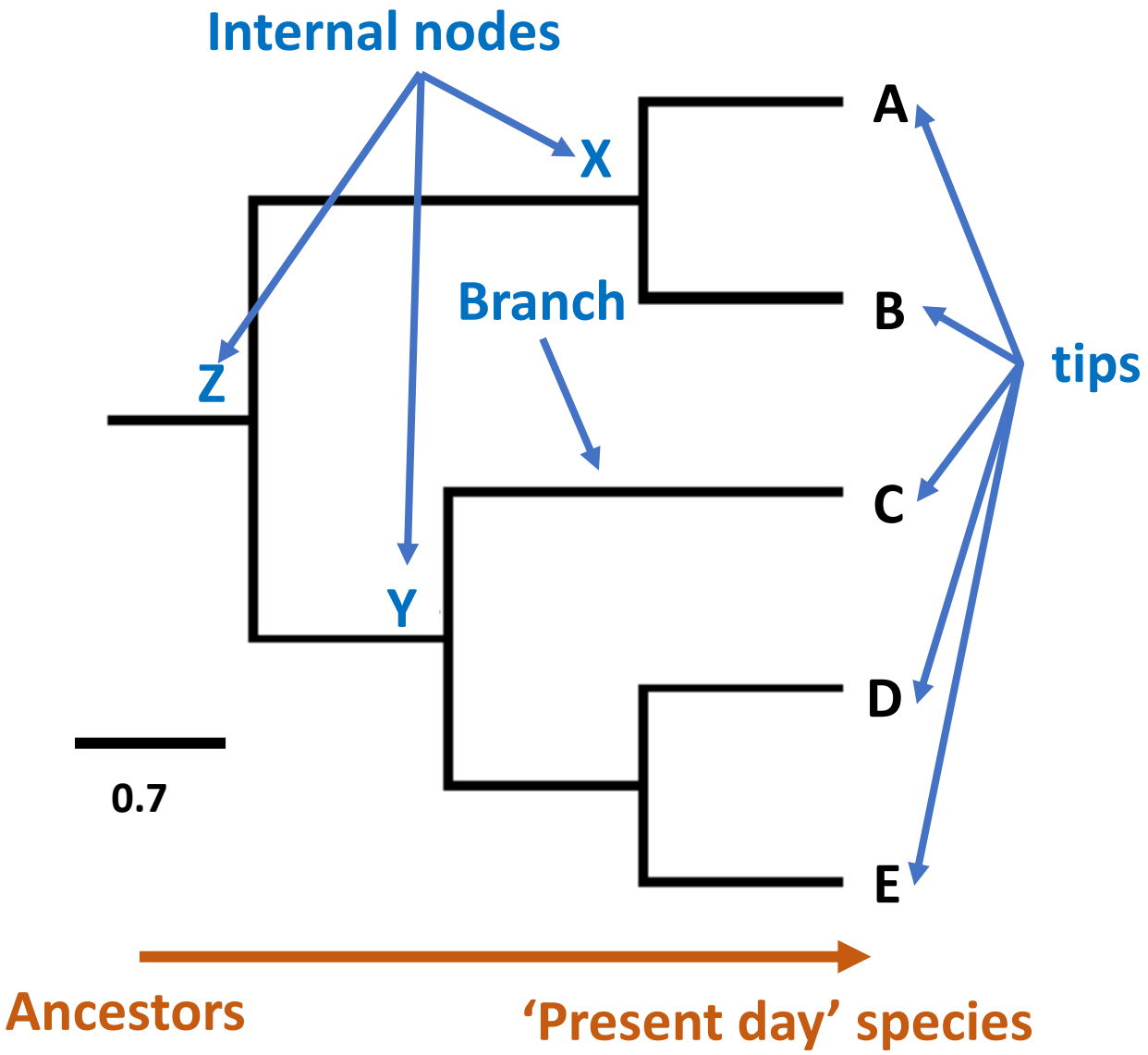
!

How to read a phylogenetic tree



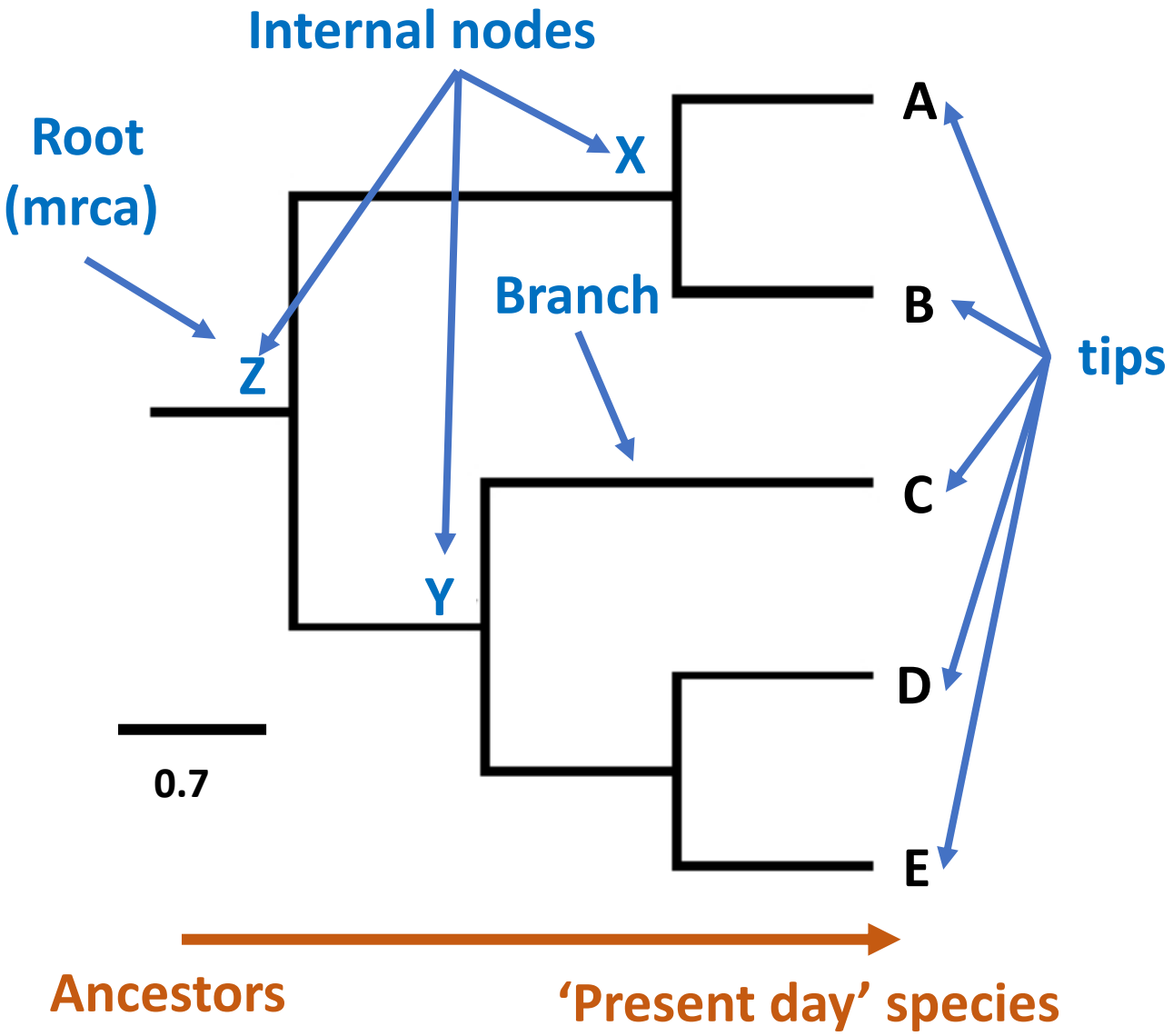
- A graph showing **evolutionary relationships** between taxa/leaves/tips/external nodes (in this case between known sequences **A,B,C,D,E**)
- **Branches** (horizontal edges) indicate genetic distance i.e. substitutions per variable site. Branches can also represent time in a dated Bayesian phylogeny
- Vertical edges – connections between taxa – nothing more!

How to read a phylogenetic tree



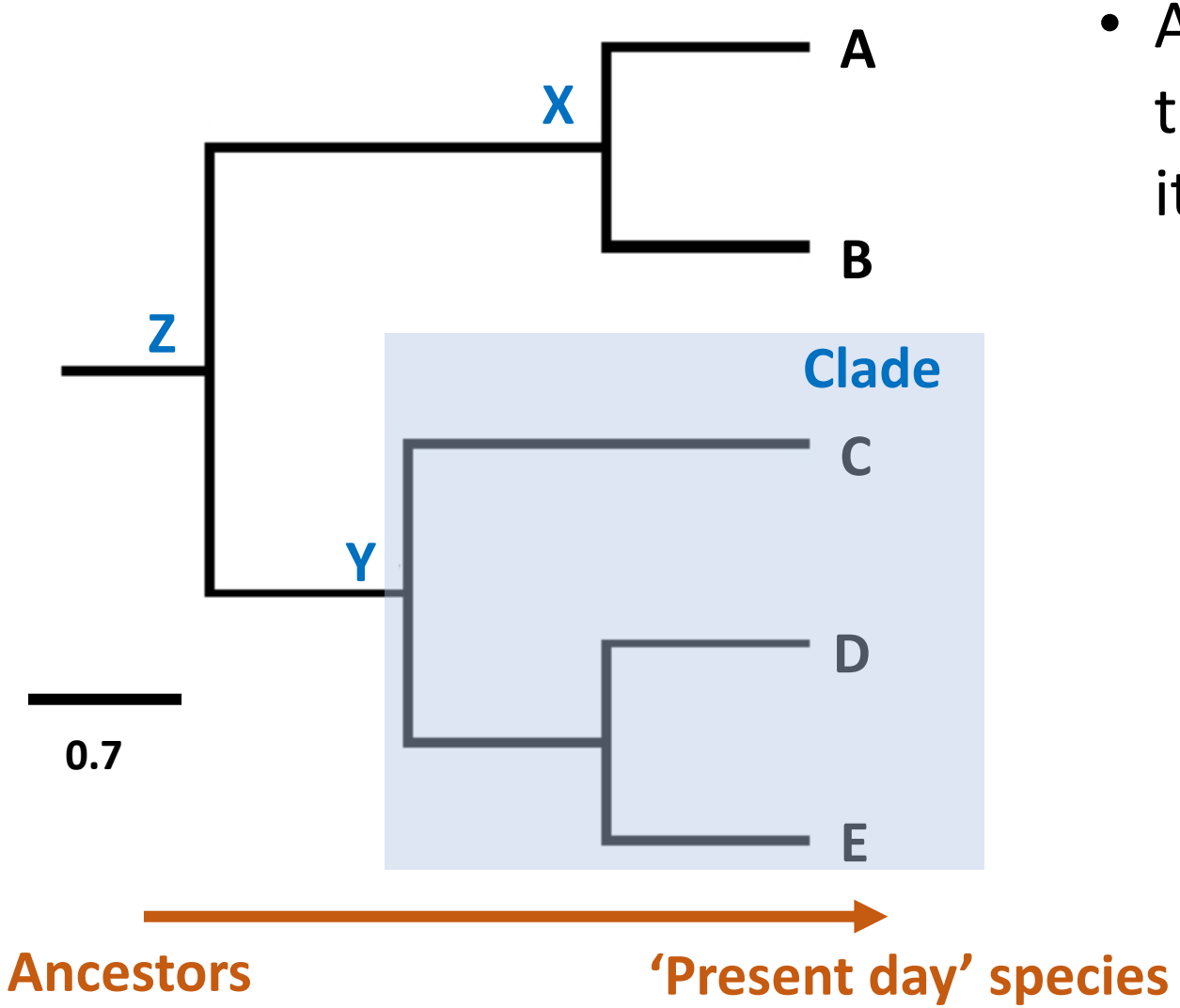
- A graph showing **evolutionary relationships** between taxa/leaves/tips/external nodes (in this case between known sequences **A,B,C,D,E**)
- **Branches** (horizontal edges) indicate genetic distance i.e. substitutions per variable site. Branches can also represent time in a dated Bayesian phylogeny
- Vertical edges – connections between taxa – nothing more!
- **Internal nodes** of a tree (e.g. **X, Y, Z**) represent shared/common ancestors e.g. **A & B** share an ancestor **X** from which they have **descended**. We do not know the internal node sequences

How to read a phylogenetic tree



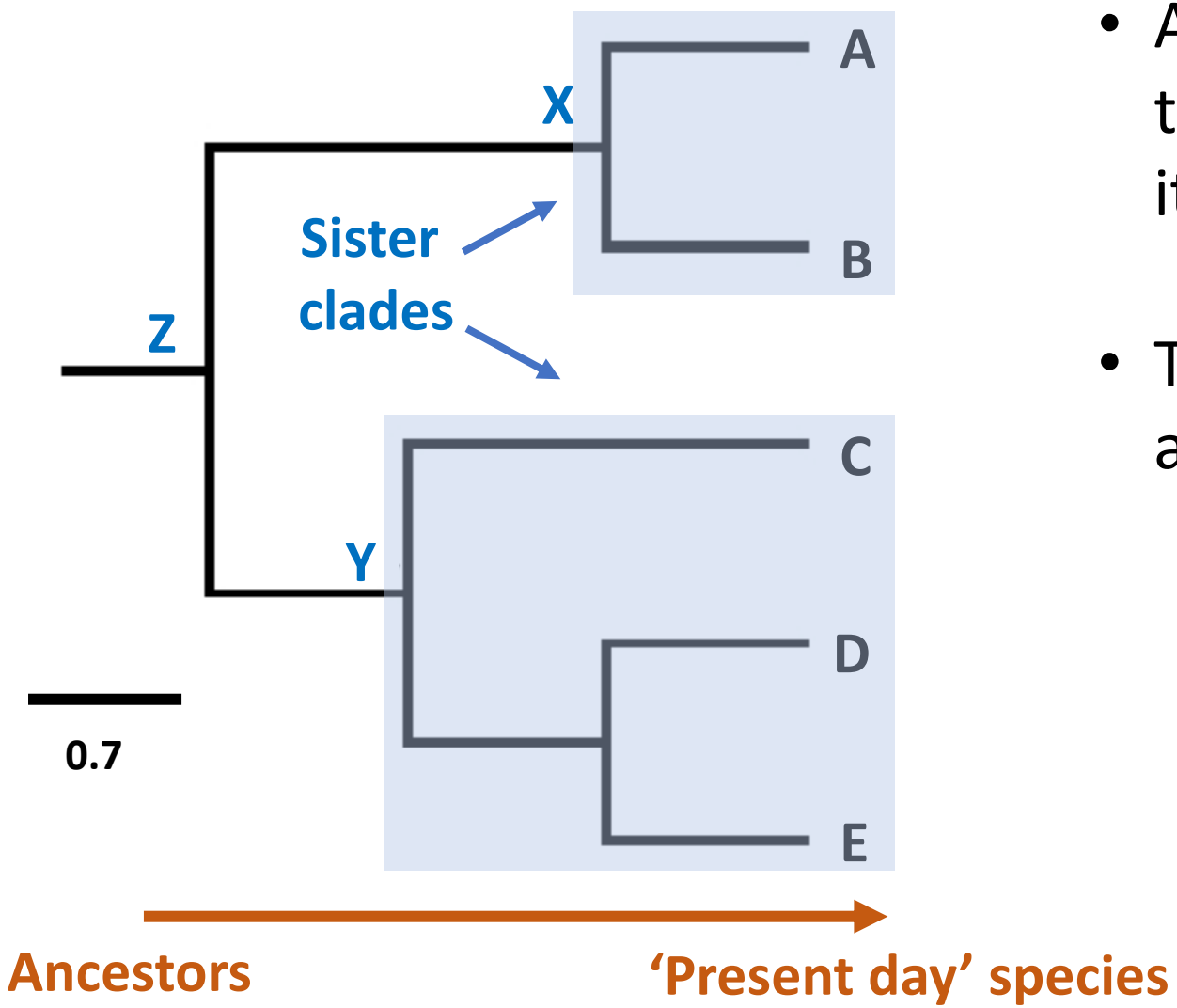
- A graph showing **evolutionary relationships** between taxa/leaves/tips/external nodes (in this case between known sequences **A,B,C,D,E**)
- **Branches** (horizontal edges) indicate genetic distance i.e. substitutions per variable site. Branches can also represent time in a dated Bayesian phylogeny
- Vertical edges – connections between taxa – nothing more!
- **Internal nodes** of a tree (e.g. **X, Y, Z**) represent shared/common ancestors e.g. **A & B** share an ancestor **X** from which they have **descended**. We do not know the internal node sequences
- All nodes (internal and external) share a **common ancestor** at the **root** of the tree (e.g. **Z**)
- **Z** is the **most recent common ancestor** (**mrca**) of all other nodes in the tree, from which they have descended

How to read a phylogenetic tree



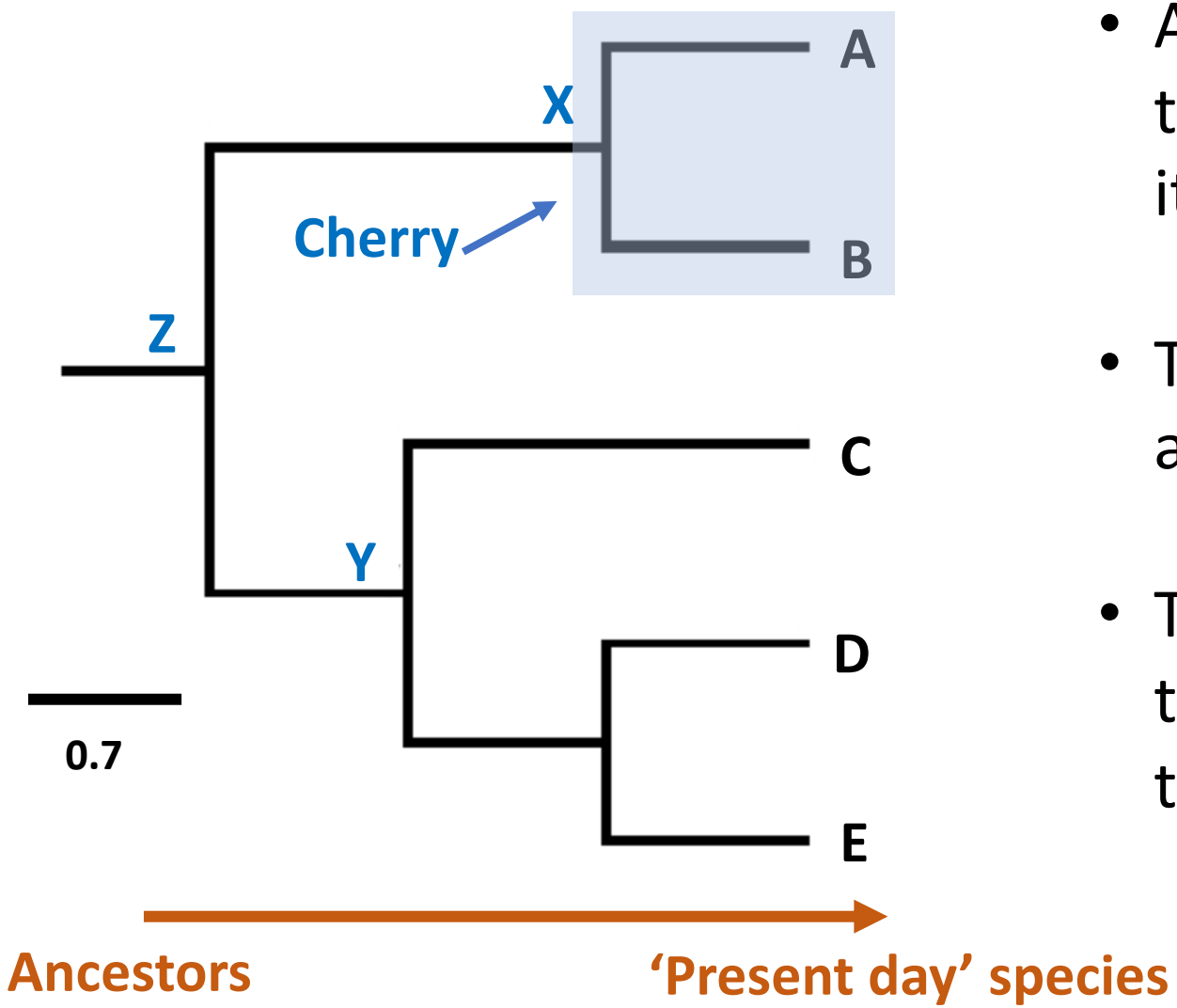
- A **clade** (or lineage) is a section of the tree consisting of an internal node and its **descendants**

How to read a phylogenetic tree



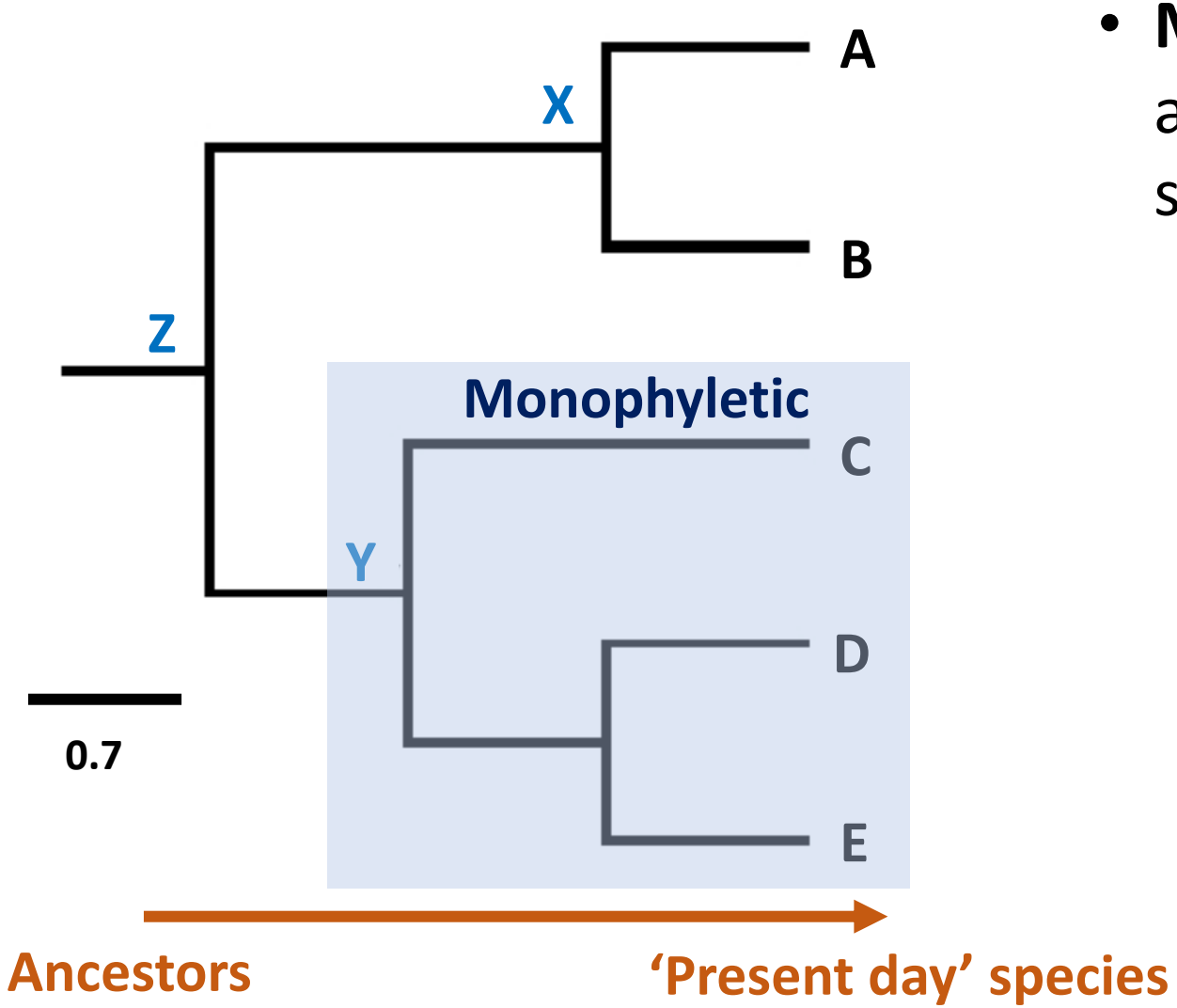
- A **clade** (or lineage) is a section of the tree consisting of an internal node and its **descendants**
- Two clades descending from the same ancestor are **sister clades**

How to read a phylogenetic tree



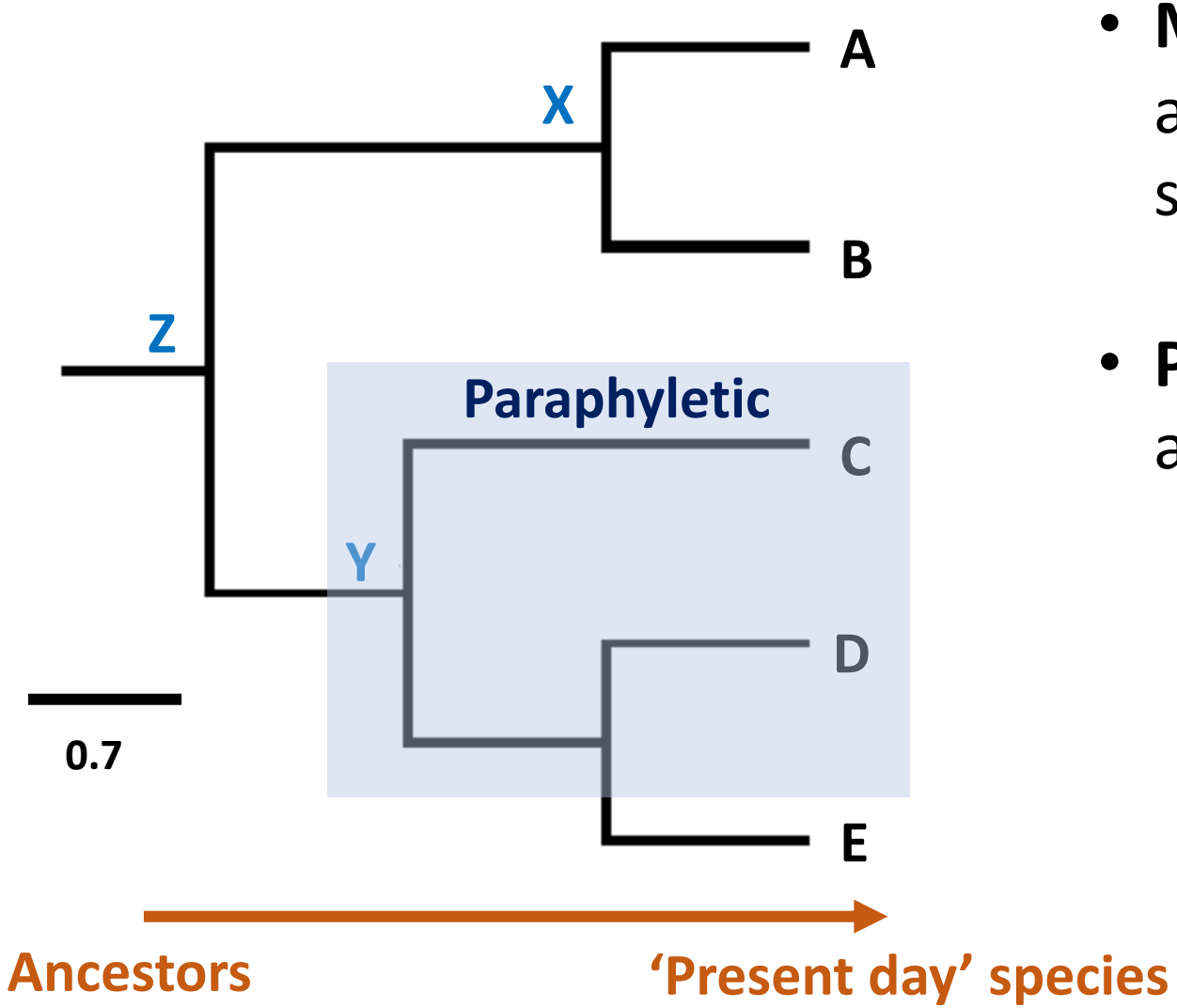
- A **clade** (or lineage) is a section of the tree consisting of an internal node and its **descendants**
- Two clades descending from the same ancestor are **sister clades**
- Two tips sharing an ancestor (adjacent tips on a tree) are sometimes referred to as **cherries**

How to read a phylogenetic tree



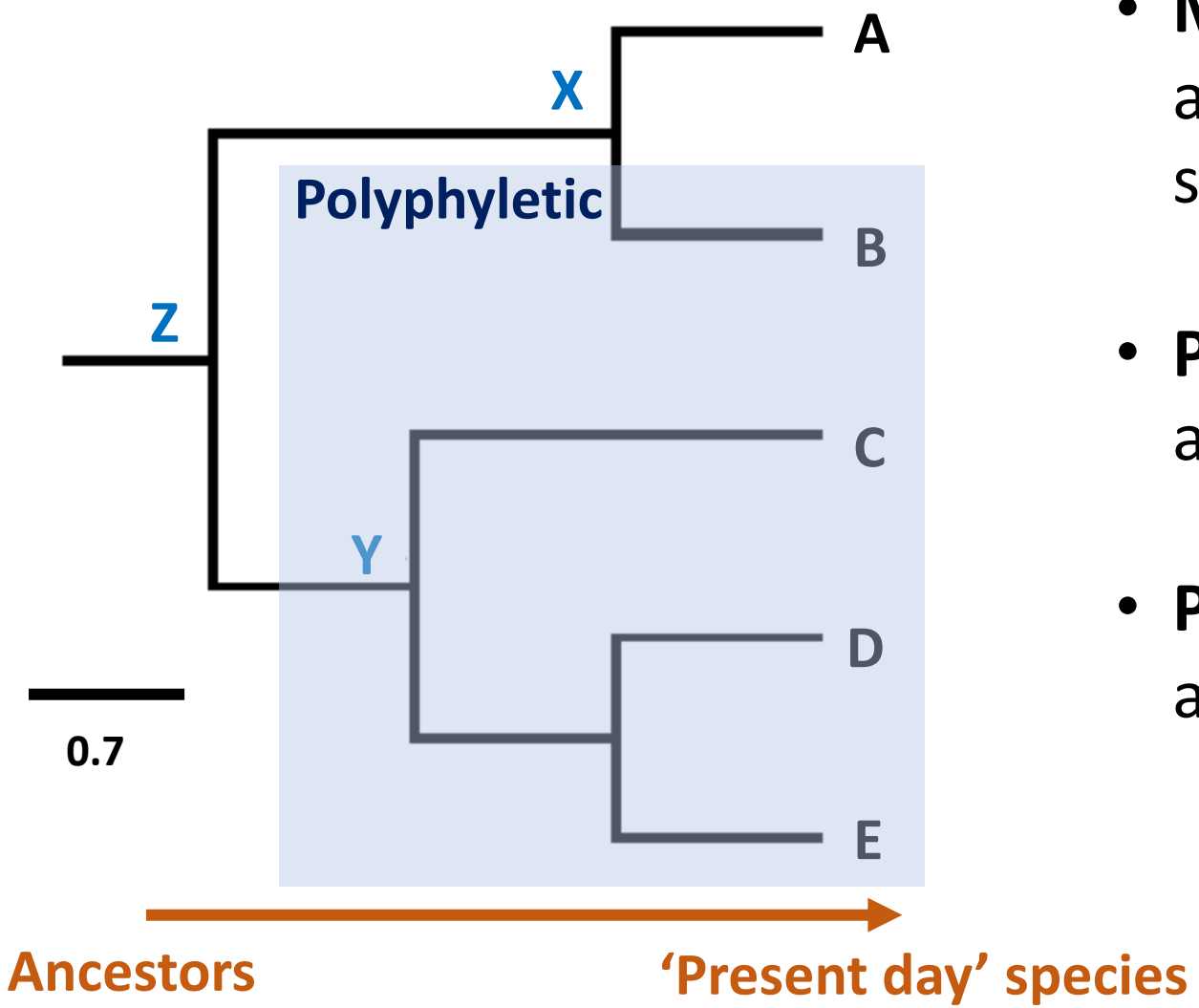
- **Monophyletic groups** include the ancestor and descendants of all sequences

How to read a phylogenetic tree

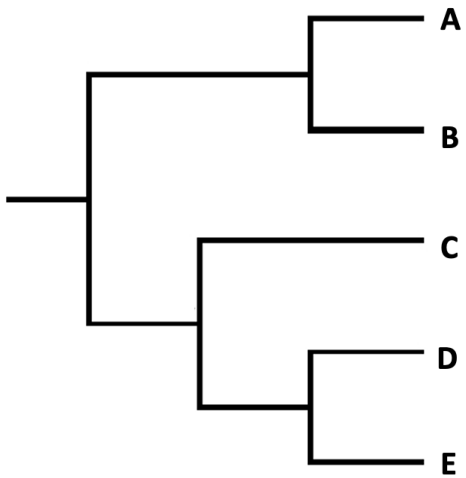


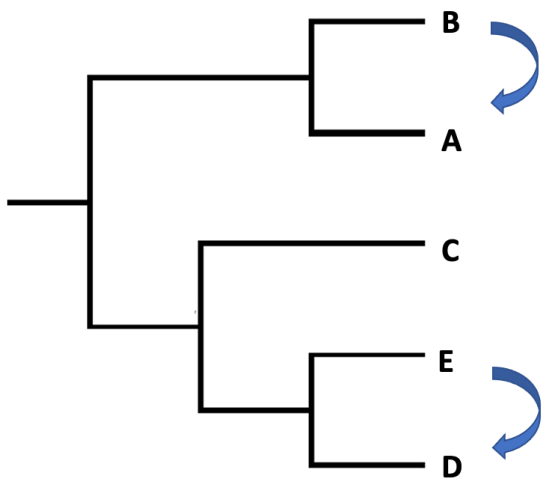
- **Monophyletic groups** include the ancestor and descendants of all sequences
- **Paraphyletic groups** include the ancestor but not all descendants

How to read a phylogenetic tree



- **Monophyletic groups** include the ancestor and descendants of all sequences
- **Paraphyletic groups** include the ancestor but not all descendants
- **Polyphyletic groups** do not include the ancestor of all sequences





Join at menti.com | use code 8692 7095

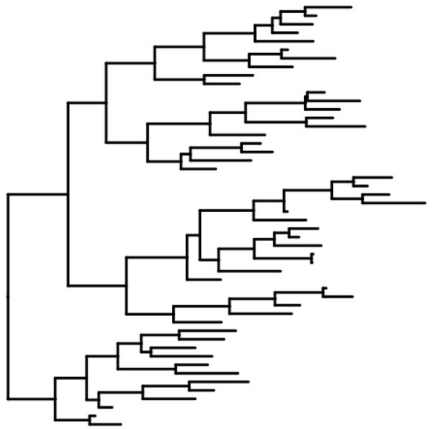
Mentimeter

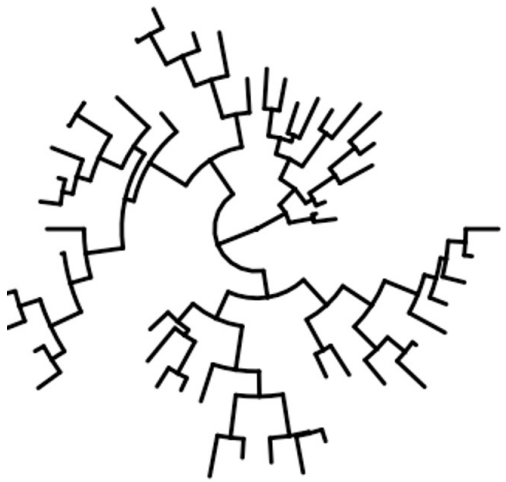
Are these the same tree?

Yes

No

Not sure





Join at menti.com | use code 8692 7095

Mentimeter

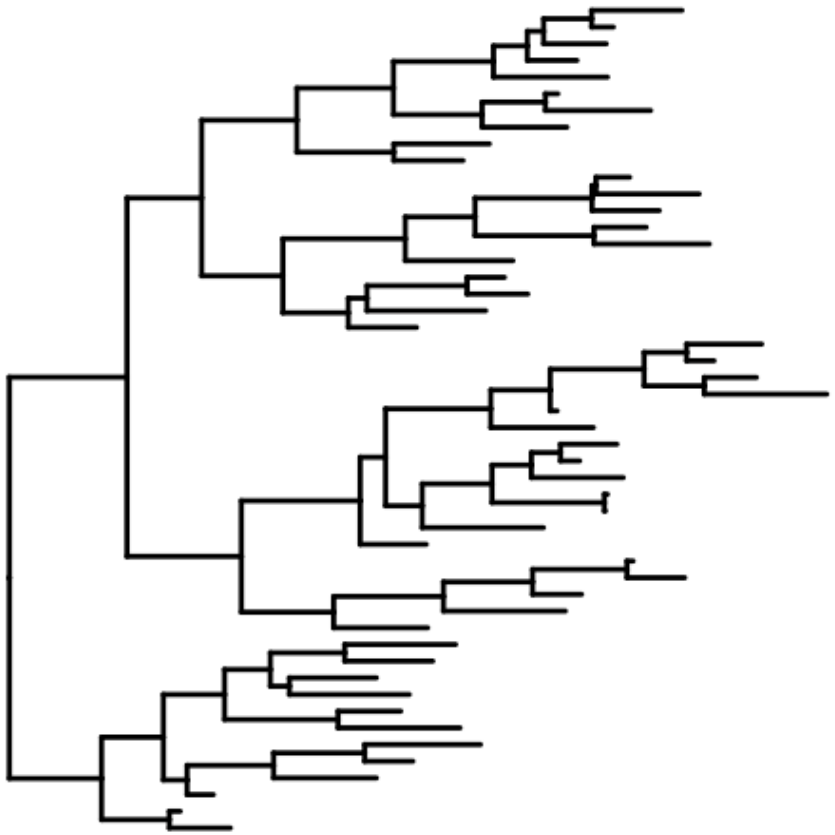
Could these be the same tree?

Yes

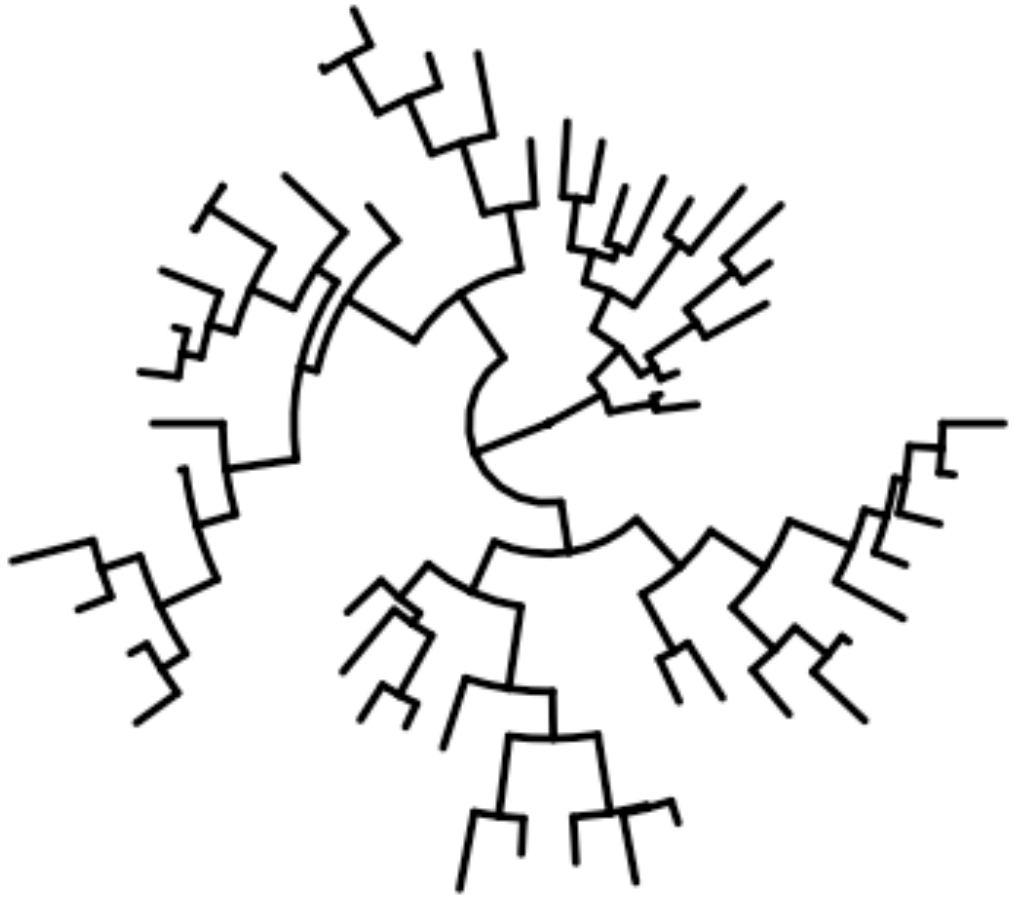
No

Not sure

Could these be the same tree?



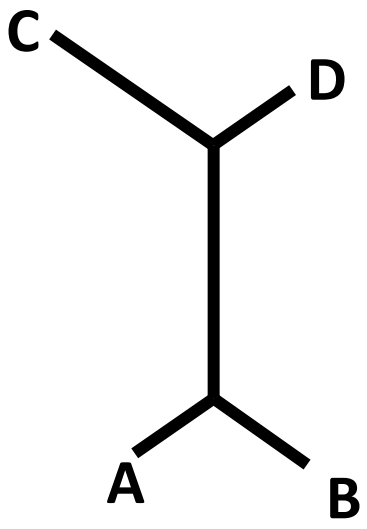
Rectangular



Circular/fan

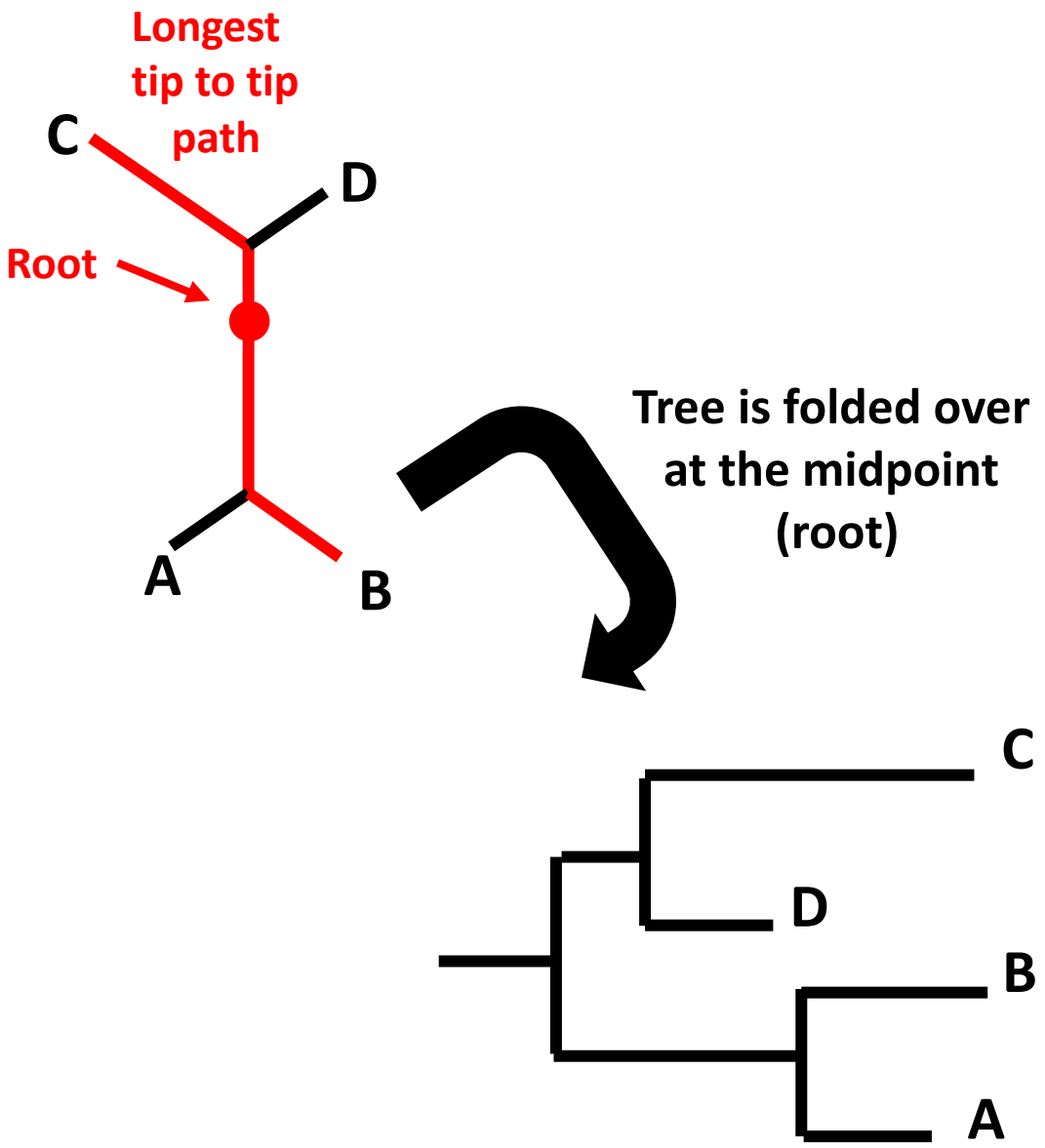
https://guangchuangyu.github.io/

Rooted & unrooted trees



- An **unrooted tree** shows relatedness of taxa without making assumptions about ancestry

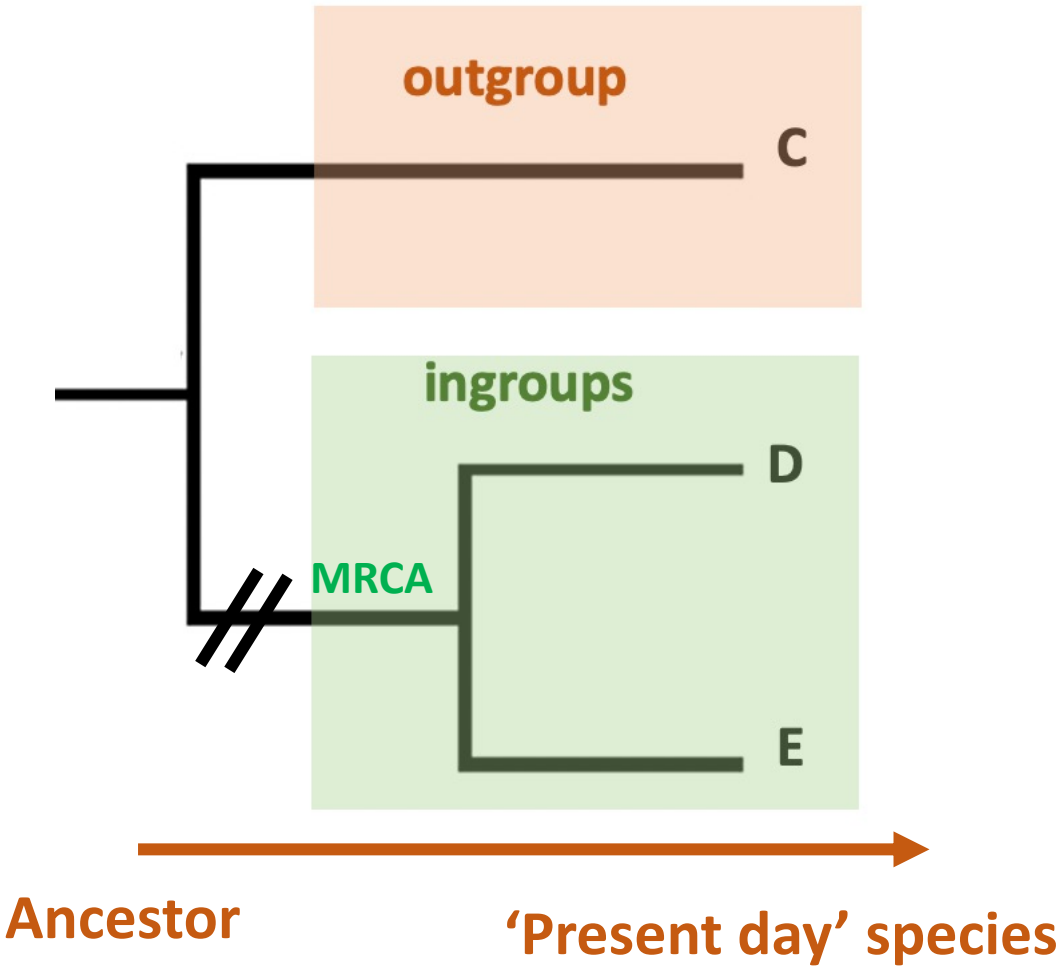
Rooted & unrooted trees



- An **unrooted tree** shows relatedness of taxa without making assumptions about ancestry

- A tree may be **midpoint rooted.** For this, the root is the midpoint between the two most distant taxa. This midpoint usually represents the ancestor

Rooted & unrooted trees



- An **unrooted tree** shows relatedness of taxa without making assumptions about ancestry

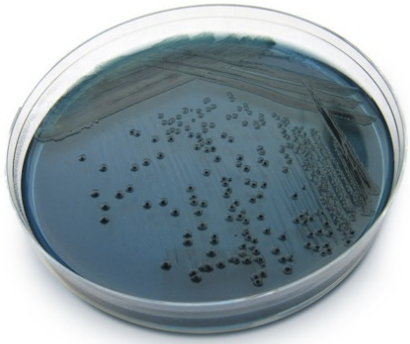
- A tree may be **midpoint rooted.** For this, the root is the midpoint between the two most distant taxa. This midpoint usually represents the ancestor

- In an **outgroup rooted tree** an **outgroup** (a more distantly related taxa) is included to root the tree. This determines the MRCA of the **ingroup** taxa, and gives the direction of evolution e.g. **C** would be an outgroup to **D & E.** The outgroup is normally removed from the tree before visualisation

Intended learning outcomes

1. Recognise the basic principles of phylogenetics
2. Interpret data on a phylogenetic tree
3. Explain the methods used to infer a phylogenetic tree from bacterial pathogen whole genome sequencing data
4. Explain core concepts related to phylodynamics and how these can provide insights into pathogen evolution and epidemiology

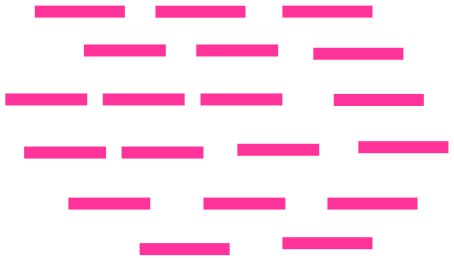
Revision: bacterial pathogen sequencing



Isolate & purify bacteria

Extract DNA & prepare library

Whole genome sequencing (Illumina short read)



Short read data

Technical limitations of short read sequencing

Genome

Short reads

Reconstructed genome

Phylogenetic inference: common workflows

Generate a core genome alignment

Option 1: Mapping-based

1. Map reads to reference
2. Call variants (SNPs/SNVs)
3. Filter recombination

Option 2: Assembly-based

1. Assemble genomes
2. Annotate genes
3. Infer pangenome

4. **Infer phylogenetic tree**

Phylogenetic inference: common workflows

Generate a core genome alignment

Option 1: Mapping-based

1. Map reads to reference
2. Call variants (SNPs/SNVs)
3. Filter recombination

4. **Infer phylogenetic tree**

Revision: Mapping

Mapping *is the process of aligning raw read data to a known* ***reference sequence. Bowtie2*** *is an example of a software tool for mapping reads to a reference sequence. A reference sequence is usually a high quality completed genome sequence that is somewhat closely related to the sequences analysed.*

Revision: Variant (SNV/SNP) calling

Variant calling *is the process of calling high confidence single base changes in the genome sequence referred to as* ***SNVs (Single Nucleotide Variants)*** *which are also often called* ***Single Nucleotide Polymorphisms (SNPs)****.* ***SamTools*** *is an example of software used for calling SNVs. Many software pipelines will carry out both mapping and variant calling such as RedDog (https://github.com/katholt/RedDog) and Snippy (https://github.com/tseemann/snippy).*

High confidence allele calls are concatenated to make a whole genome sequence alignment.

Phylogenetic inference: common workflows

Generate a core genome alignment

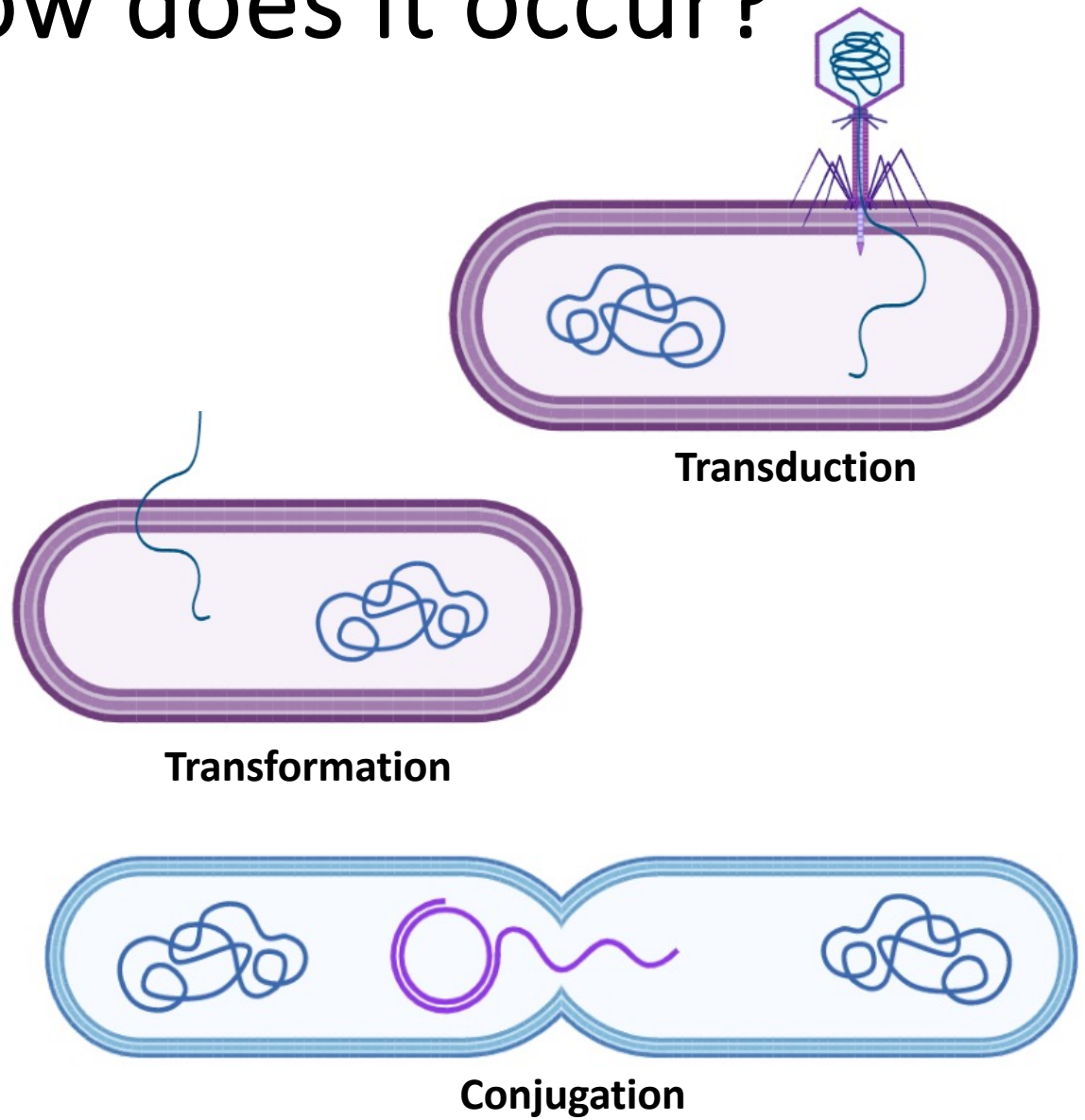
Option 1: Mapping-based

1. Map reads to reference
2. Call variants (SNPs/SNVs)
3. Filter recombination

4. **Infer phylogenetic tree**

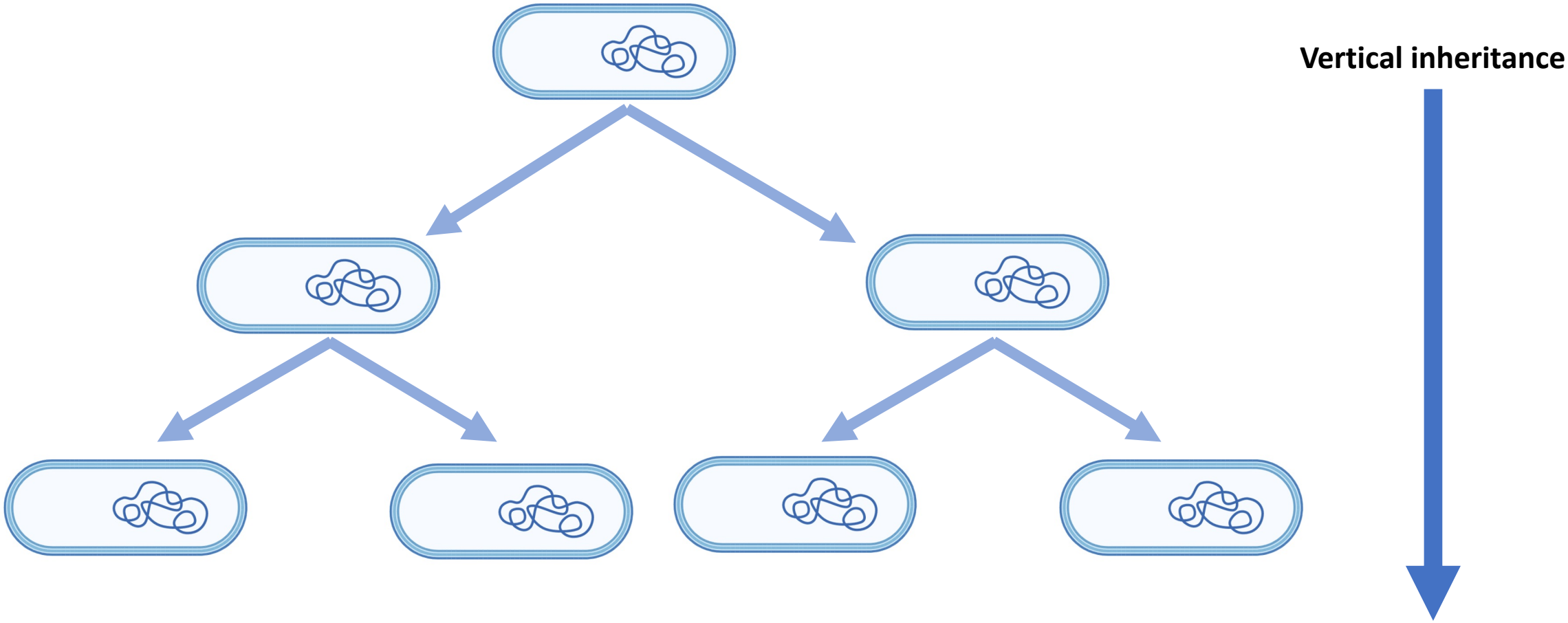
What is recombination & how does it occur?

- **Horizontal transfer** of genetic material from a donor to an acceptor (recipient) cell
- **Mobile genetic elements** (e.g. plasmids, bacteriophages) have their own evolutionary history which may be different that of their bacterial host
- Recombination introduces clusters of SNVs that can **confound phylogenetic inference**



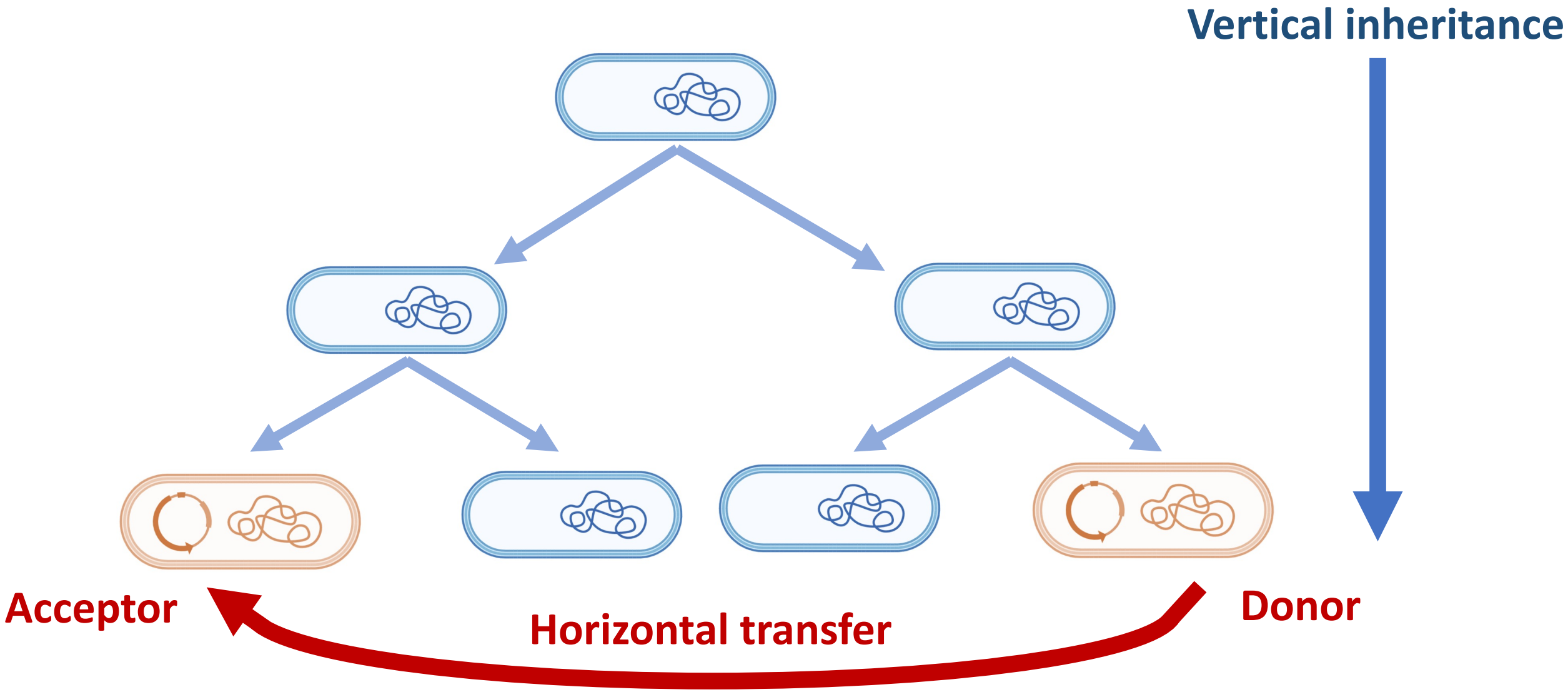
Mechanisms of recombination

Vertical inheritance



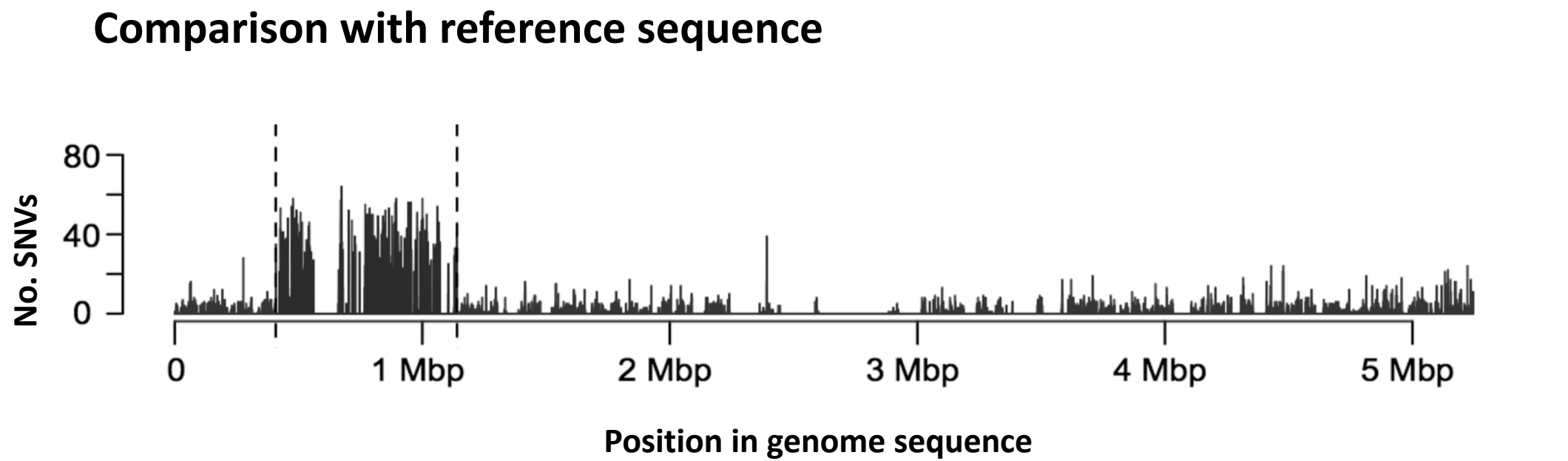
Substitutions (SNVs) occur randomly at a measurable rate during cell division due to errors ”typos” made by the molecular machinery responsible for DNA replication

Recombination



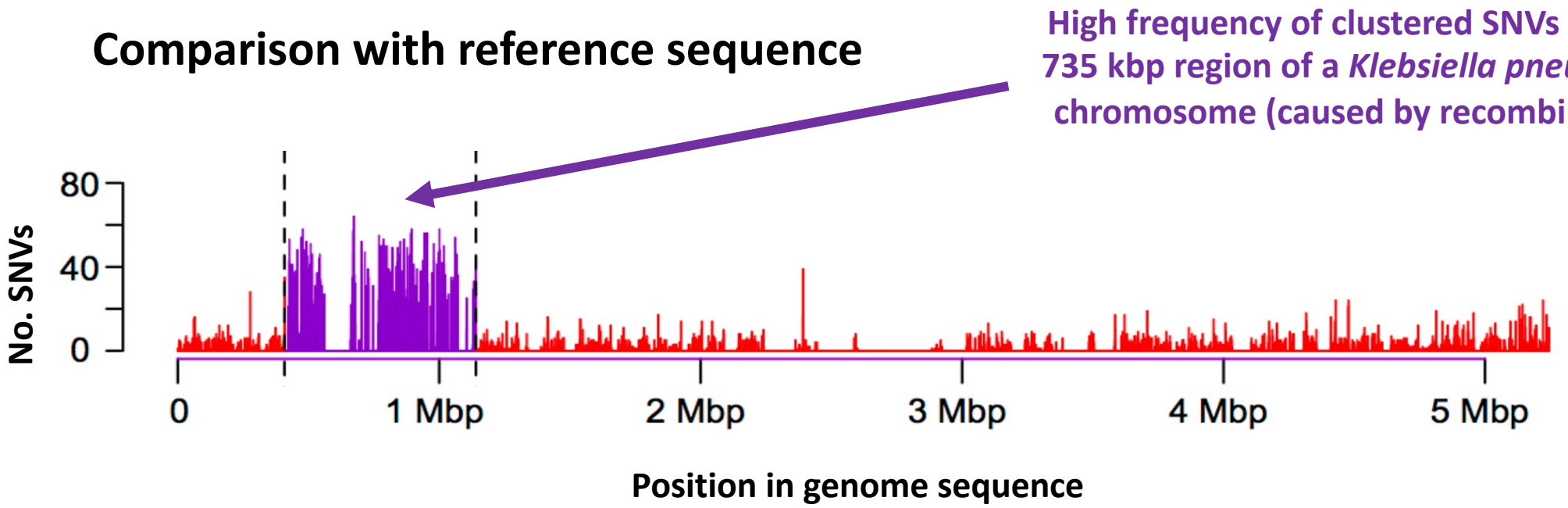
Introduction of clustered substitutions due to horizontal acquisitions

What does recombination look like?



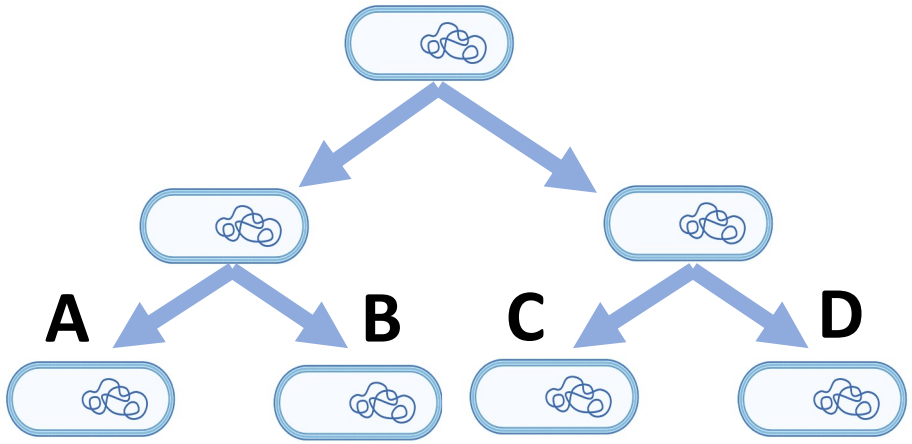
Holt K.E. *et al.* 2015, PNAS

What does recombination look like?



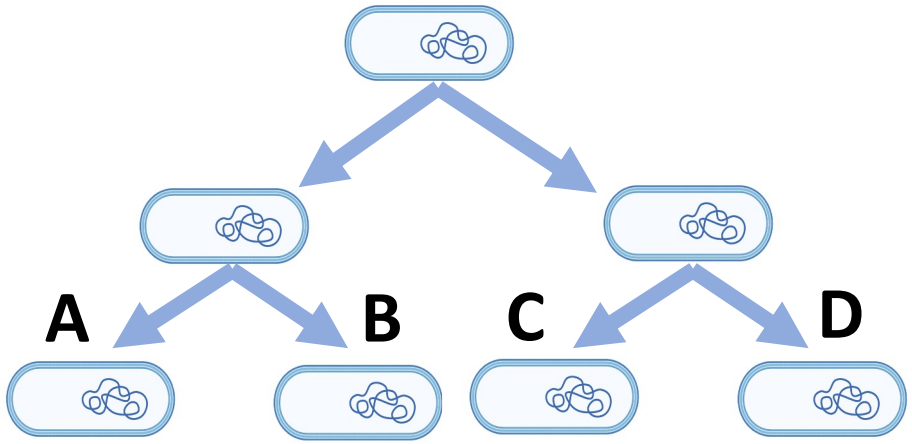
Holt K.E. *et al.* 2015, PNAS

Why does recombination matter?

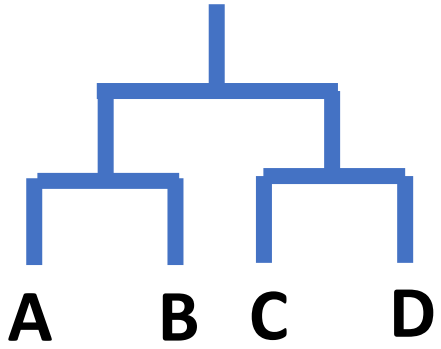


No recombination

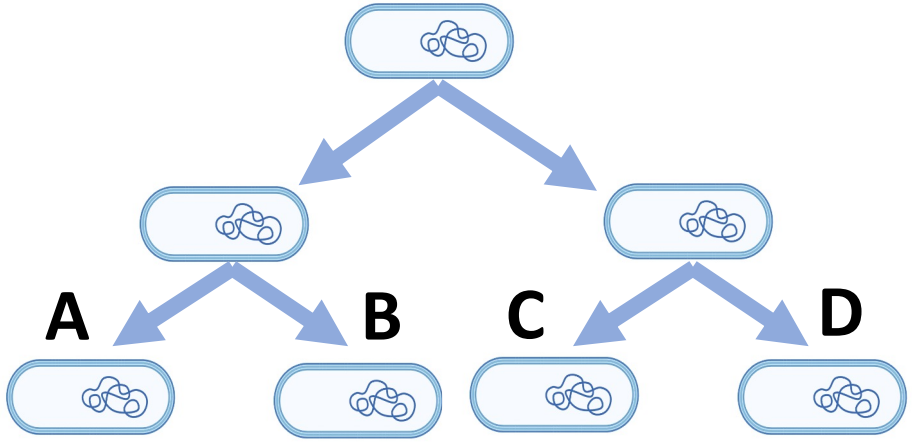
Why does recombination matter?



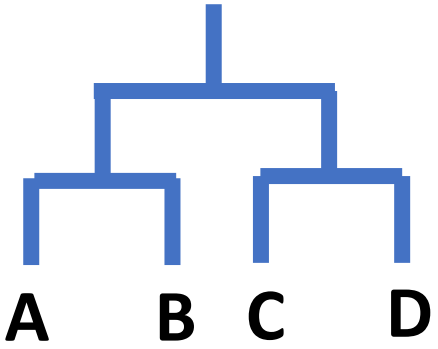
No recombination

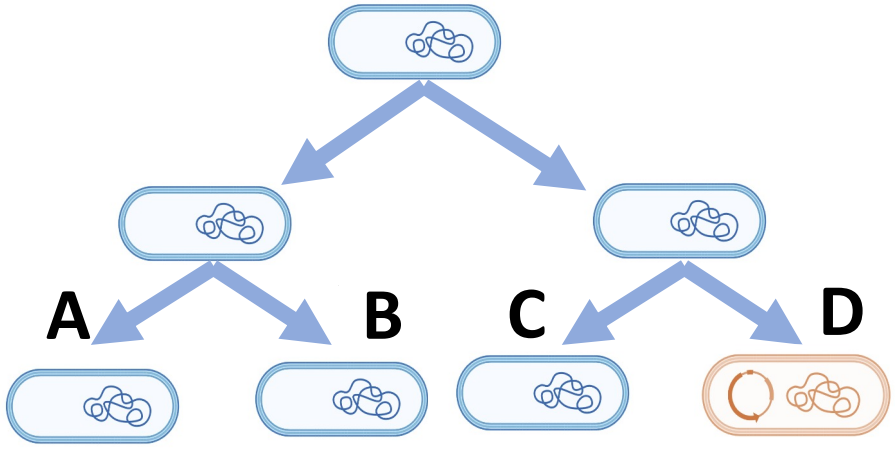


Why does recombination matter? Branch lengths!



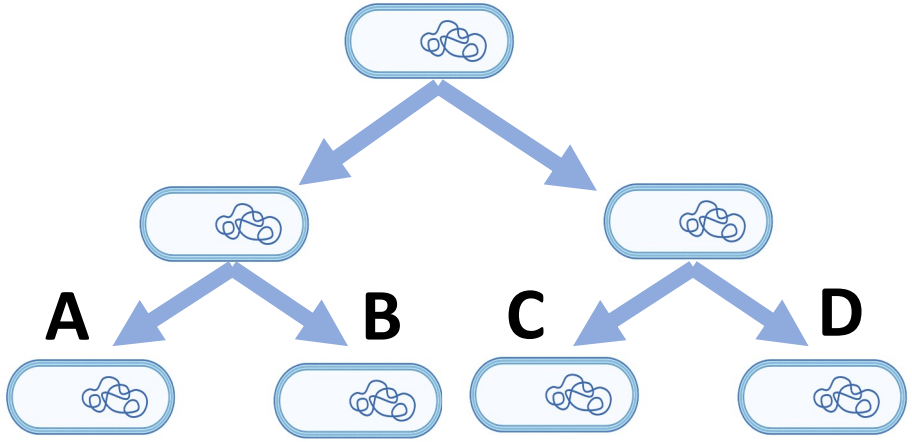
No recombination



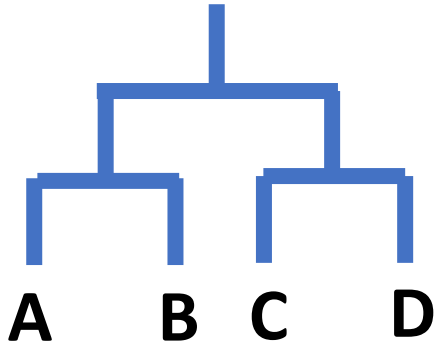


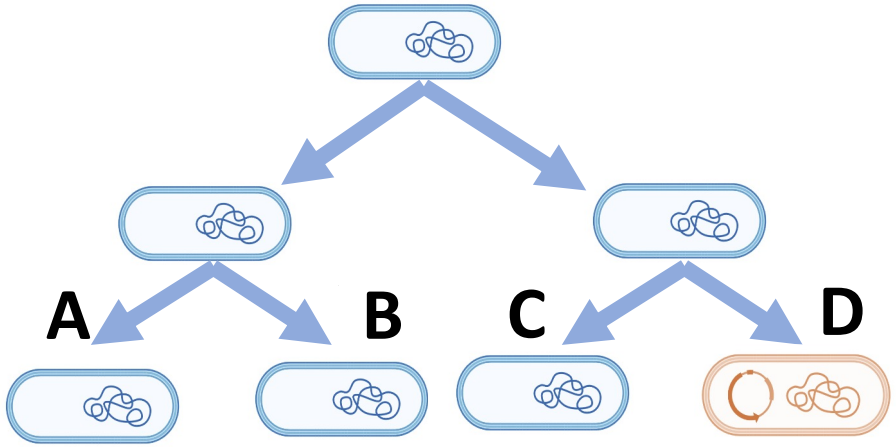
D horizontally acquires DNA

Why does recombination matter? Branch lengths!

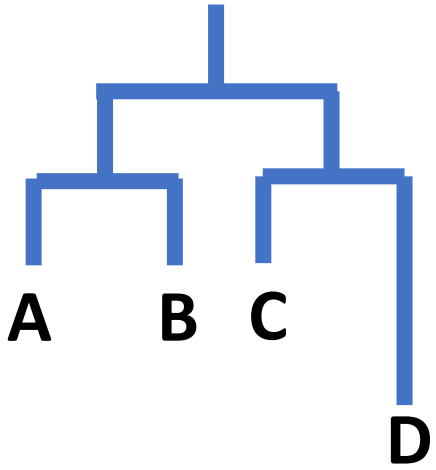


No recombination



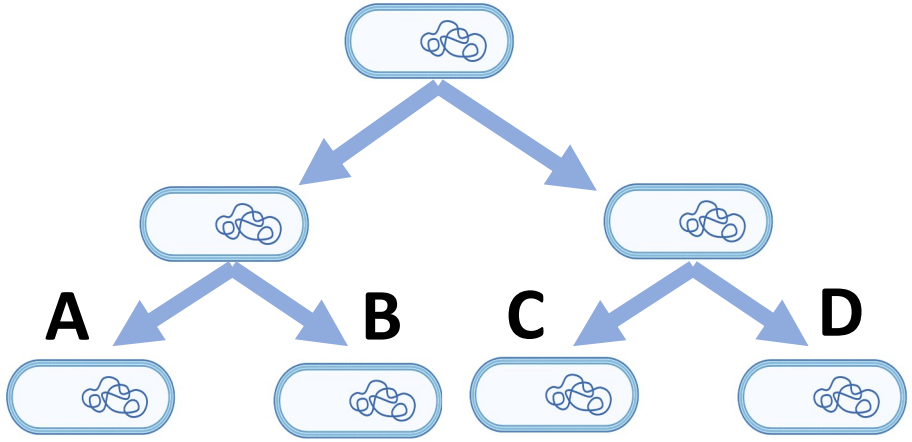


D horizontally acquires DNA

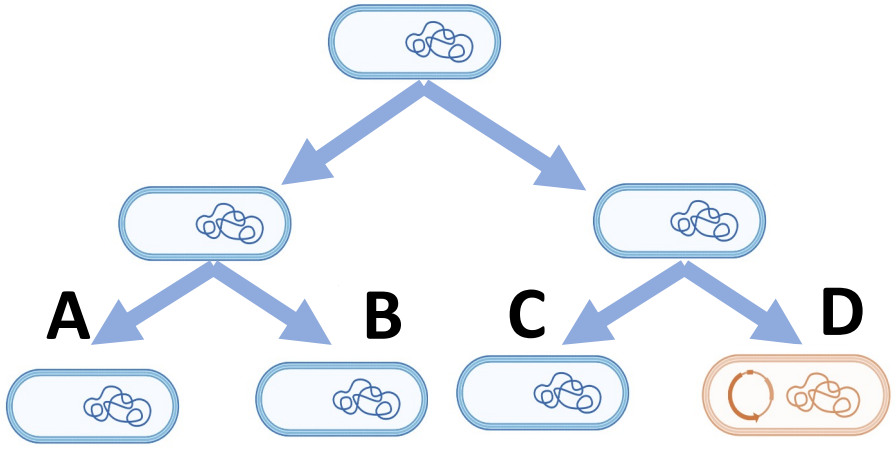


Inflated branch lengths (D)

Why does recombination matter? Temporal signal!

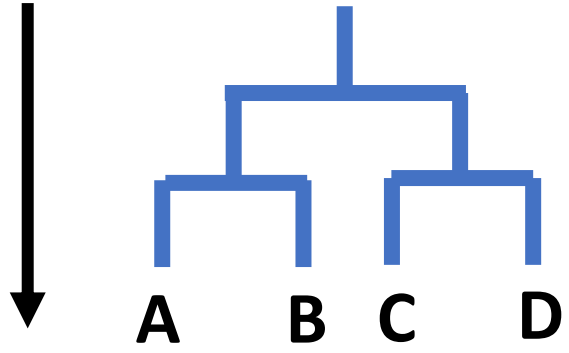


No recombination

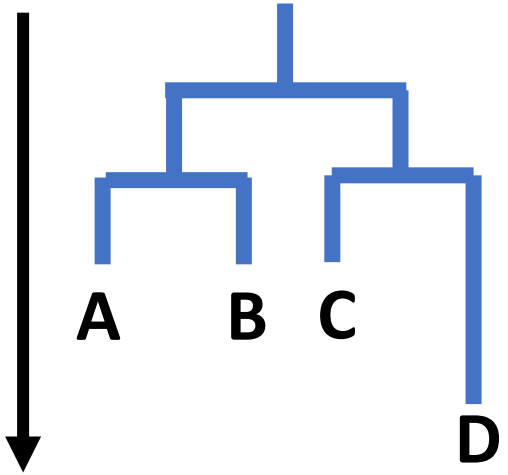


D horizontally acquires DNA

10 years

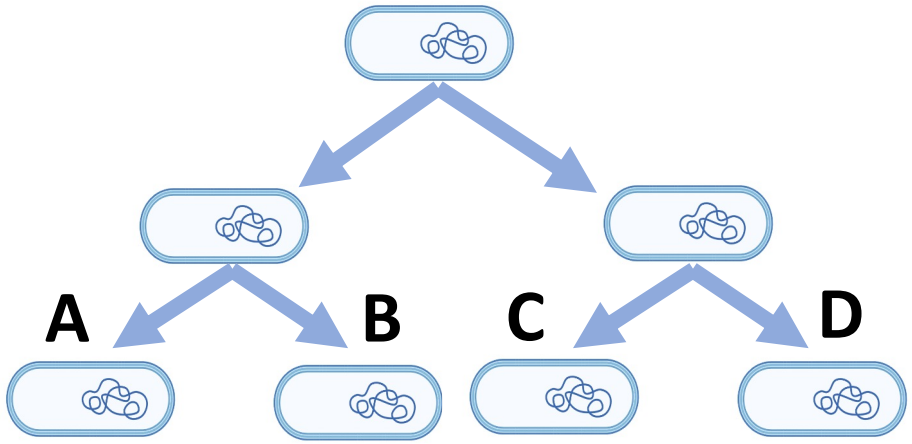


100 years

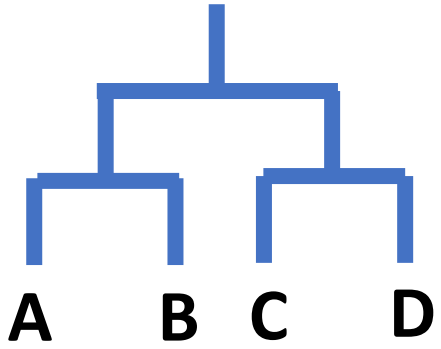


Destroys temporal signal!

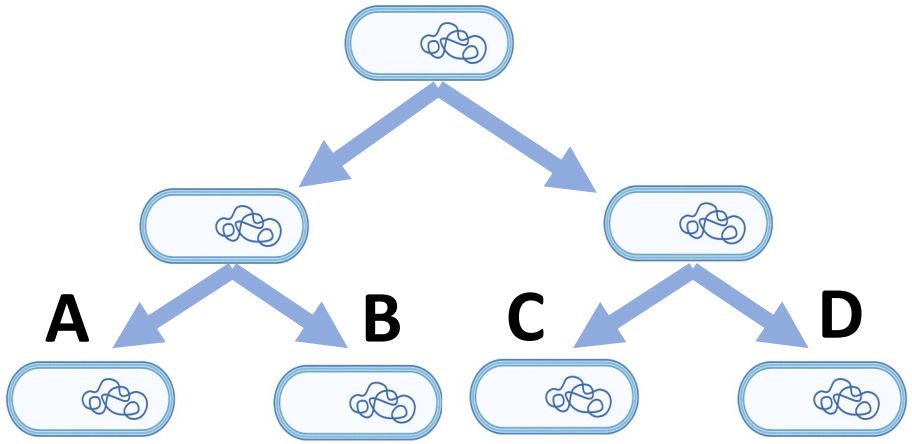
Why does recombination matter?



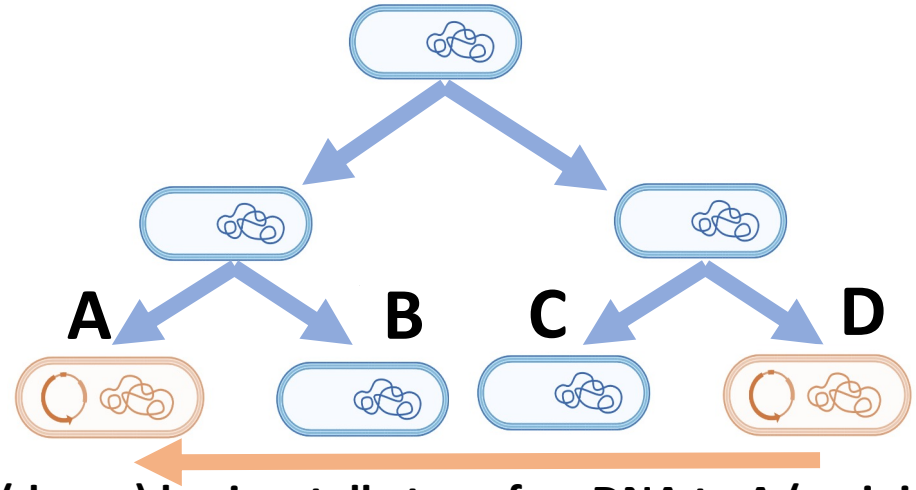
No recombination



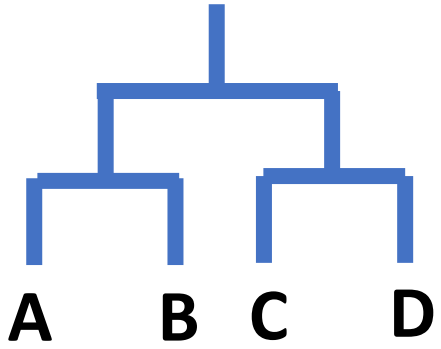
Why does recombination matter?



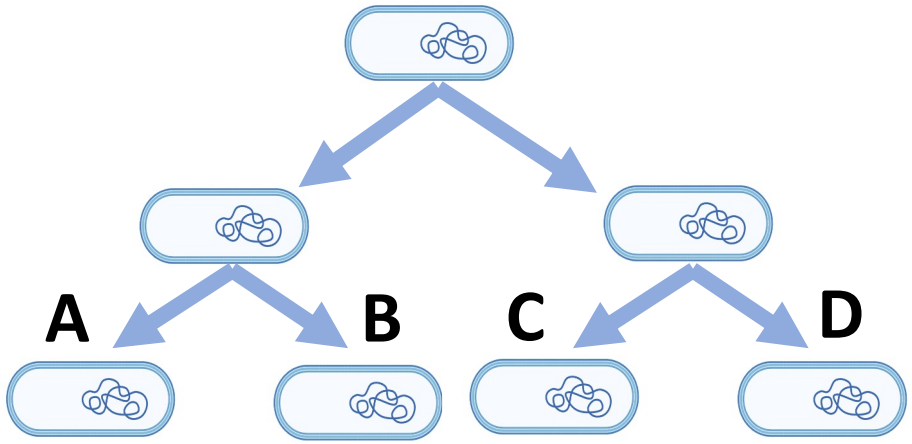
No recombination



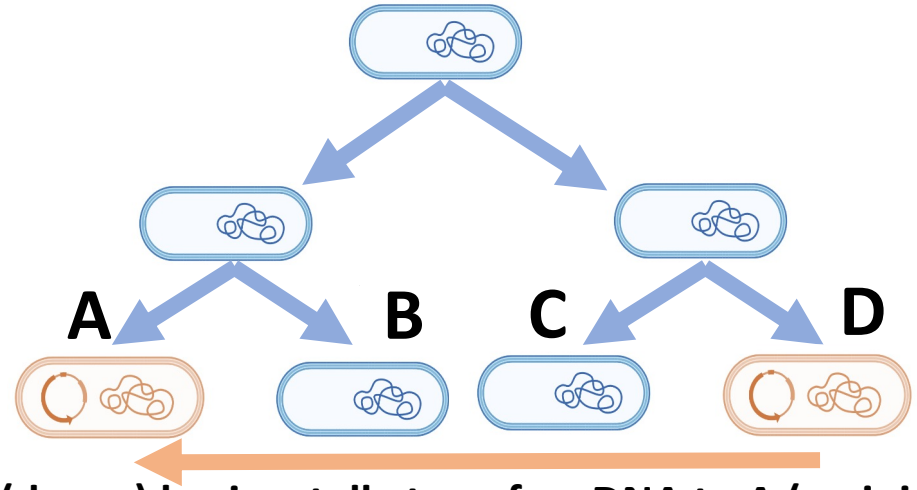
D (donor) horizontally transfers DNA to A (recipient)



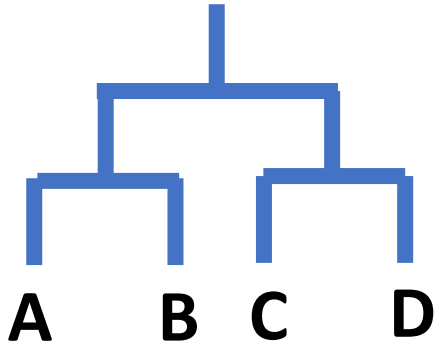
Why does recombination matter? Tree topology!

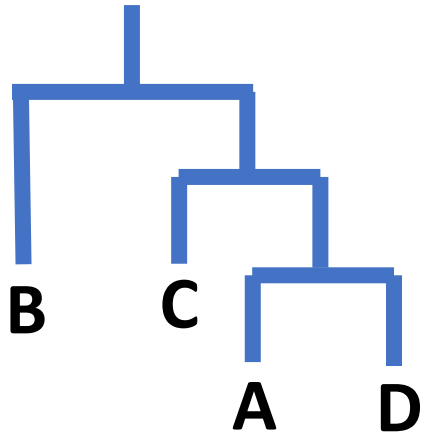


No recombination



D (donor) horizontally transfers DNA to A (recipient)





Incorrect tree topology (A&D)!

How do you remove recombination?

PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes

Xavier Didelot[1]*, Daniel J. Wilson[2,3]*

1 Department of Infectious Disease Epidemiology, Imperial College, L
Department of Medicine, University of Oxford, John Radcliffe Hospital,
Trust Centre for Human Genetics, Oxford, United Kingdom

Published online 20 November 2014

Nucleic Acids Research, 2015, Vol. 43, No. 3 ***e15***
doi: 10.1093/nar/gku1196

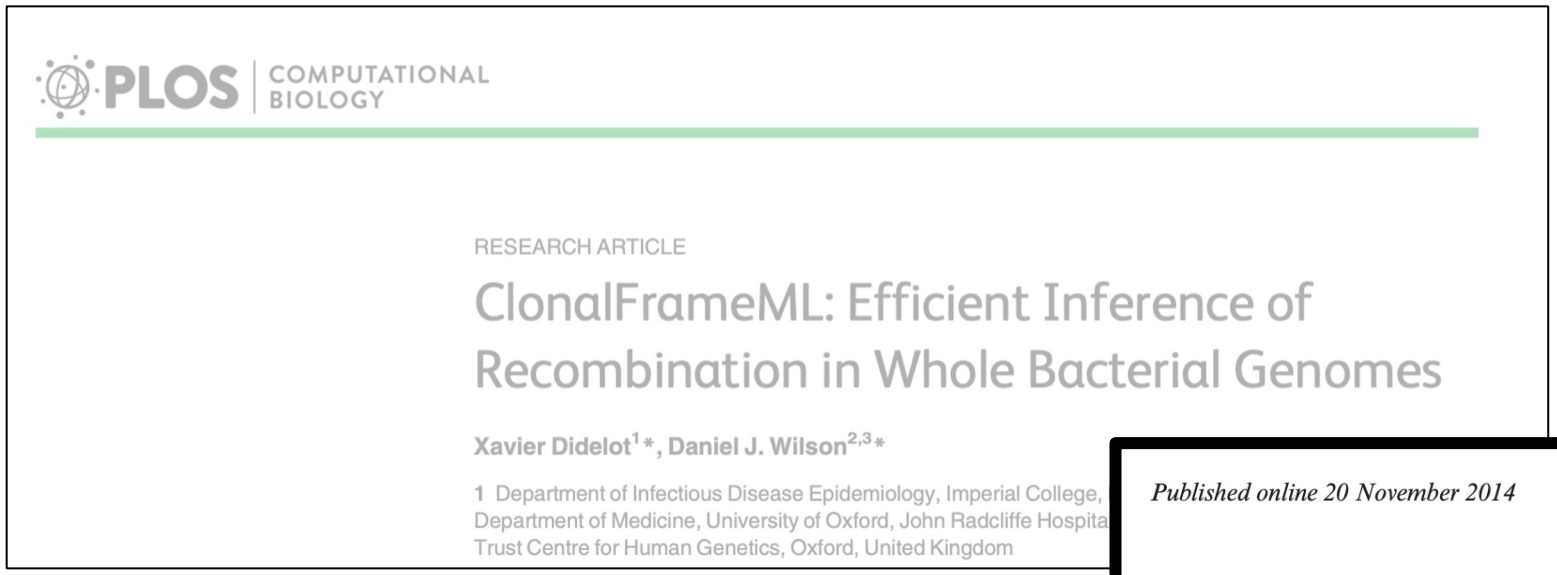
Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins

Nicholas J. Croucher[1,2,3], Andrew J. Page[1], Thomas R. Connor[1,4], Aidan J. Delaney[5], Jacqueline A. Keane[1], Stephen D. Bentley[1,6], Julian Parkhill[1] and Simon R. Harris[1,*]

[1]Pathogen Genomics, The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK, [2]Center for Communicable Disease Dynamics, Harvard School of Public Health, 677 Longwood Avenue, Boston, MA 02115, USA, [3]Department of Infectious Disease Epidemiology, Imperial College London, St. Mary's Campus, Norfolk Place, London W2 1PG, UK, [4]Cardiff School of Biosciences, Sir Martin Evans Building, Museum Avenue, Cardiff CF10 3AX, UK, [5]School of Computing, Engineering and Mathematics, University of Brighton, Brighton BN2 4GJ, UK and [6]Department of Medicine, University of Cambridge, Addenbrooke's Hospital, Cambridge CB2 0SP, UK

Specialised software tools for detection of clustered SNVs driven by recombination

How do you remove recombination?



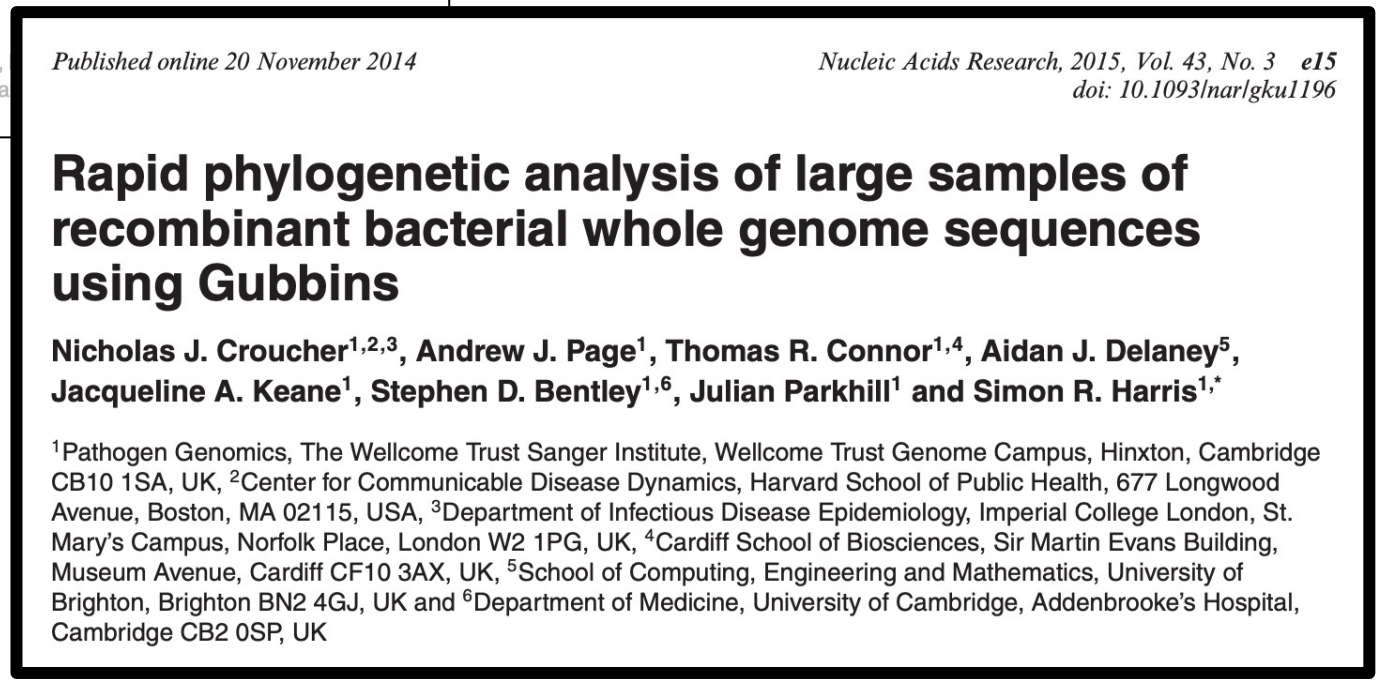
PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes

Xavier Didelot[1]*, Daniel J. Wilson[2,3]*

1 Department of Infectious Disease Epidemiology, Imperial College,
Department of Medicine, University of Oxford, John Radcliffe Hospita
Trust Centre for Human Genetics, Oxford, United Kingdom



Published online 20 November 2014

Nucleic Acids Research, 2015, Vol. 43, No. 3 ***e15***
doi: 10.1093/nar/gku1196

Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins

Nicholas J. Croucher[1,2,3], Andrew J. Page[1], Thomas R. Connor[1,4], Aidan J. Delaney[5], Jacqueline A. Keane[1], Stephen D. Bentley[1,6], Julian Parkhill[1] and Simon R. Harris[1,*]

[1]Pathogen Genomics, The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK, [2]Center for Communicable Disease Dynamics, Harvard School of Public Health, 677 Longwood Avenue, Boston, MA 02115, USA, [3]Department of Infectious Disease Epidemiology, Imperial College London, St. Mary's Campus, Norfolk Place, London W2 1PG, UK, [4]Cardiff School of Biosciences, Sir Martin Evans Building, Museum Avenue, Cardiff CF10 3AX, UK, [5]School of Computing, Engineering and Mathematics, University of Brighton, Brighton BN2 4GJ, UK and [6]Department of Medicine, University of Cambridge, Addenbrooke's Hospital, Cambridge CB2 0SP, UK

Gubbins:

- Rapid automated detection of recombination
- Detects recombination within closely related bacterial sequences (i.e. within a clone)

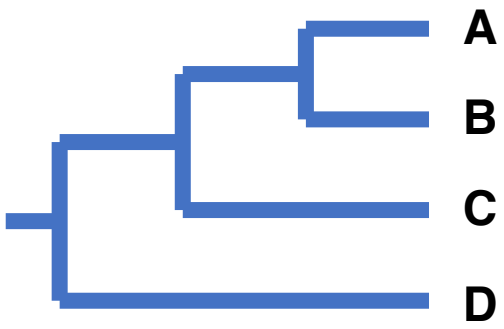
Specialised software tools for detection of clustered SNVs driven by recombination

How does Gubbins remove recombination?

Input: chromosomal whole genome sequence alignment

```
Sequence A: ...CGTTAGTACACT...
Sequence B: ...CGATAGTTCACT...
Sequence C: ...CGTTGGTTTACG...
Sequence D: ...CCTTGGTTTACG...
```

Step 1: Infer phylogenetic tree (e.g. RAxML)

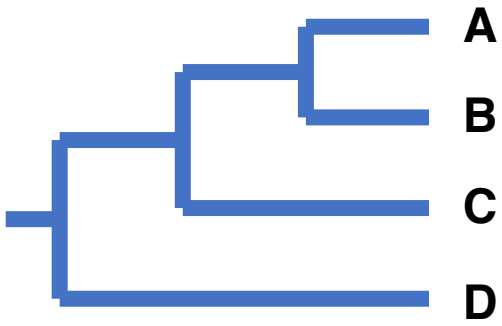


How does Gubbins remove recombination?

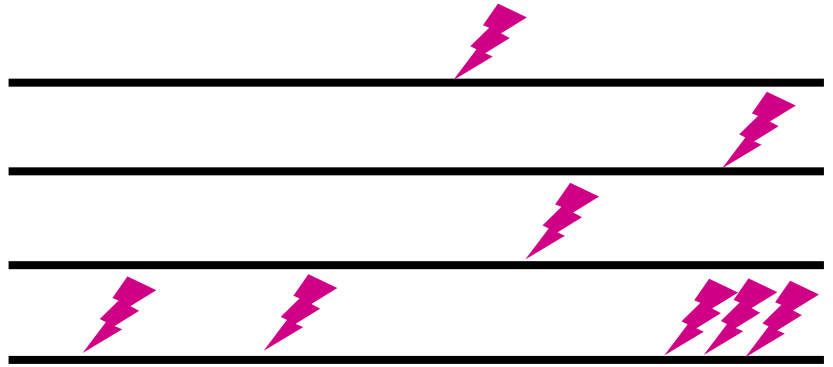
Input: chromosomal whole genome sequence alignment

```
Sequence A: ...CGTTAGTACACT...
Sequence B: ...CGATAGTTCACT...
Sequence C: ...CGTTGGTTTACG...
Sequence D: ...CCTTGGTTTACG...
```

Step 1: Infer phylogenetic tree (e.g. RAxML)



Step 2: Map SNVs to tree (Ancestral State Reconstruction)

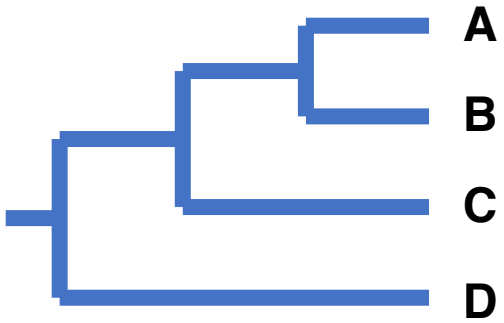


How does Gubbins remove recombination?

Input: chromosomal whole genome sequence alignment

```
Sequence A: ...CGTTAGTACACT...
Sequence B: ...CGATAGTTCACT...
Sequence C: ...CGTTGGTTTACG...
Sequence D: ...CCTTGGTTTACG...
```

Step 1: Infer phylogenetic tree (e.g. RAxML)



Step 2: Map SNVs to tree (Ancestral State Reconstruction)

Recombination!

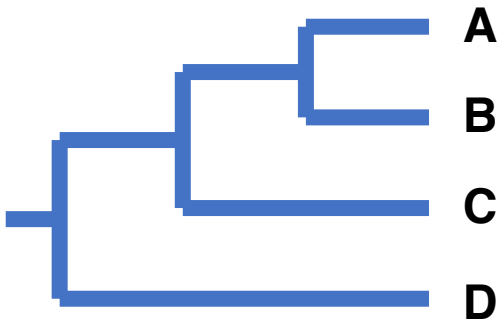
Step 3: Spatial statistics to detect windows with a high density of SNVs

How does Gubbins remove recombination?

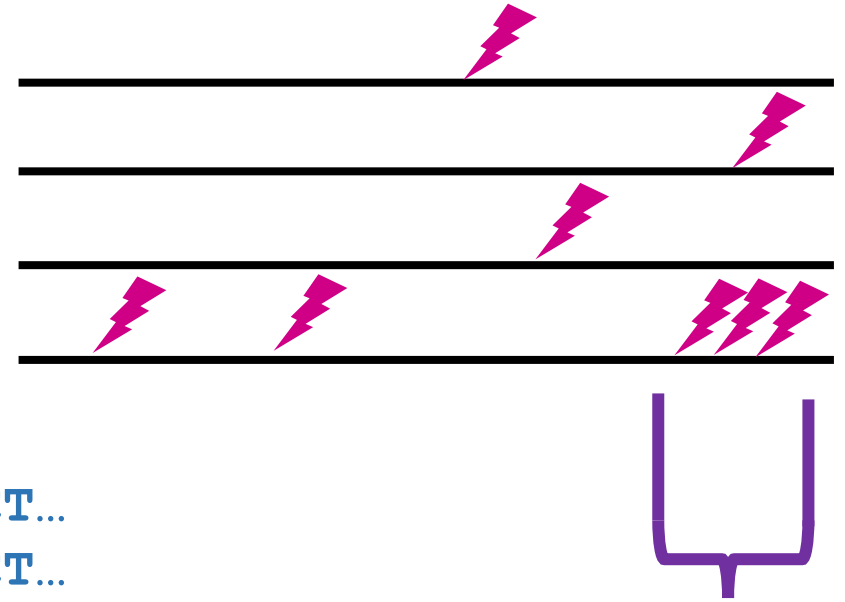
Input: chromosomal whole genome sequence alignment

```
Sequence A: ...CGTTAGTACACT...
Sequence B: ...CGATAGTTCACT...
Sequence C: ...CGTTGGTTTACG...
Sequence D: ...CCTTGGTTTACG...
```

Step 1: Infer phylogenetic tree (e.g. RAxML)



Step 2: Map SNVs to tree (Ancestral State Reconstruction)



Recombination!

Step 3: Spatial statistics to detect windows with a high density of SNVs

```
...ATCTTTGCGGAATCTCT...
...ATCTTTGCGGAATCTCT...
...ATCTTTGCGGAATCTCT...
...ATCTTTGCGGA---TCT...
```

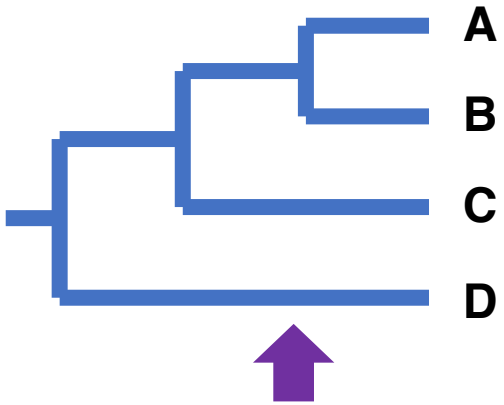
Step 4: Mask recombinant region(s) from alignment.

How does Gubbins remove recombination?

Input: chromosomal whole genome sequence alignment

```
Sequence A: ...CGTTAGTACACT...
Sequence B: ...CGATAGTTCACT...
Sequence C: ...CGTTGGTTTACG...
Sequence D: ...CCTTGGTTTACG...
```

Step 1: Infer phylogenetic tree (e.g. RAxML)




```
...ATCTTTGCGGAATCTCT...
...ATCTTTGCGGAATCTCT...
...ATCTTTGCGGAATCTCT...
...ATCTTTGCGGA---TCT...
```

Step 2: Map SNVs to tree (Ancestral State Reconstruction)

Recombination!

Step 3: Spatial statistics to detect windows with a high density of SNVs

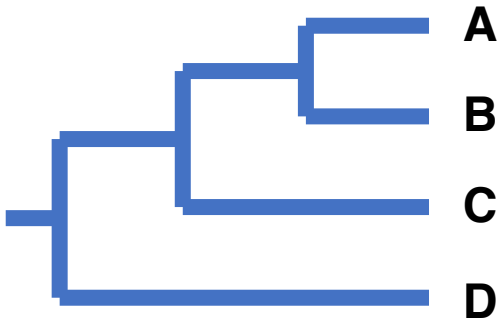
Step 4: Mask recombinant region(s) from alignment. **Repeat steps 1-4 until no further recombination is detected.**

How does Gubbins remove recombination?

Input: chromosomal whole genome sequence alignment

```
Sequence A: ...CGTTAGTACACT...
Sequence B: ...CGATAGTTCACT...
Sequence C: ...CGTTGGTTTACG...
Sequence D: ...CCTTGGTTTACG...
```

Step 1: Infer phylogenetic tree (e.g. RAxML)



Step 2: Map SNVs to tree (Ancestral State Reconstruction)

Recombination!

Step 3: Spatial statistics to detect windows with a high density of SNVs

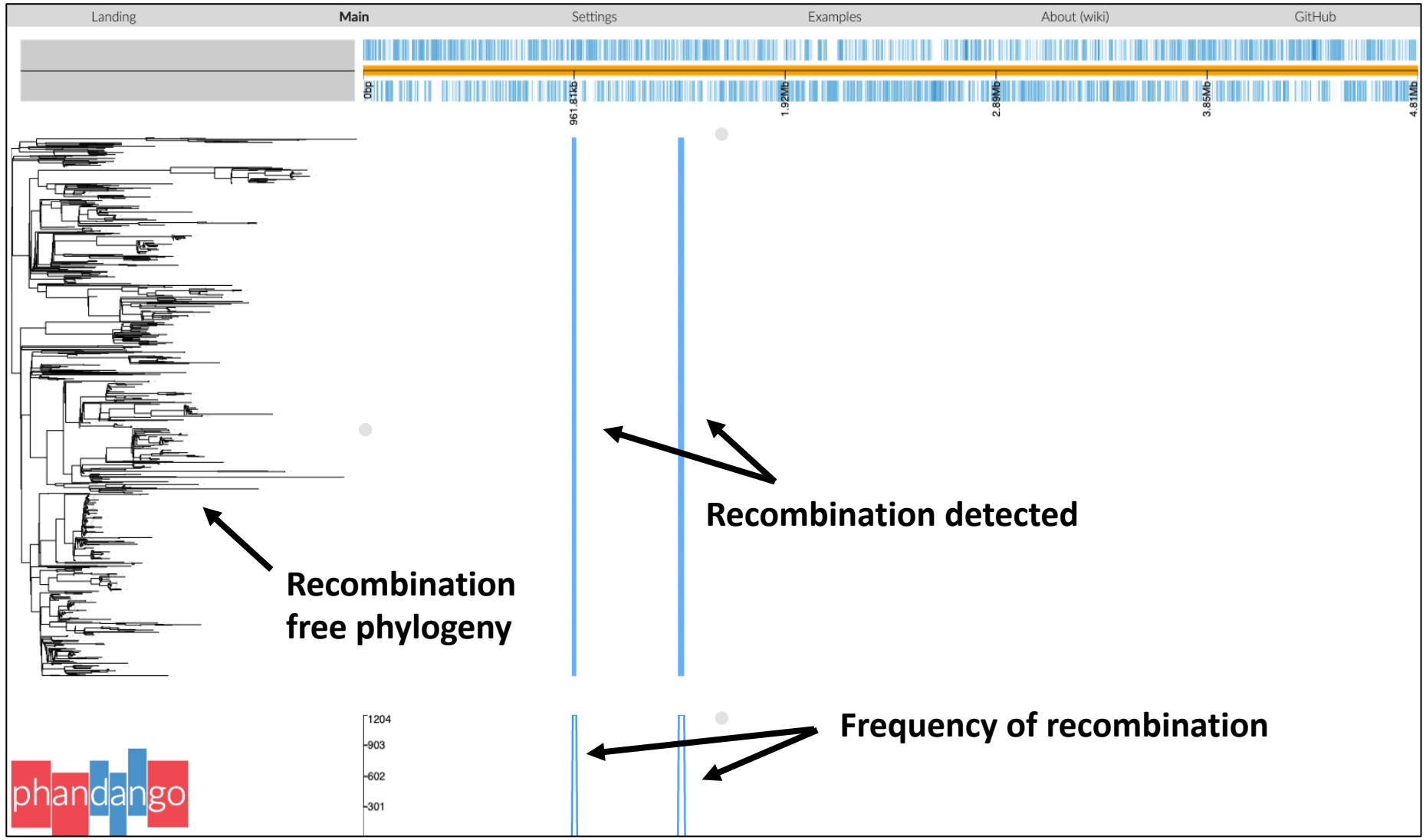
Step 4: Mask recombinant region(s) from alignment. **Repeat steps 1-4 until no further recombination is detected.**

```
...ATCTTTGCGGAATCTCT...
...ATCTTTGCGGAATCTCT...
...ATCTTTGCGGAATCTCT...
...ATCTTTGCGGA---TCT...
```

Output: recombination filtered SNV alignment & details of recombinant regions

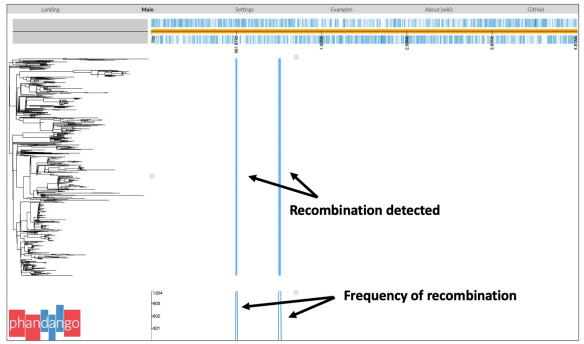
Visualisation of Gubbins output

e.g. *Salmonella enterica* serovar Typhi genotype H58 (low recombination)



Kariuki *et al.* 2021, eLife.

Phandango data visualization tool: https://jameshadfield.github.io/phandango/



Join at menti.com | use code 8692 7095

Mentimeter

How might this impact tree topology?

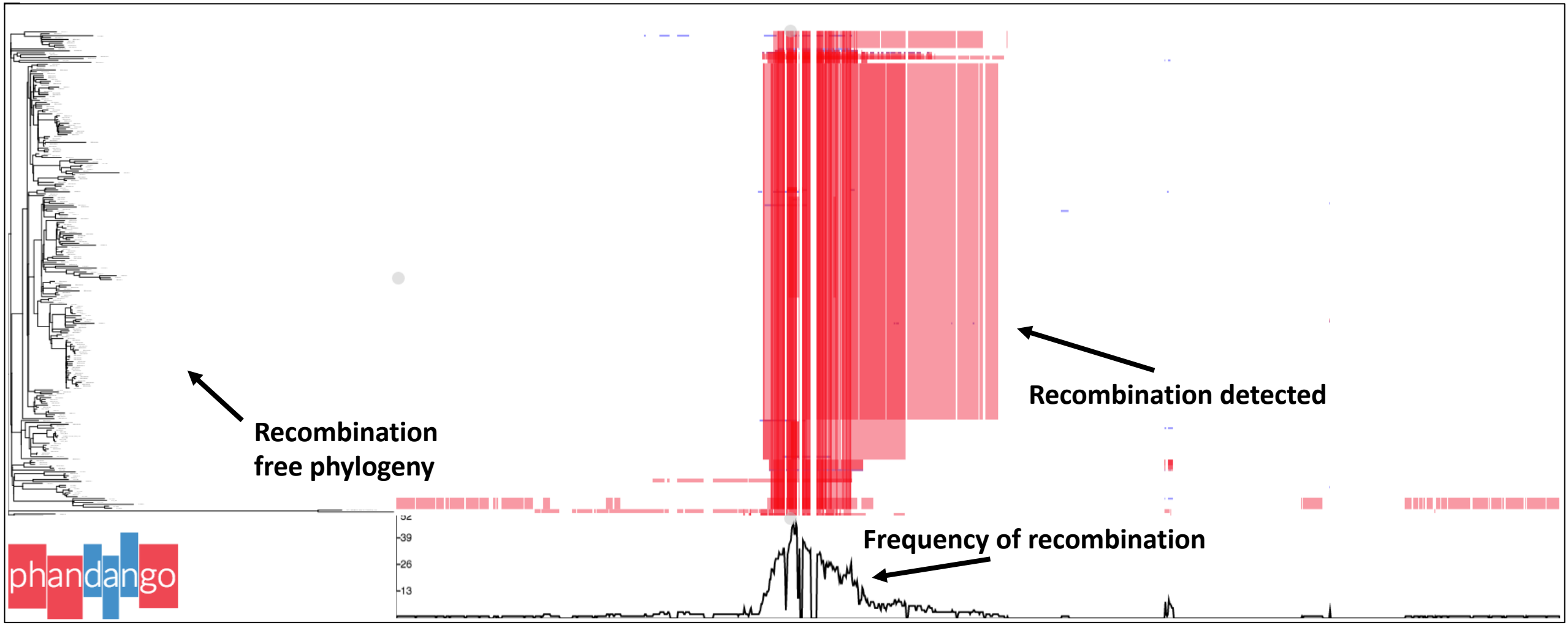
Inflated branch lengths

Incorrect tree topology

No temporal signal

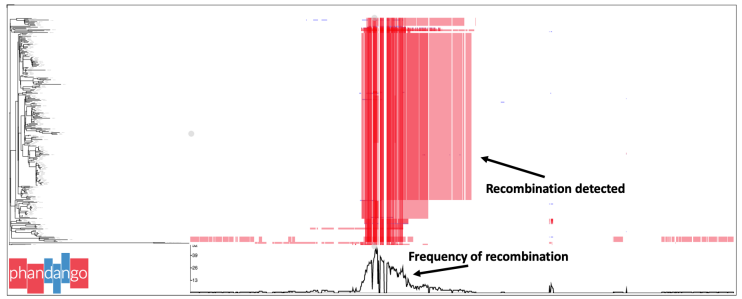
Visualisation of Gubbins output

e.g. *Klebsiella pneumoniae* (high recombination)



Wyres *et al.* 2019, PLoS Genetics.

Phandango data visualization tool: https://jameshadfield.github.io/phandango/



Join at menti.com | use code 8692 7095

Mentimeter

How might this impact tree topology?

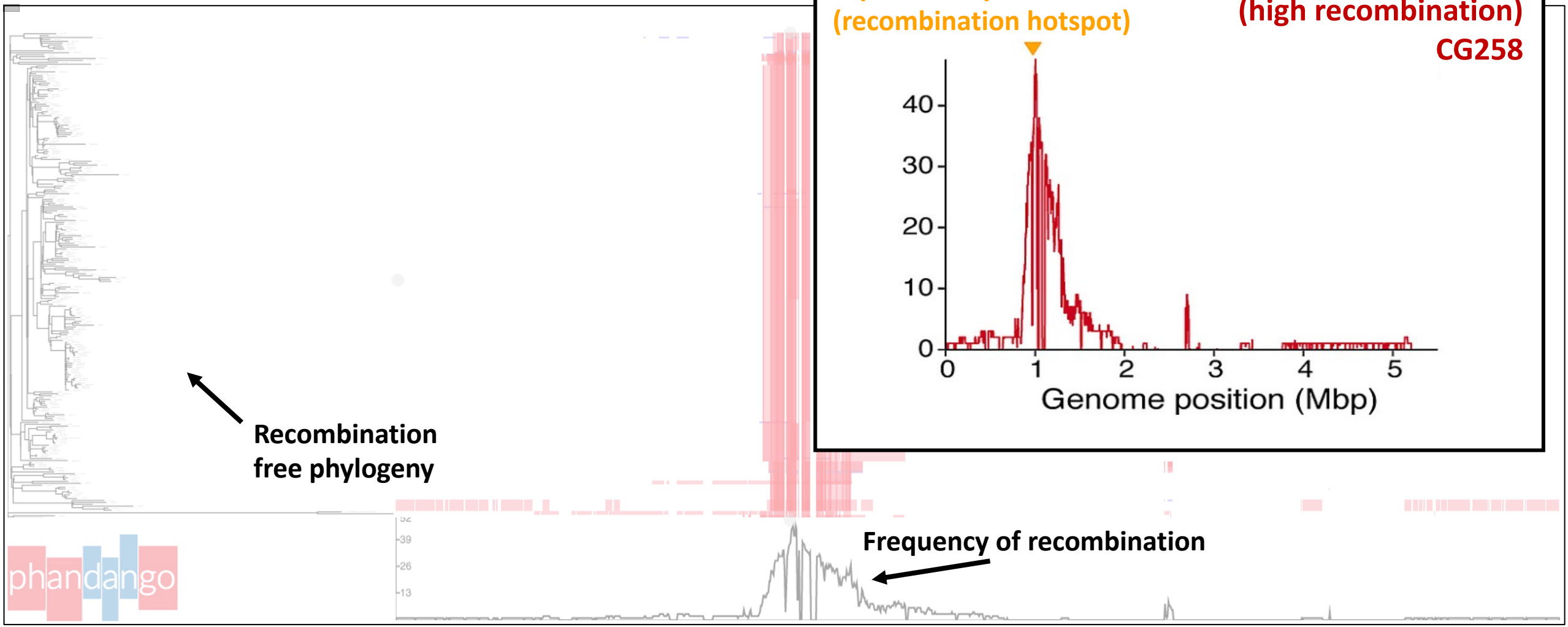
Inflated branch lengths

Incorrect tree topology

No temporal signal

Recombination can be biologically informative

e.g. *Klebsiella pneumoniae* (high recombination)



Wyres *et al.* 2019, PLoS Genetics.

Phandango data visualization tool: https://jameshadfield.github.io/phandango/

Phylogenetic inference: common workflows

Generate a core genome alignment

Option 1: Mapping-based

1. Map reads to reference
2. Call variants (SNPs/SNVs)
3. Filter recombination

Option 2: Assembly-based

1. Assemble genomes
2. Annotate genes
3. Infer pangenome

4. **Infer phylogenetic tree**

Phylogenetic inference: common workflows

Generate a core genome alignment

Option 2: Assembly-based

1. Assemble genomes
2. Annotate genes
3. Infer pangenome

4. **Infer phylogenetic tree**

Phylogenetic inference: common workflows

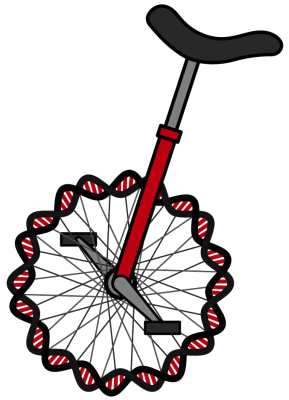
Generate a core genome alignment

Option 2: Assembly-based

1. Assemble genomes
2. Annotate genes
3. Infer pangenome

4. **Infer phylogenetic tree**

Tools for assembling genomes from reads



Unicycler

https://github.com/rrwick/Unicycler

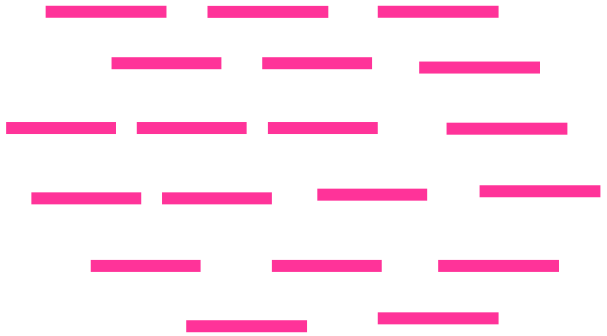
SPAdes

https://github.com/ablab/spades

Velvet

https://www.ebi.ac.uk/~zerbino/velvet/

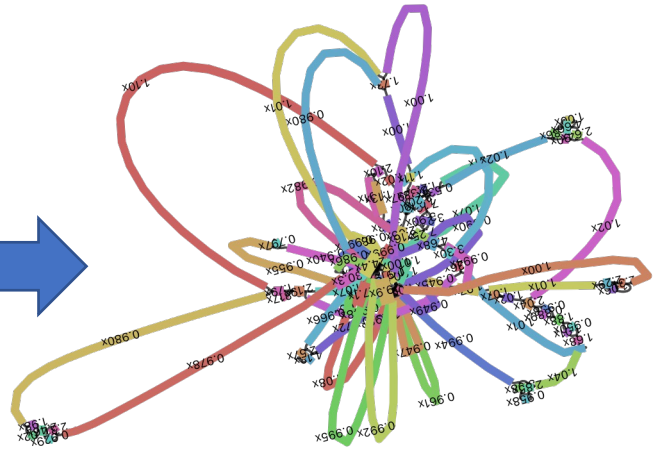
Genome assembly and annotation



Illumina short read data

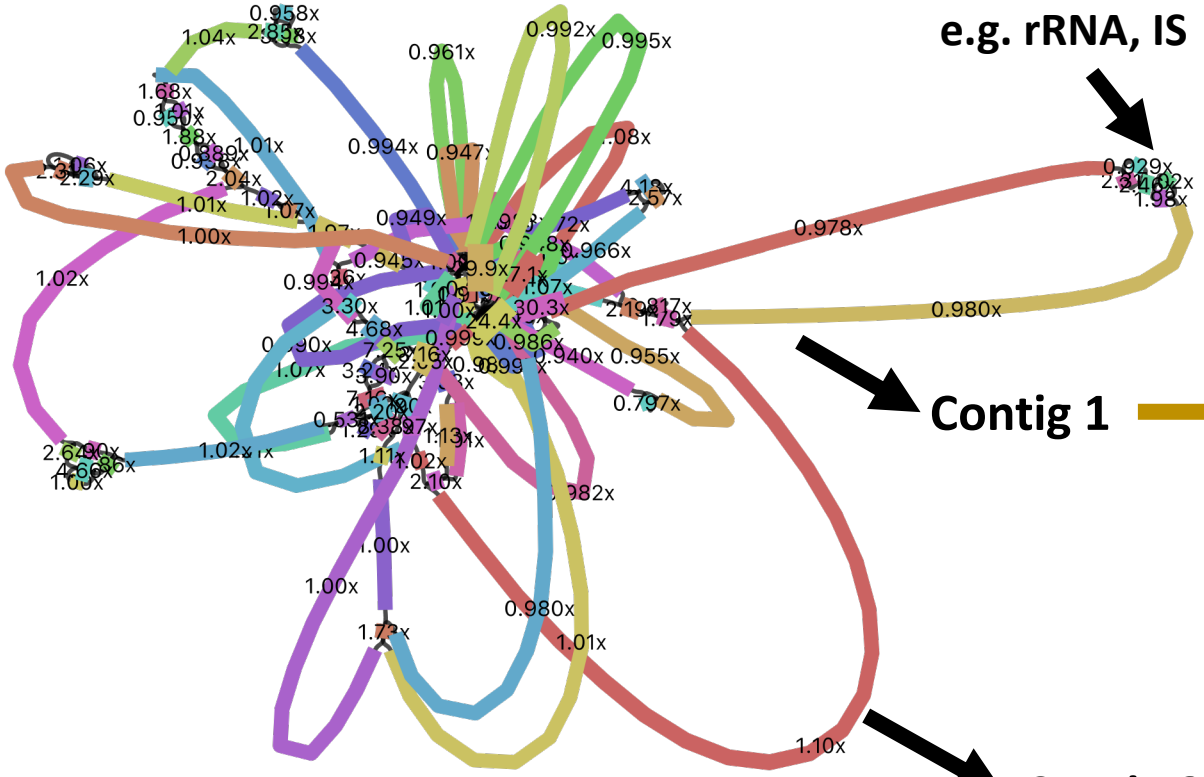
Unicycler

Assembly software



Assembled genome
(shown as ‘De Brujin` graph)

What does an assembled genome look like?



Assembled genome shown as a graph

Repetitive regions
e.g. rRNA, IS

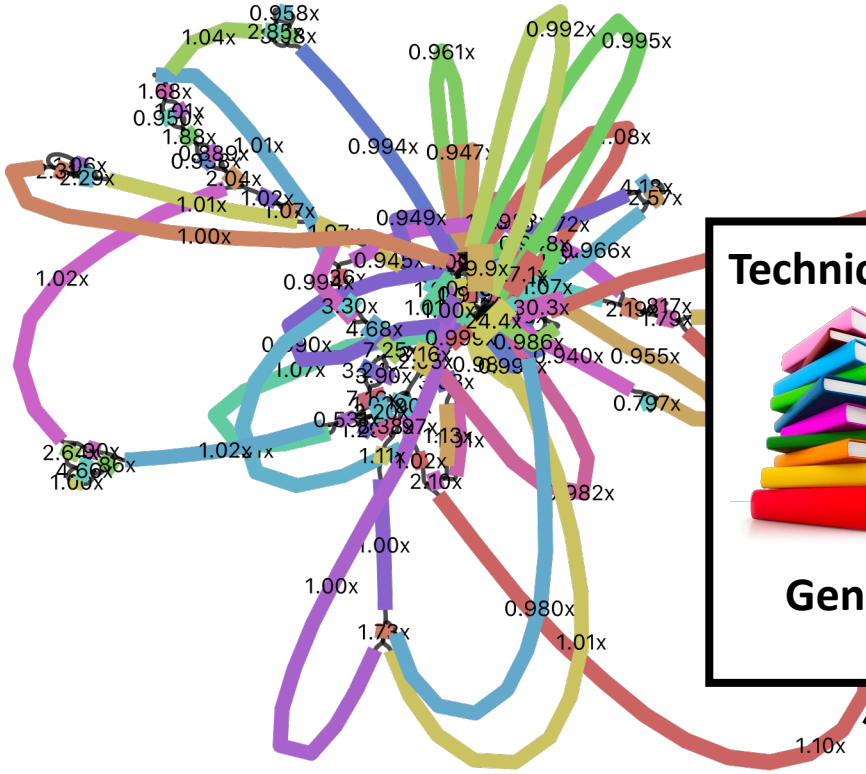
Contig 1

Contig 2

Contigs in a fasta file format:

```
>contig_1_header
TGGGCGGATGGTGATTATACCCTGAGTGTGTCAGTC
GAAGATAAAGCGGGGAACACCAGCCATTCTGCAT
CGCTGACGGTGACGGTGGACACGCAAATCGCCATT
AATAACATTGAACTGGTCAATGACAGCGGTATTCC
CAACGATAATCTGACTAATAATGTGCGTCCA
>contig_2_header
GGGCAGCCCGTTCGCTGCGCGCCCGGTCTGTCCAA
CTGGCTGCGCCAGTTGTCGAACCCCGGTCGGTGGT
TCTCATCCCCCCTTGGTTTGGGGGATACATATAAGCA
AAAAGCCTGTACTTCTGTACAGGCTCTCAACTT
GAAGATGGCGGTGAGGGGGGGATT
>contig_3_header
...
```

What does an assembled genome look like?



Assembled genome shown as a graph

Repetitive regions
e.g. rRNA, IS

Technical limitations of short read sequencing

Genome → Short reads → Reconstructed genome

Contig 2

Contigs in a fasta file format:

```
>contig_1_header
TGGGCGGATGGTGATTATACCCTGAGTGTGTCAGTC
GAAGATAAAGCGGGGAACACCAGCCATTCTGCAT
GTGACGGTGGACACGCAAATCGCCATT
GAACTGGTCAATGACAGCGGTATTCC
ATCTGACTAATAATGTGCGTCCA
header
CGTTCGCTGCGCGCCCGGTCTGTCCAA
GCCAGTTGTCGAACCCCGGTCGGTGGT
CCCTTGGTTTGGGGGATACATATAAGCA
GTACTTCTGTACAGGCTCTCAACTT
CGGTGAGGGGGGGATT
>contig_3_header
...
```

Phylogenetic inference: common workflows

Generate a core genome alignment

Option 2: Assembly-based

1. Assemble genomes
2. Annotate genes
3. Infer pangenome

4. **Infer phylogenetic tree**

Tools for annotating genomes: identifying genes and predicting their function

PROKKA

https://github.com/tseemann/prokka

RAST Rapid Annotation using Subsystem Technology

https://rast.nmpdr.org/

https://github.com/oschwengers/bakta

Genome annotation: identify putative genes

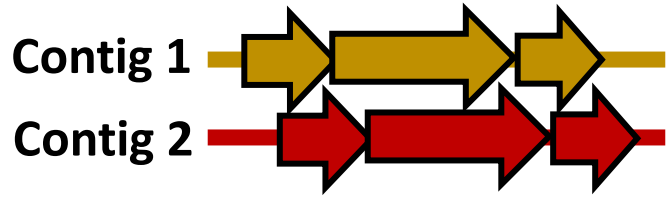
Assembled contigs:

Annotation software:

PROKKA

https://github.com/tseemann/prokka

Annotated genes:



Tool (reference)	Features predicted
Prodigal (Hyatt 2010)	Coding sequence (CDS)
RNAmmer (Lagesen *et al.* , 2007)	Ribosomal RNA genes (rRNA)
Aragorn (Laslett and Canback, 2004)	Transfer RNA genes
SignalP (Petersen *et al.* , 2011)	Signal leader peptides
Infernal (Kolbe and Eddy, 2011)	Non-coding RNA

Phylogenetic inference: common workflows

Generate a core genome alignment

Option 2: Assembly-based

1. Assemble genomes
2. Annotate genes
3. Infer pangenome

4. **Infer phylogenetic tree**

Pangenome inference software

Bioinformatics, 31(22), 2015, 3691–3693
doi: 10.1093/bioinformatics/btv421
Advance Access Publication Date: 20 July 2015
Applications Note

OXFORD

Sequence analysis

Roary: rapid large-scale prokaryote pan genome analysis

Andrew J. Page[1,*], Carla A. Cummins[1], Martin Hunt[1], Vanessa K. Wong[1,2], Sandra Reuter[2], Matthew T.G. Holden[3], Maria Fookes[1], Daniel Falush[4], Jacqueline A. Keane[1] and Julian Parkhill[1]

https://sanger-pathogens.github.io/Roary/

Tonkin-Hill *et al. Genome Biology* (2020) 21:180
https://doi.org/10.1186/s13059-020-02090-4

Genome Biology

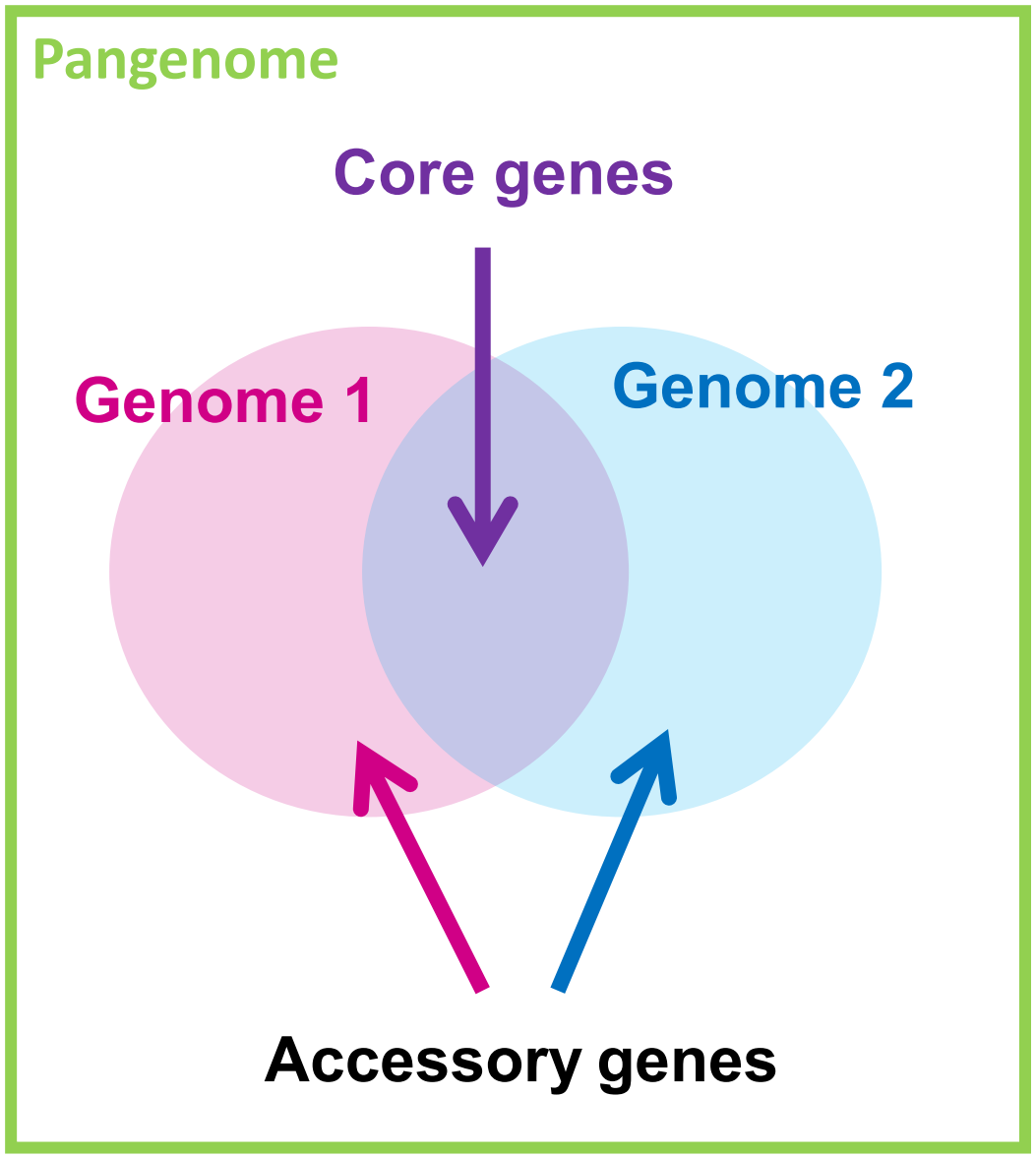
SOFTWARE Open Access

Producing polished prokaryotic pangenomes with the Panaroo pipeline

Gerry Tonkin-Hill[1,2*], Neil MacAlasdair[1,3], Christopher Ruis[3,4,5], Aaron Weimann[3,4,5,6], Gal Horesh[1], John A. Lees[7], Rebecca A. Gladstone[2], Stephanie Lo[1], Christopher Beaudoin[8], R. Andres Floto[4,9], Simon D.W. Frost[10,11], Jukka Corander[1,2,12†], Stephen D. Bentley[1†] and Julian Parkhill[3†]

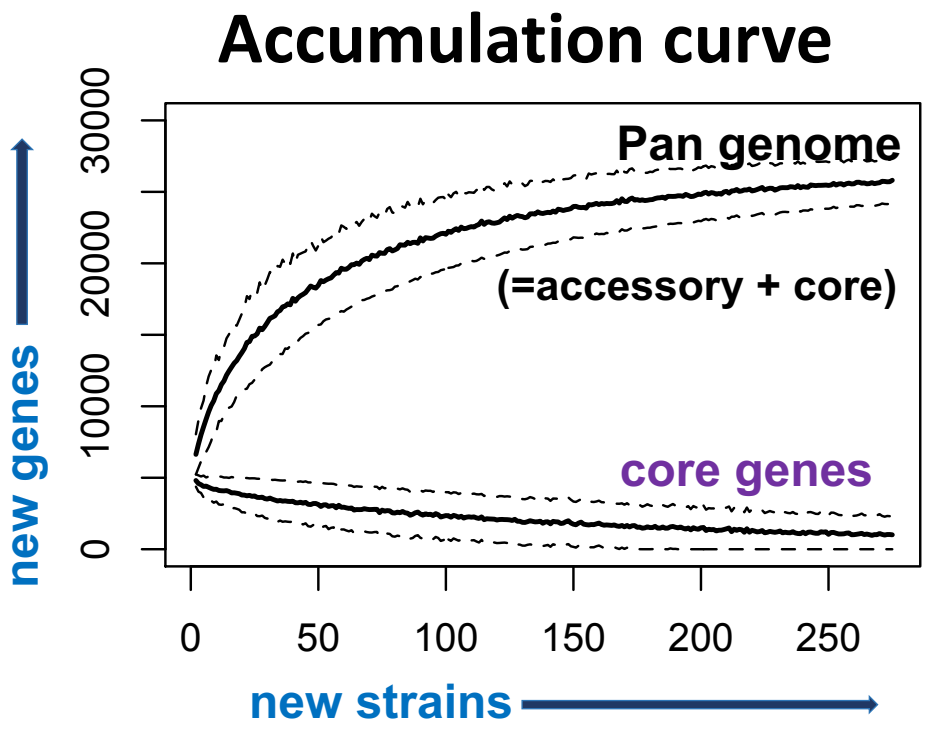
https://github.com/gtonkinhill/panaroo

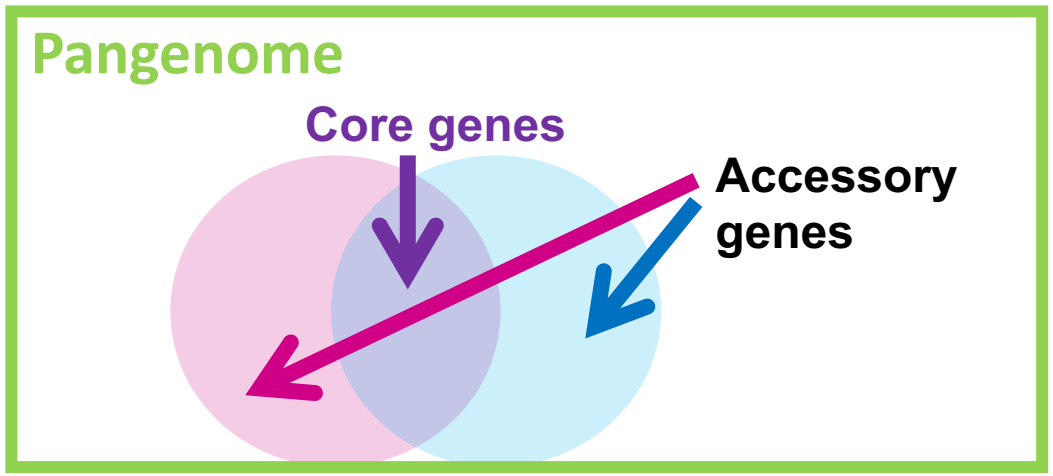
Pangenome analysis: identifying a core genome



- A **pangenome** is the set of all genes that have been found in a collection of genomes as a whole
- The **core genome** is the set of genes present in all members of a species, however, in practice this is usually defined using a threshold e.g. genes shared across 95-99% of genomes
- Non-core genes form the **accessory genome**

Pangenome analysis: identifying a core genome





- Therefore, the **pangenome embraces both core and accessory genes**
- As more genomes are sequenced, additional gene content is discovered, until the true diversity is largely sampled, and the accumulation curve plateaus
- As more genomes sequenced and in the pangenome analysis the number of core genes decreases
- **Core genes can be concatenated to give a core genome alignment free of recombination**, however, data loss occurs as non-coding regions are excluded

Phylogenetic inference: common workflows

Generate a core genome alignment

Option 1: Mapping-based

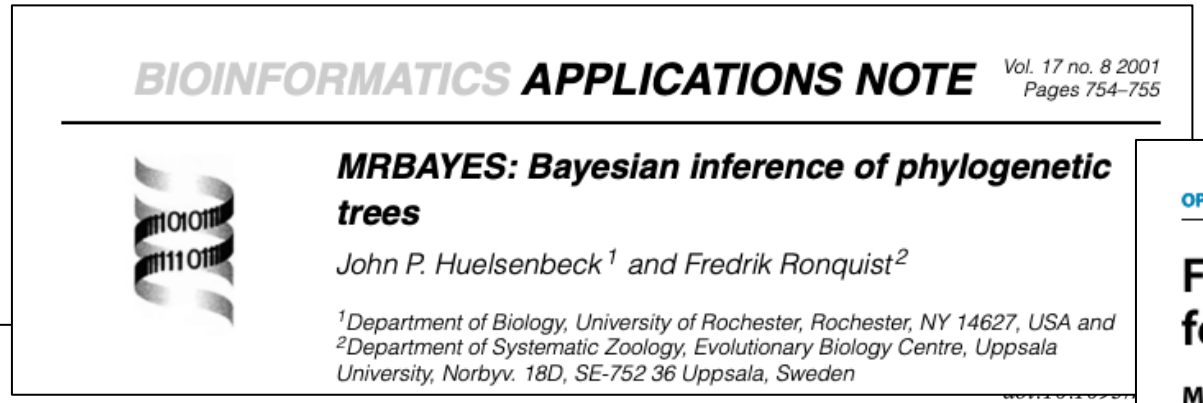
1. Map reads to reference
2. Call variants (SNPs/SNVs)
3. Filter recombination

Option 2: Assembly-based

1. Assemble genomes
2. Annotate genes
3. Infer pangenome

4. Infer phylogenetic tree

Inferring a phylogenetic tree

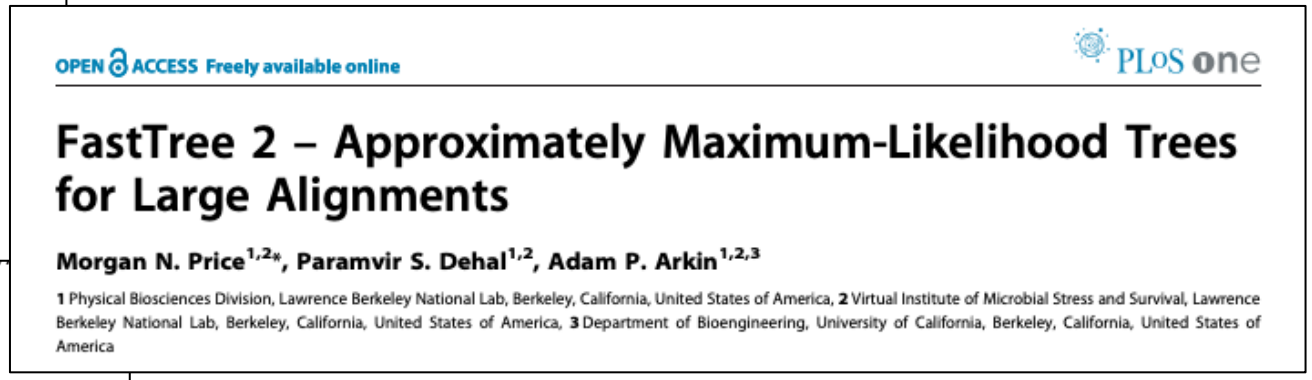


BIOINFORMATICS APPLICATIONS NOTE Vol. 17 no. 8 2001 Pages 754–755

MRBAYES: Bayesian inference of phylogenetic trees

John P. Huelsenbeck[1] and Fredrik Ronquist[2]

[1]Department of Biology, University of Rochester, Rochester, NY 14627, USA and [2]Department of Systematic Zoology, Evolutionary Biology Centre, Uppsala University, Norbyv. 18D, SE-752 36 Uppsala, Sweden

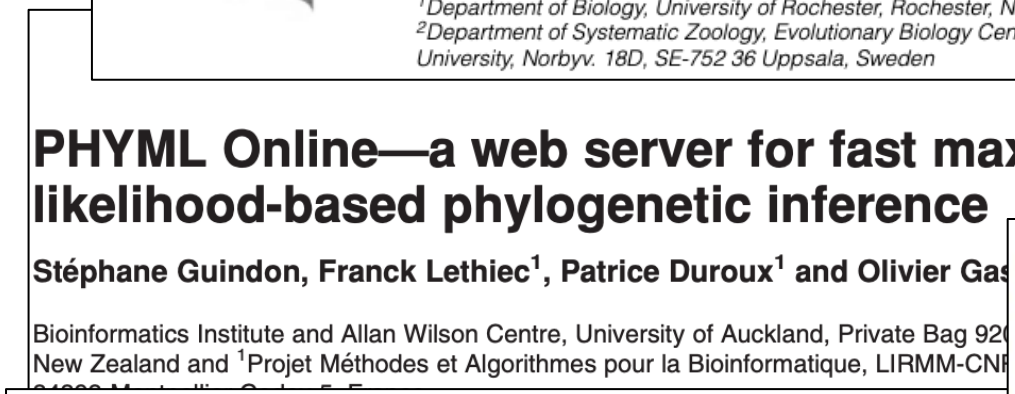


OPEN ACCESS Freely available online PLoS one

FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments

Morgan N. Price[1,2]*, Paramvir S. Dehal[1,2], Adam P. Arkin[1,2,3]

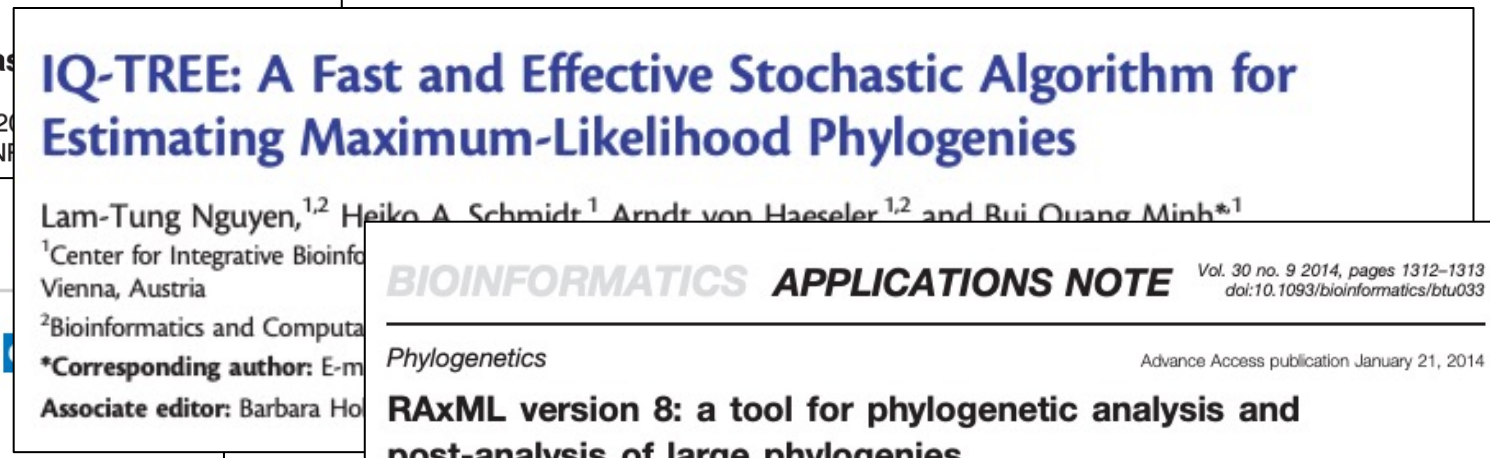
1 Physical Biosciences Division, Lawrence Berkeley National Lab, Berkeley, California, United States of America, 2 Virtual Institute of Microbial Stress and Survival, Lawrence Berkeley National Lab, Berkeley, California, United States of America, 3 Department of Bioengineering, University of California, Berkeley, California, United States of America



PHYML Online—a web server for fast maximum likelihood-based phylogenetic inference

Stéphane Guindon, Franck Lethiec[1], Patrice Duroux[1] and Olivier Gas

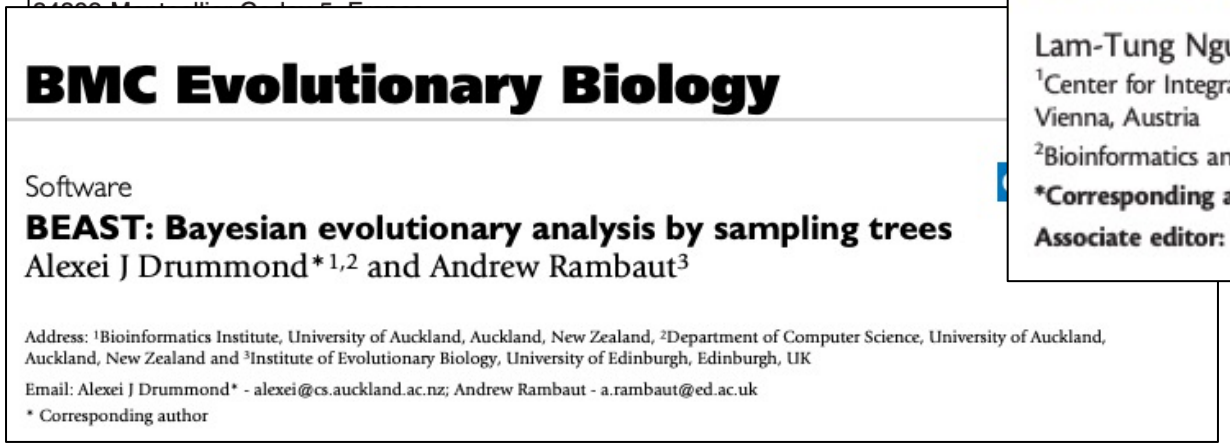
Bioinformatics Institute and Allan Wilson Centre, University of Auckland, Private Bag 92
New Zealand and [1]Projet Méthodes et Algorithmes pour la Bioinformatique, LIRMM-CNR



IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies

Lam-Tung Nguyen,[1,2] Heiko A. Schmidt,[1] Arndt von Haeseler[1,2] and Bui Quang Minh*[,1]

[1]Center for Integrative Bioinfo
Vienna, Austria
[2]Bioinformatics and Computa
*Corresponding author: E-m
Associate editor: Barbara Ho



BMC Evolutionary Biology

Software

BEAST: Bayesian evolutionary analysis by sampling trees

Alexei J Drummond*[1,2] and Andrew Rambaut[3]

Address: [1]Bioinformatics Institute, University of Auckland, Auckland, New Zealand, [2]Department of Computer Science, University of Auckland, Auckland, New Zealand and [3]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK

Email: Alexei J Drummond* - alexei@cs.auckland.ac.nz; Andrew Rambaut - a.rambaut@ed.ac.uk

* Corresponding author

BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 9 2014, pages 1312–1313 doi:10.1093/bioinformatics/btu033

Phylogenetics Advance Access publication January 21, 2014

RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies

Alexandros Stamatakis[1,2]

[1]Scientific Computing Group, Heidelberg Institute for Theoretical Studies, 69118 Heidelberg and [2]Department of Informatics, Institute of Theoretical Informatics, Karlsruhe Institute of Technology, 76128 Karlsruhe, Germany

Associate Editor: Jonathan Wren

Many different specialised software tools exist for phylogenetic inference!!!

Methods of phylogenetic inference

1. **Distance based methods –** Calculate the distance between pairs of sequences and infer a tree that best describes the distances observed

   e.g. Neighbour joining (NJ)

2. **Character based methods** – Examine each SNV, one at a time, and search for the tree with the best “score”

   e.g. Maximum Likelihood (ML)

Methods of phylogenetic inference

1. **Distance based methods –** Calculate the distance between pairs of sequences and infer a tree that best describes the distances observed

   e.g. Neighbour joining (NJ)

2. **Character based methods** – Examine each SNV, one at a time, and search for the tree with the best "score"

   e.g. Maximum Likelihood (ML)

Distance based methods of phylogenetic inference

e.g. Neighbour joining

```
Sequence A: ...CGTTAGTACACT...
Sequence B: ...CGATAGTTCACT...
Sequence C: ...CGTTGGTTTACG...
Sequence D: ...CCTTGGTTTACG...
```

1. Recombination filtered alignment

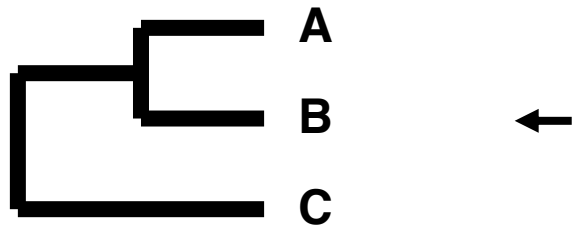
	A	B	C	D
A		1	3	4
B			2	3
C				4
D				

2. Compute distance matrix & correct for different rates

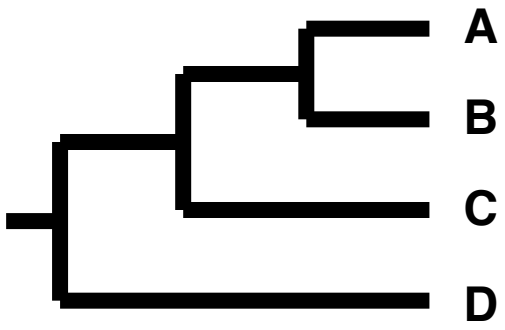
3. Join closest nodes (A & B) & calculate branch lengths

	A&B	C	D
A&B		2	3
C			4
D			

4. Update distance matrix & corrections



5. Join closest nodes (A/B & C) & calculate branch lengths



6. Phylogenetic tree (4-5 repeat until all nodes joined)

Methods of phylogenetic inference

1. **Distance based methods –** Calculate the distance between pairs of sequences and infer a tree that best describes the distances observed

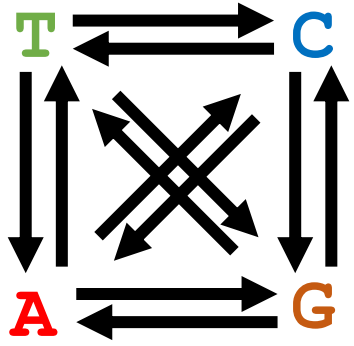
   e.g. Neighbour joining (NJ)

2. **Character based methods** – Examine each SNV, one at a time, and search for the tree with the best "score"

   e.g. Maximum Likelihood (ML)

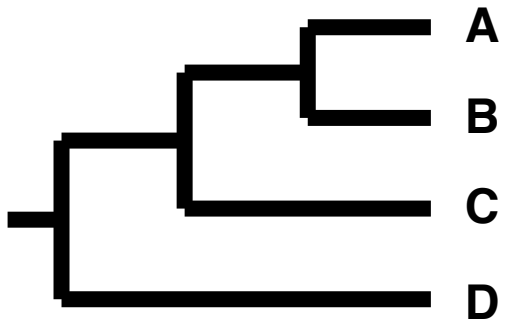
Maximum likelihood (ML) phylogenetic inference

Given



Model of evolution

+



Phylogenetic tree (evolutionary history)

Probability of?

```
Sequence A: ...CGTTAGTACACT...
Sequence B: ...CGATAGTTCACT...
Sequence C: ...CGTTAGTTTACC...
Sequence D: ...CATTGGTTTACT...
```

Recombination filtered alignment (data)

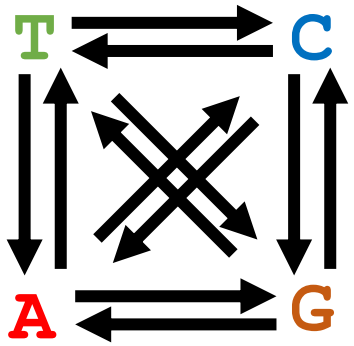
Likelihood *(L)*= *P(D|M)*

the probability *(P)* of the observed data *(D)*, given the model *(M)*

Finds the model that maximises the likelihood of the observed data

Model of evolution (ML)

e.g. Maximum Likelihood



Rate matrix (Q)

+

$$\pi_A + \pi_C + \pi_G + \pi_T = 1$$

Base frequencies

+

$$+ I + G$$

Site rates

Common substitution models include:
GTR, HKY85, JC69, K80

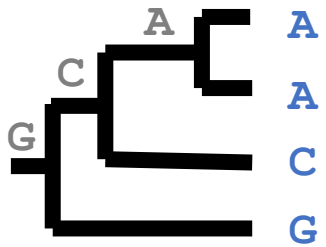
Transitions (Ts) = A<->G, C<->T
Transversions (Tv) = G <-> T, C<->A

ML phylogenetic inference

Recombination filtered alignment

```
                  j
Sequence A: ...CGTTAGTTCACT...
Sequence B: ...CGATAGTTCACT...
Sequence C: ...CGTTCGTTTACT...
Sequence D: ...CGTTGGTTTACT...
```

Examine each SNV site (e.g. site *j*)



Infer every possible tree for site *j*

$$L_j = P\left[\begin{matrix}A\\A\\C\\G\end{matrix}\right] * P\left[\begin{matrix}A\\A\\C\\G\end{matrix}\right] * P\left[\begin{matrix}C\\A\\A\\G\end{matrix}\right] * P\left[\begin{matrix}G\\A\\C\\A\end{matrix}\right] * \ldots$$

Calculate likelihood for site (L_j)

Likelihood of the tree is the product of likelihoods for each site:

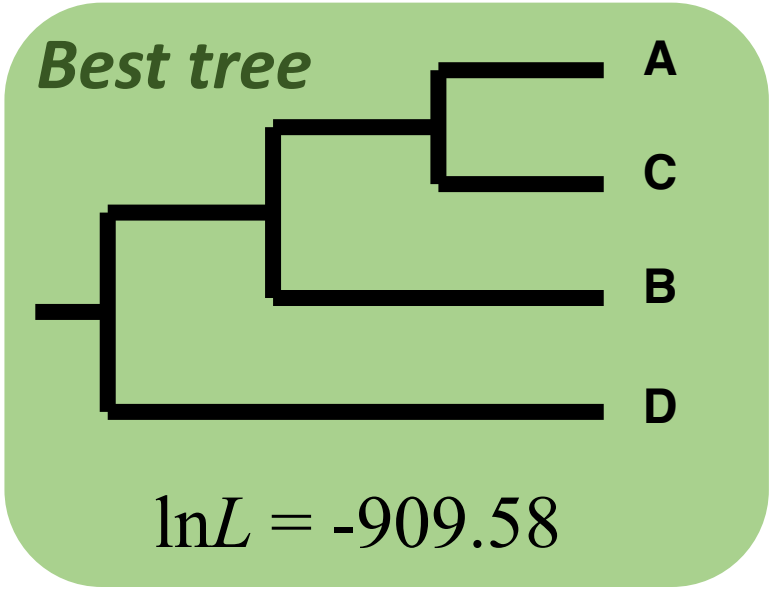
$$L = L_1 * L_2 * L_3 * \ldots * L_n$$

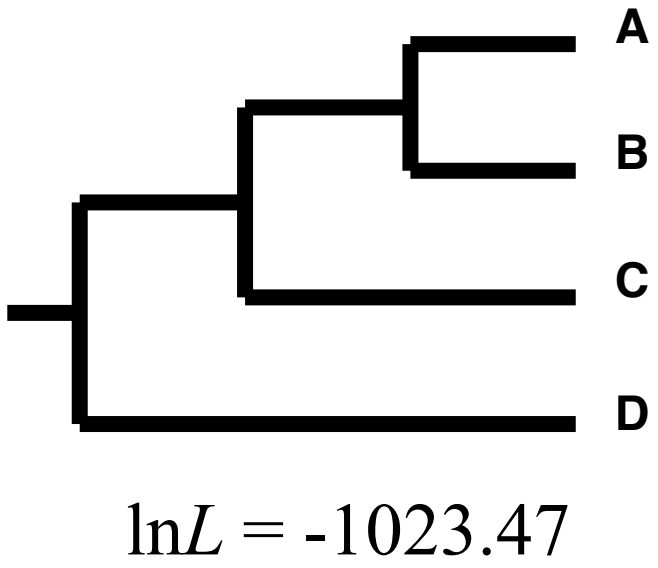
Usually evaluated as the sum of the log (ln) likelihoods (L):

$$\ln L = \ln L_1 + \ln L_2 + \ln L_3 + \ldots + \ln L_n$$

ML phylogenetic inference – the best tree

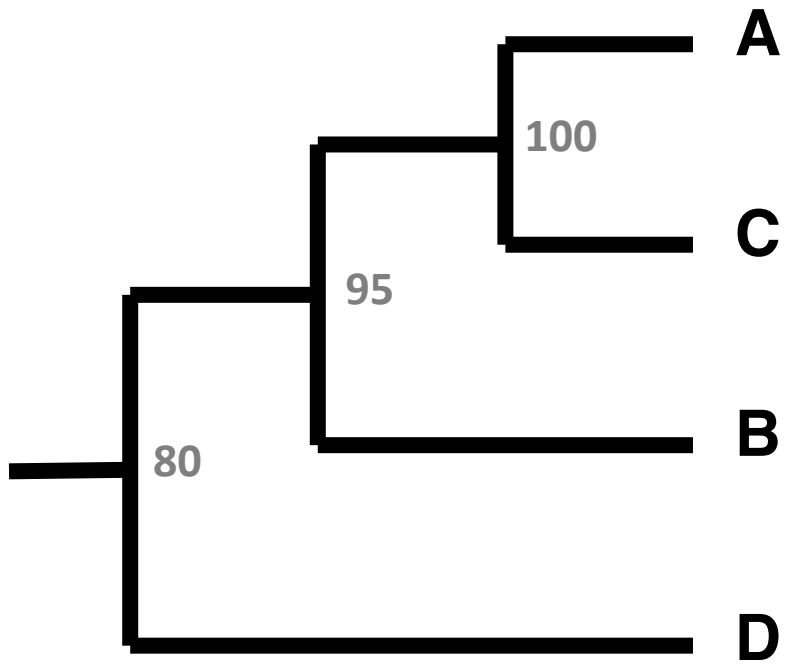
- The most likely (best) tree has the highest likelihood (score)
- ML evaluates all possible ancestral states, at each site, for all tree topologies
- In practice, most ML tree inference software implementations use other heuristics to keep inference computationally tractable





ML bootstrap support values

- In order to understand the confidence of the inferred tree topology we compute **bootstrap support** values
- For this we pseudo-sample (randomly subsample) our alignment data (by site, with replacement) & run many **pseudoreplicate** trees (usually 100)
- We then assess how many times the same clades are inferred



Grey numbers are bootstrap support values.

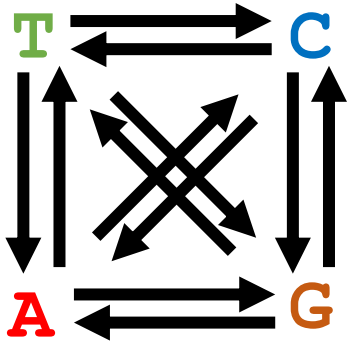
Sequences A & C were clustered together 100% of the time in all pseudoreplicates.

Inferring ML trees with software

```
Sequence A: ...CGTTAGTACACT...
Sequence B: ...CGATAGTTCACT...
Sequence C: ...CGTTAGTTTACC...
Sequence D: ...CATTGGTTTACT...
```

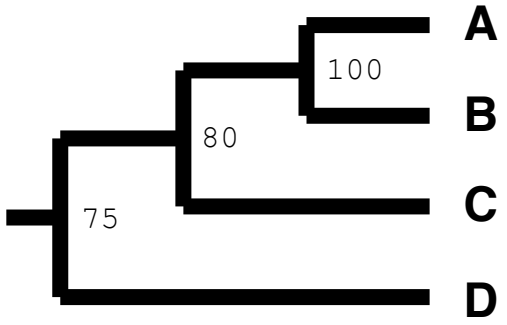
Recombination filtered alignment

+



Model of evolution

Tree inference software
e.g. RAxML



Newick tree file

```
(D:60,(C:40,(A:20,B:20)100:20)80:20)75:10
```

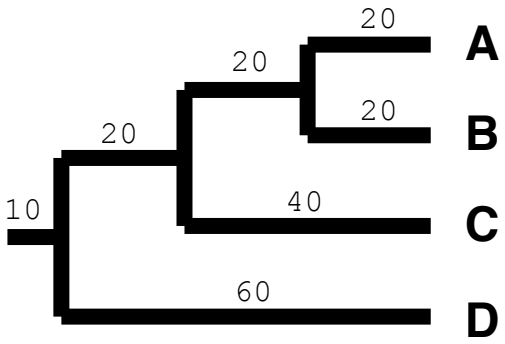
RAxML: Stamatakis 2014, Bioinformatics

Newick tree file format

Trees are usually output & stored in a Newick file (filename.nwk) - a small text file containing all the information required to construct the graph. Newick files can be opened & visualized with interactive tools e.g. FigTree.

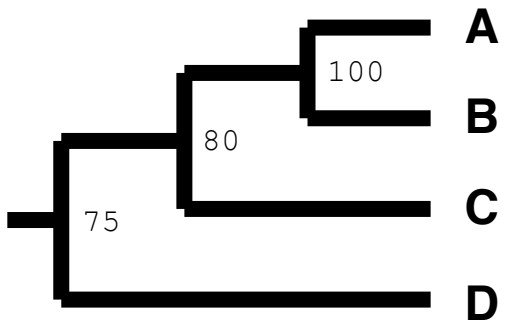
e.g. for a rooted tree:

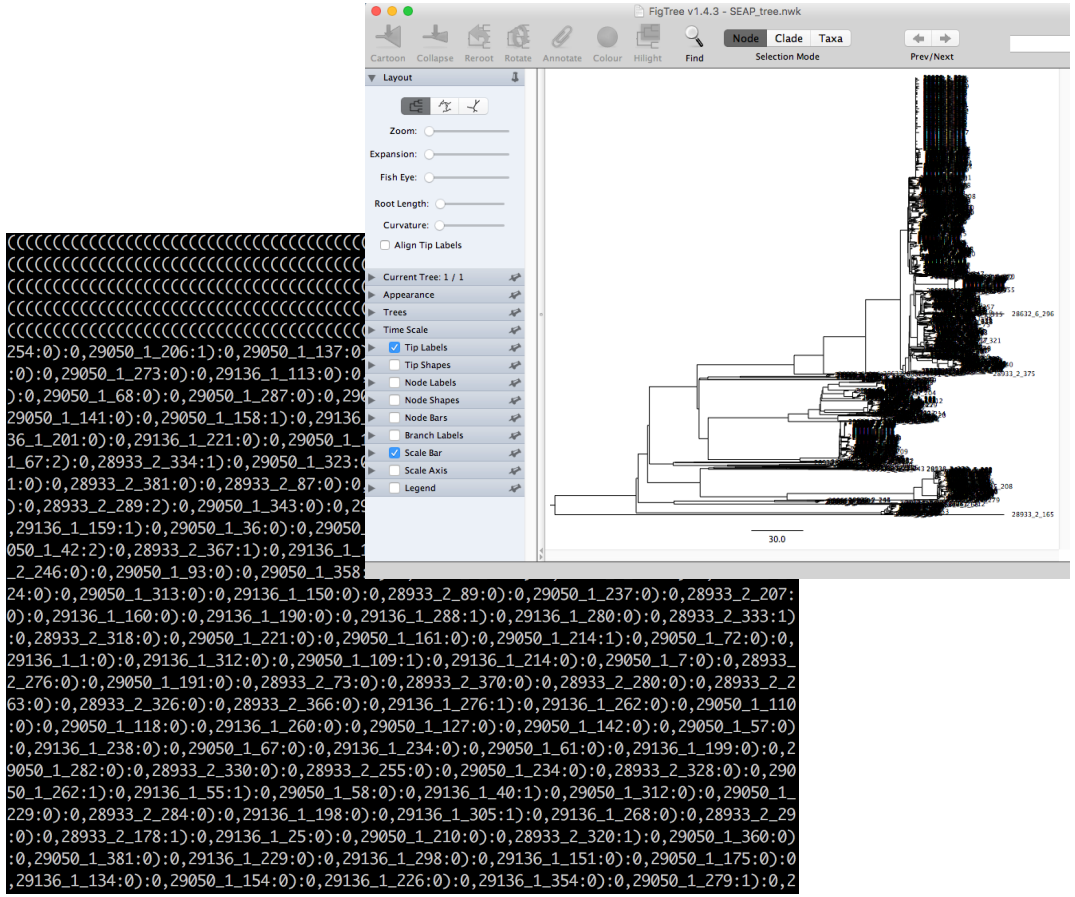
```
(D:60,(C:40,(A:20,B:20):20):20):10);
```



e.g. for the same rooted tree with bootstrap values

```
(D:60,(C:40,(A:20,B:20)100:20)80:20)75:10);
```





http://tree.bio.ed.ac.uk/software/figtree/

Distance based vs. character based methods

Method	Advantages	Disadvantages
Distance based e.g. Neighbour Joining	• Fast • Computationally efficient • Good for large datasets • Good where distances are small • One tree inferred	• Information loss when converting characters to distances • May be inaccurate for large distances • Branch lengths not interpretable • One tree inferred
Character based e.g. Maximum Likelihood	• Accurate • Explicit & more complex models • Suitable for dissimilar sequences • Branch lengths interpretable • Bootstrapping for confidence	• Slow • Requires lots of computational resources

Join at menti.com | use code 8692 7095

Mentimeter

Multiple ChoiceWhat method would you use for a final analysis?

Distance based (e.g. Neighbour Joining)	Character based (e.g. Maximum Likelihood)	Not sure

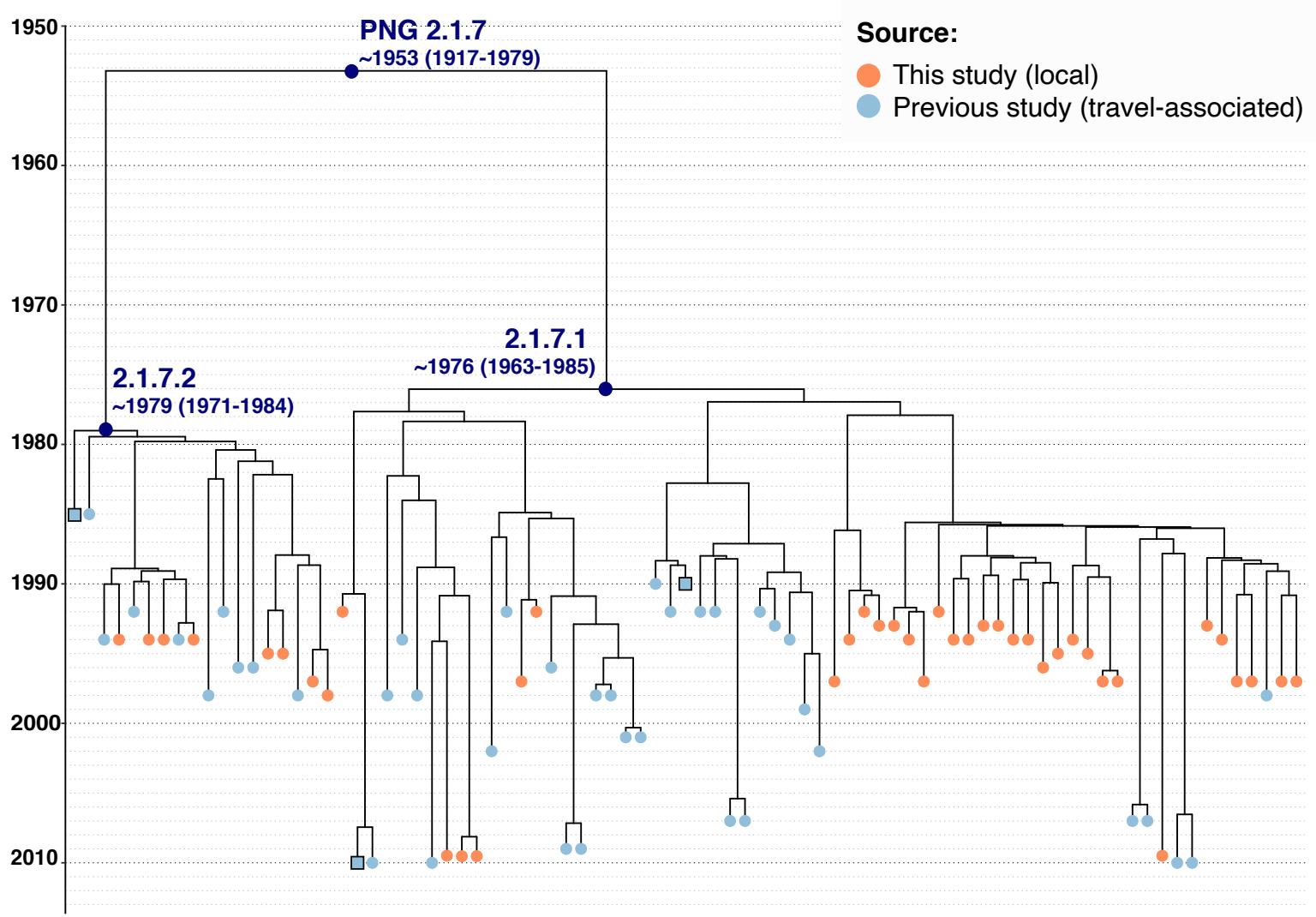
Intended learning outcomes

1. Recognise the basic principles of phylogenetics
2. Interpret data on a phylogenetic tree
3. Explain the methods used to infer a phylogenetic tree from bacterial pathogen whole genome sequencing data
4. Explain core concepts related to phylodynamics and how these can provide insights into pathogen evolution and epidemiology

What is phylodynamics?

- **Phylodynamics** involves linking mathematical epidemiology & statistical phylogenetics in order to study the **dynamics of pathogen evolution and epidemiology**

- Phylodynamics often involves inferring a **dated phylogeny** and modeling various aspects of the pathogen population under study



Dyson *et al.* 2022, PLoS NTDs

The molecular clock and temporal signal

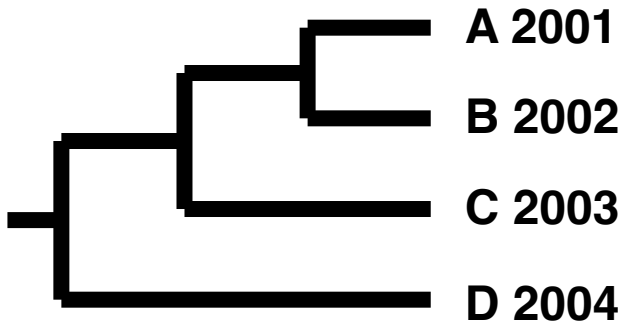
- The **molecular clock** is the concept that evolution occurs at a measurable rate over time which can then be used to infer the dates of past events

- When analysing bacterial pathogen sequence data we commonly **calibrate the molecular clock** with **sampling/tip dates** (the isolation dates for a sequenced samples). A molecular clock model can be **strict** (substitution rate is constant) or **relaxed** (allows for substitution rates to vary among lineages)

- **Temporal signal** refers to a sufficient accumulation of genetic change between sampling times to recognize the relationship between genetic divergence and time

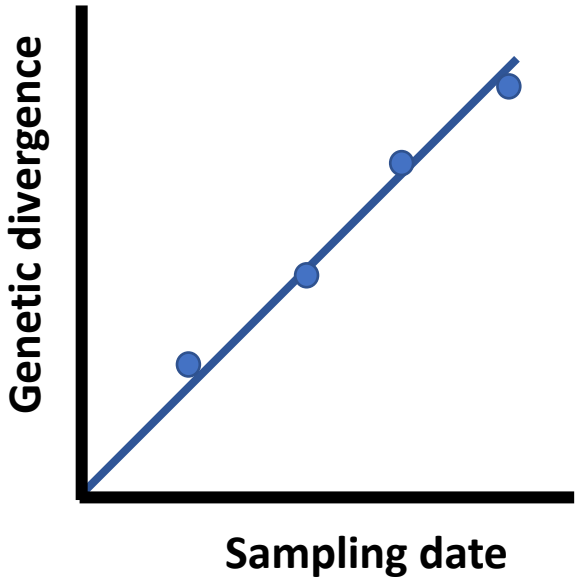
- A **root-to-tip regression analysis** can be used to carry out a preliminary evaluation of a temporal signal across a dataset. This analysis can be used to determine if it is appropriate to attempt to infer a dated phylogeny

Root-to-tip regression analysis



Input: ML phylogeny & sampling dates

Root-to-tip regression software
e.g. TempEst



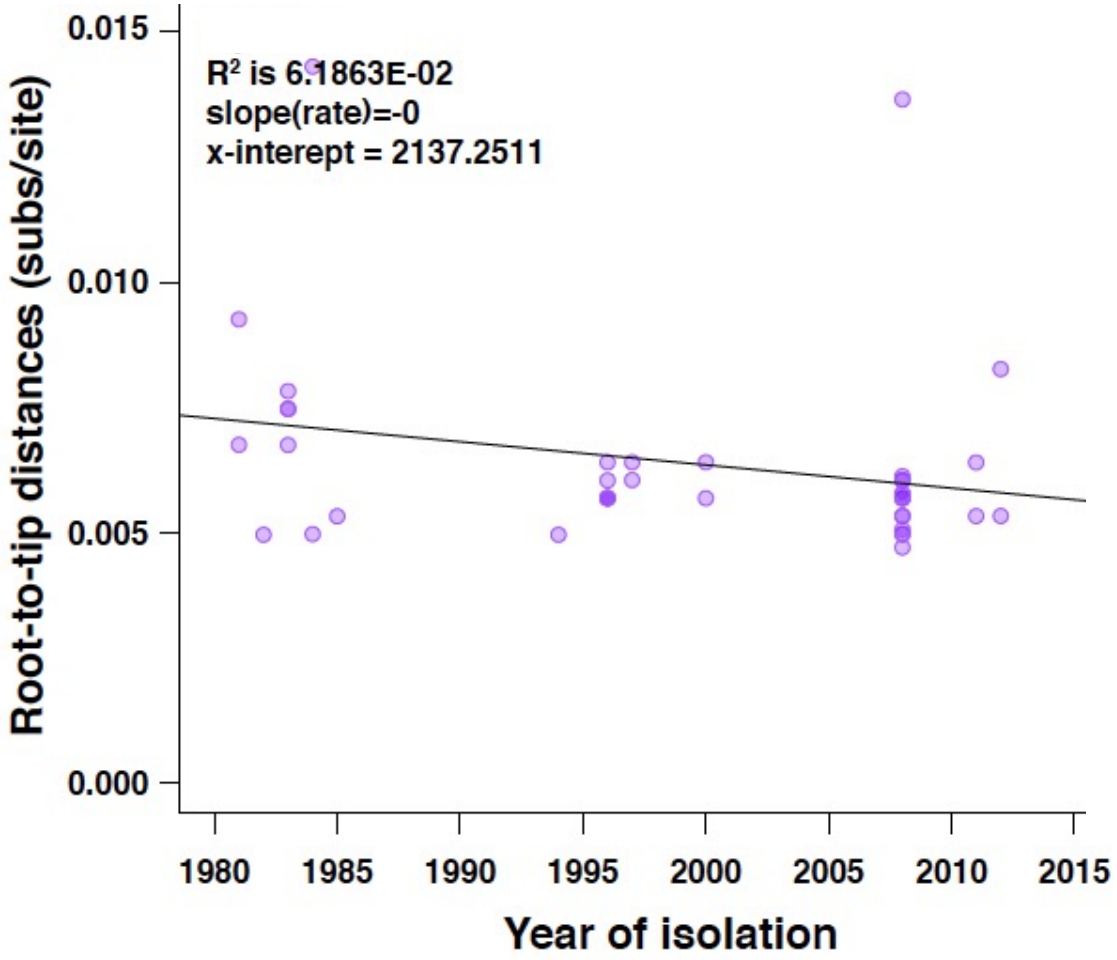
Output: Regression model

Regression analysis of genetic divergence (branch lengths from input ML tree) and sampling (isolation) dates

TempEst: Rambaut et al. 2016, Virus Evol.

Root-to-tip regression analysis

No temporal signal!

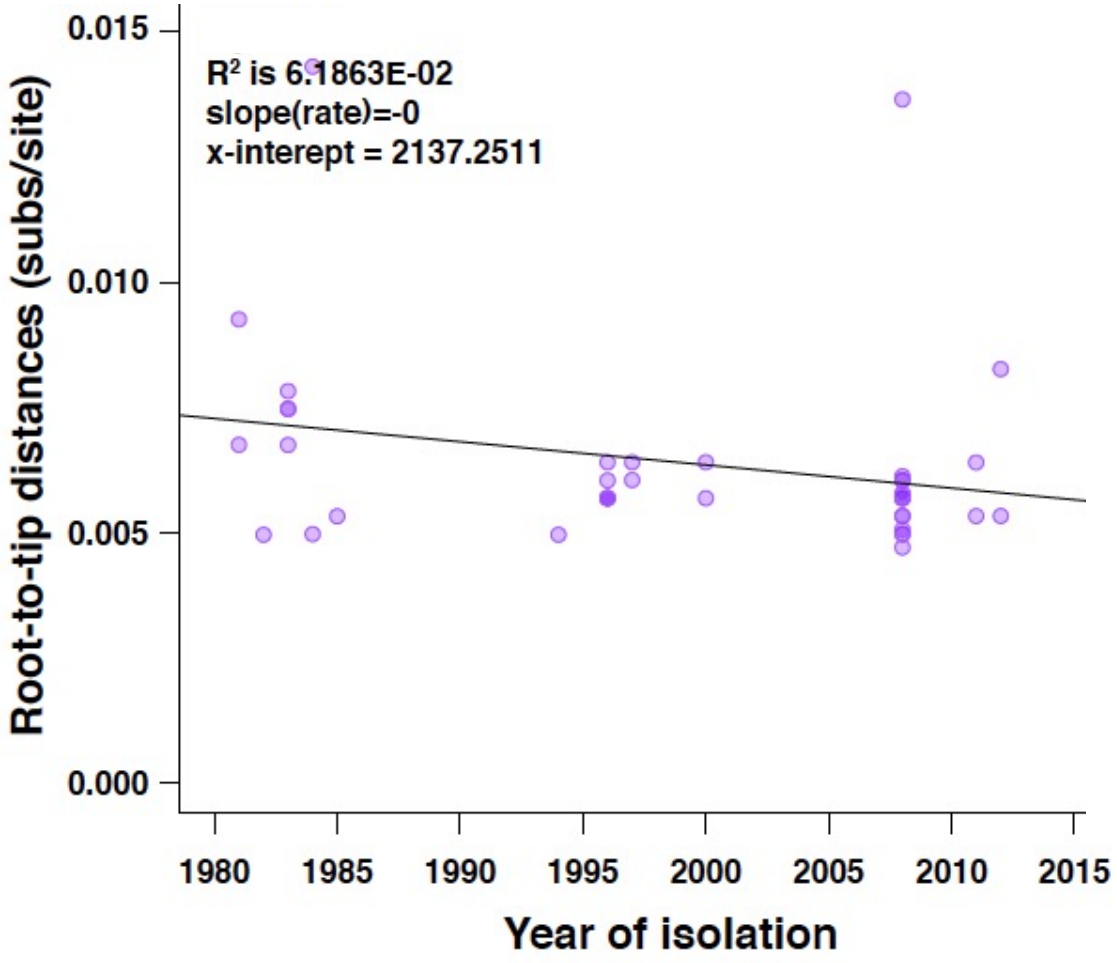


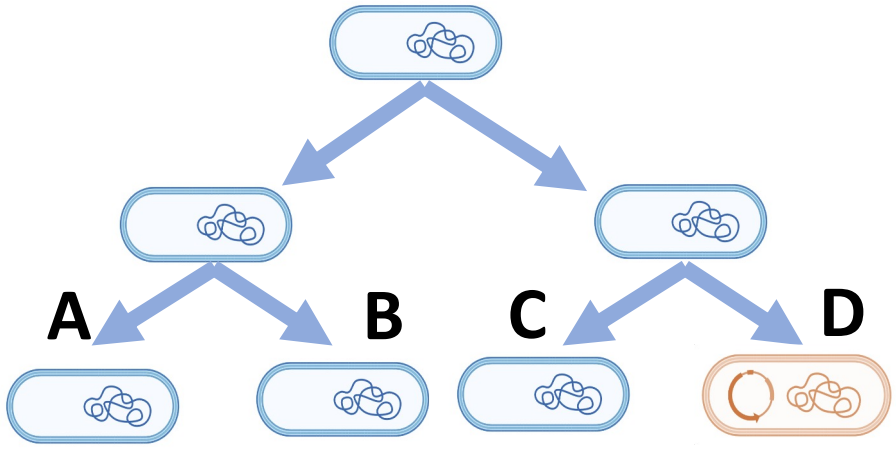
x-intercept = divergence date/age of the root node
slope = crude estimate of the **substitution rate** (usually substitutions/site/year)

Britto/Dyson *et al.* 2018, PLoS NTDs
TempEst: Rambaut et al. 2016, Virus Evol.

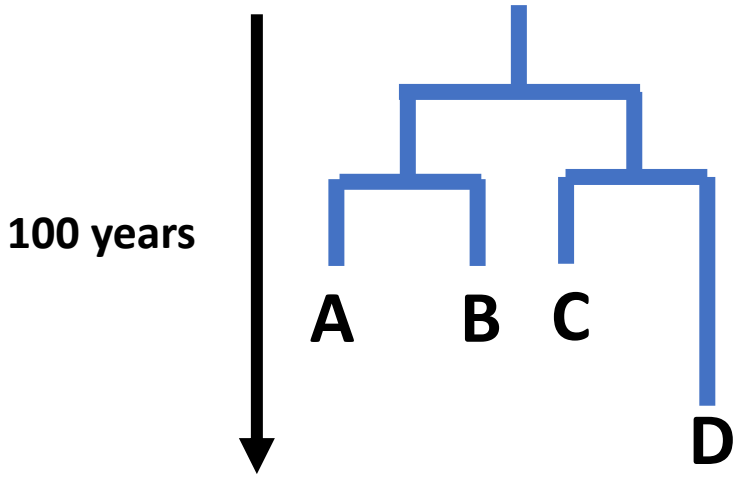
Root-to-tip regression analysis

No temporal signal!





D horizontally acquires DNA

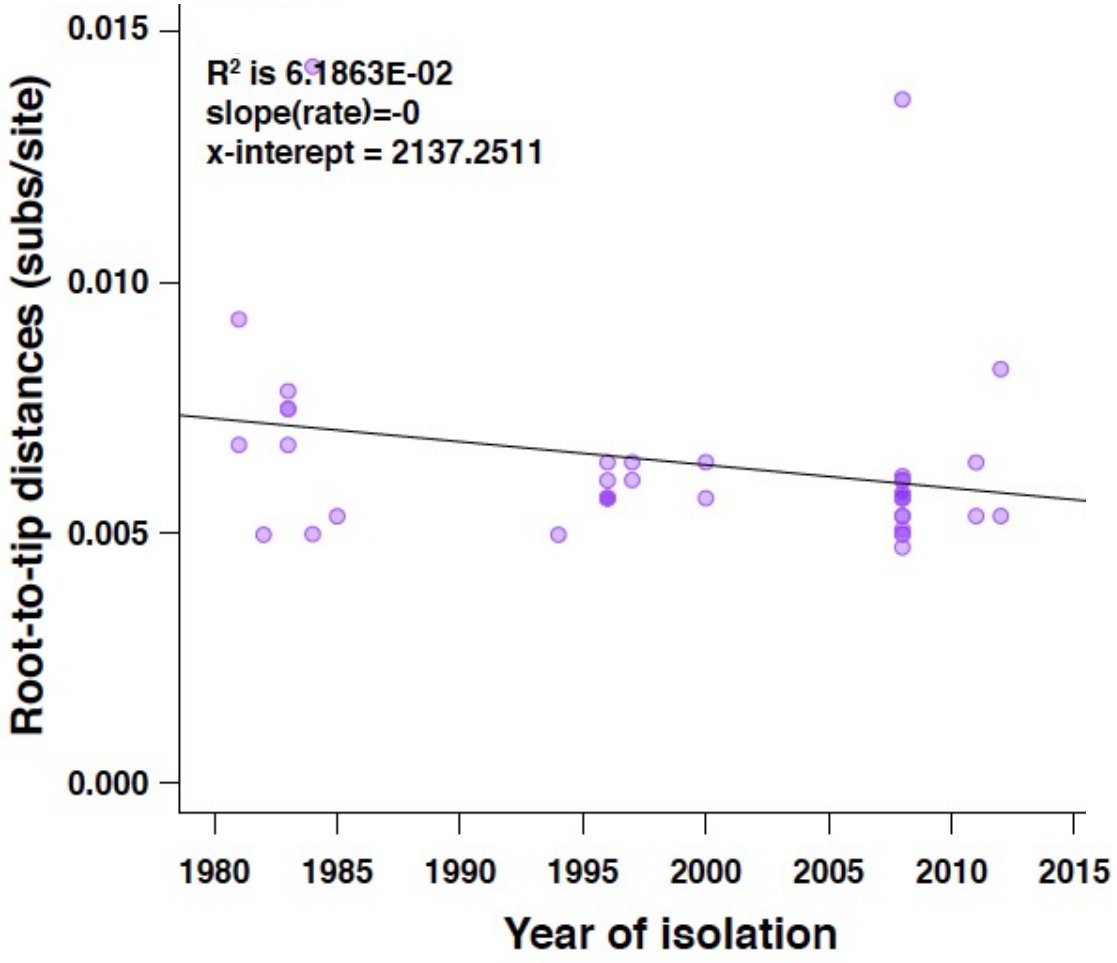


Destroys temporal signal!

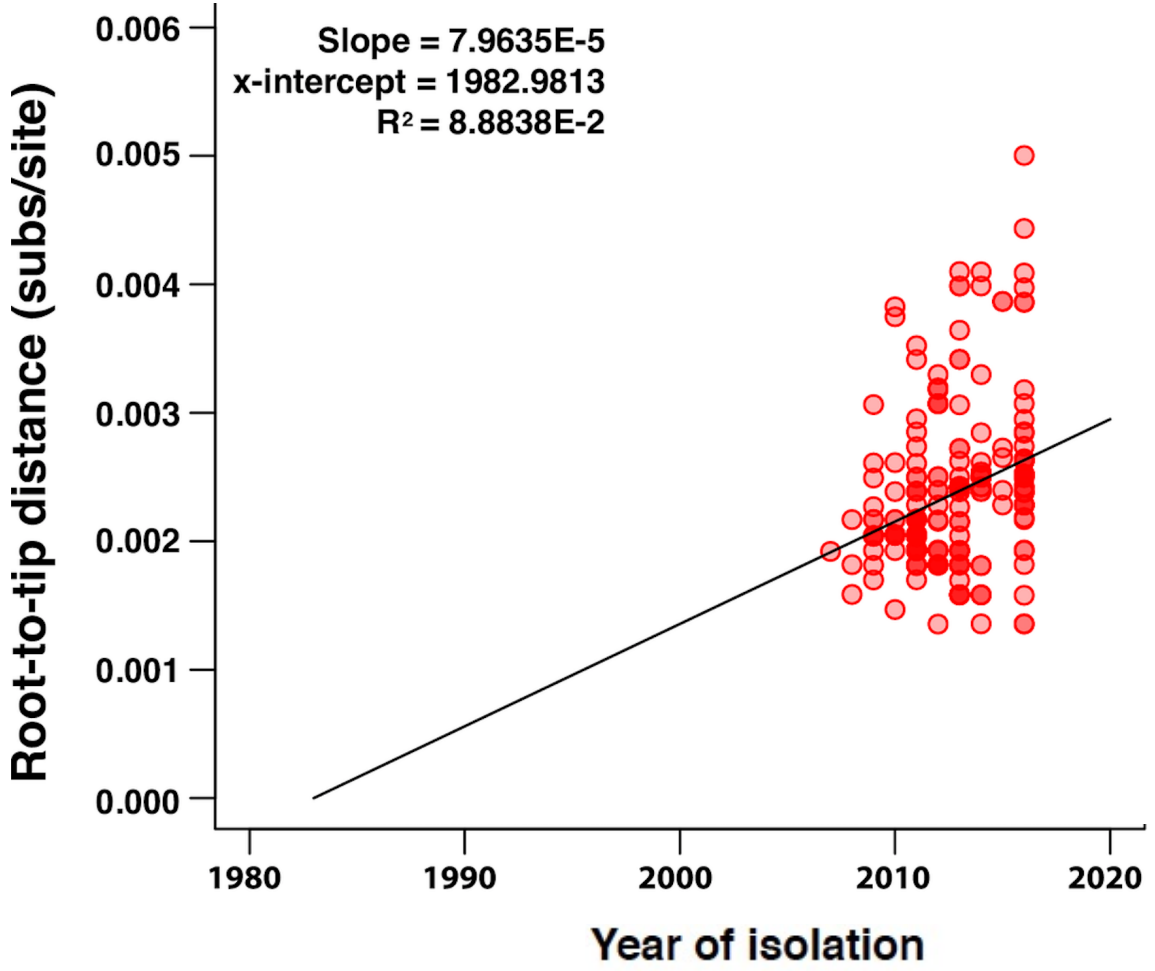
x-intercept = divergence date/age of the root node
slope = crude estimate of the **substitution rate** (usually substitutions/site/year)

Root-to-tip regression analysis

No temporal signal!



Temporal signal!



x-intercept = divergence date/age of the root node
slope = crude estimate of the **substitution rate** (usually substitutions/site/year)

Britto/Dyson *et al.* 2018, PLoS NTDs
TempEst: Rambaut et al. 2016, Virus Evol.

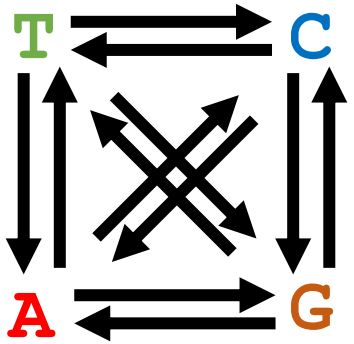
Bayesian phylogenomic inference

Given

Probability of?

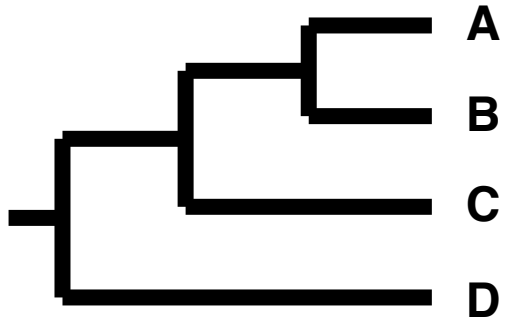
```
Sequence A: ...CGTTAGTACACT...
Sequence B: ...CGATAGTTCACT...
Sequence C: ...CGTTAGTTTACC...
Sequence D: ...CATTGGTTTACT...
```

Recombination filtered alignment (data)



Model of evolution

+



Phylogenetic tree (evolutionary history)

P(M|D)

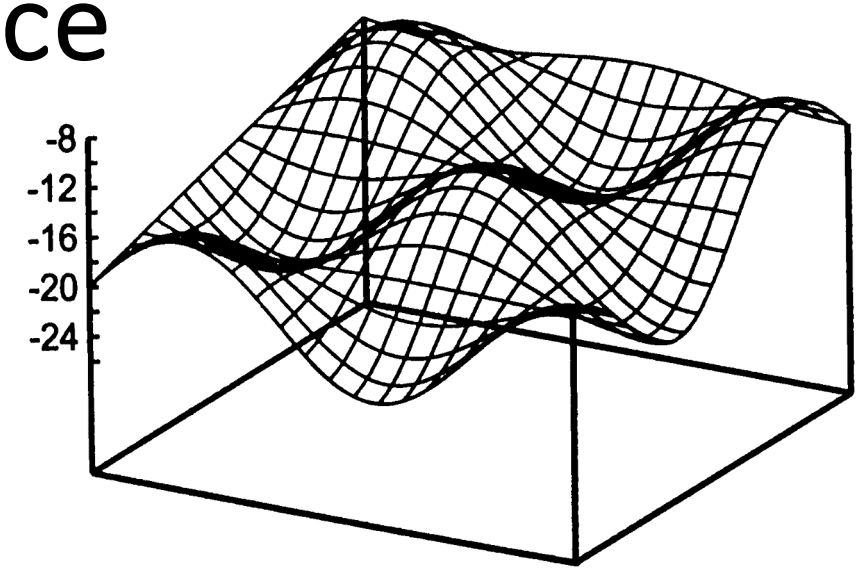
updated probability *(P)* of the model parameters *(M)*
in light of the observed data *(D)*

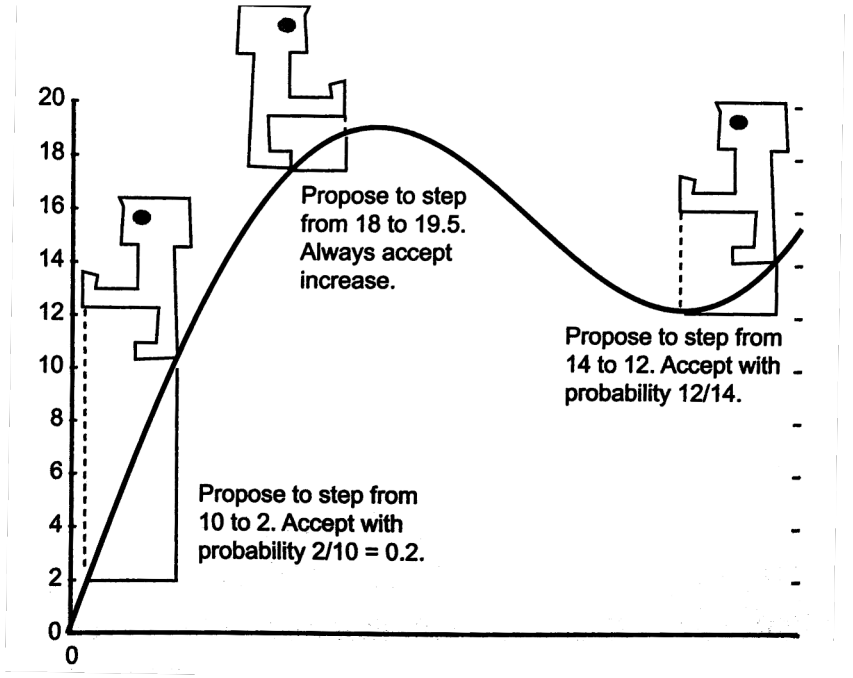
Bayesian phylogenomic inference

- Each model parameter has a **prior distribution**, which is set before the data are observed by the model

- The **posterior distribution** contains the results (e.g. trees, substitution rate estimates)

- The posterior is obtained via the **Markov chain Monte Carlo** (MCMC) sampling technique which improves the model through a series of **steps**





Drummond & Bouckaert, 2015

Bayesian dated phylogenetic inference

BMC Evolutionary Biology

Software

BEAST: Bayesian evolutionary analysis by sampling trees

Alexei J Drummond*[1,2] and Andrew Rambaut[3]

Address: [1]Bioinformatics Institute, University of Auckland, Auckland, New Zealand, [2]Department of Computer Science, University Auckland, New Zealand and [3]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK

Email: Alexei J Drummond* - alexei@cs.auckland.ac.nz; Andrew Rambaut - a.rambaut@ed.ac.uk

* Corresponding author

OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

BEAST 2: A Software Platform for Bayesian Evolutionary Analysis

Remco Bouckaert[1]*, Joseph Heled[1], Denise Kühnert[1,2], Tim Vaughan[1,3], Chieh-Hsi Wu[1], Dong Xie[1], Marc A. Suchard[4,5], Andrew Rambaut[6], Alexei J. Drummond[1,7]*

1 Computational Evolution Group, Department of Computer Science, University of Auckland, Auckland, New Zealand, **2** Department of Environmental Systems Science, ETH Zürich, Zürich, Switzerland, **3** Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand, **4** Departments of Biomathematics and Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, California, United States of America, **5** Department of Biostatistics, School of Public Health, University of California, Los Angeles, Los Angeles, California, United States of America, **6** Institute of Evolutionary of Edinburgh, Edinburgh, United Kingdom, **7** Allan Wilson Centre for Molecular Ecology and Evolution, University of Auckland, Auckland, New Zealand

Published online 3 September 2018

Nucleic Acids Research, 2018, Vol. 46, No. 22 e134
doi: 10.

Bayesian inference of ancestral dates on bacte phylogenetic trees

Xavier Didelot[1,*], Nicholas J. Croucher[1], Stephen D. Bentley[2], Simon R. Harris[2] J. Wilson[3]

[1]Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, L
[2]The Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK and [3]Big Data
Nuffield Department of Population Health, University of Oxford, Oxford, UK

BIOINFORMATICS **APPLICATIONS NOTE** *Vol. 17 no. 8 2001*
Pages 754–755

MRBAYES: Bayesian inference of phylogenetic trees

John P. Huelsenbeck[1] and Fredrik Ronquist[2]

[1]Department of Biology, University of Rochester, Rochester, NY 14627, USA and
[2]Department of Systematic Zoology, Evolutionary Biology Centre, Uppsala University, Norbyv. 18D, SE-752 36 Uppsala, Sweden

Character based method of phylogenetic inference using a **Bayesian** framework. Many software tools available; BEAST & BEAST2 are popular.

BEAST2 (Bayesian evolutionary analysis by sampling trees)

OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

BEAST 2: A Software Platform for Bayesian Evolutionary Analysis

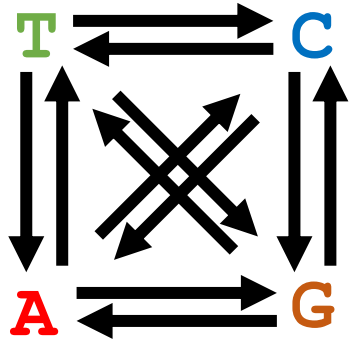
Remco Bouckaert[1]*, Joseph Heled[1], Denise Kühnert[1,2], Tim Vaughan[1,3], Chieh-Hsi Wu[1], Dong Xie[1], Marc A. Suchard[4,5], Andrew Rambaut[6], Alexei J. Drummond[1,7]*

1 Computational Evolution Group, Department of Computer Science, University of Auckland, Auckland, New Zealand, **2** Department of Environmental Systems Science, ETH Zürich, Zürich, Switzerland, **3** Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand, **4** Departments of Biomathematics and Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, California, United States of America, **5** Department of Biostatistics, School of Public Health, University of California, Los Angeles, Los Angeles, California, United States of America, **6** Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, United Kingdom, **7** Allan Wilson Centre for Molecular Ecology and Evolution, University of Auckland, Auckland, New Zealand

Suite of Graphical User Interface (GUI) software tools for inferring & evaluating Bayesian phylogenies:
BEATUi, BEAST2, Tracer & TreeAnnotator

BEAST2 workflow for inferring dated phylogenies

```
Sequence A 2001: ...CGTTAGTACACT...
Sequence B 2002: ...CGATAGTTCACT...
Sequence C 2003: ...CGTTAGTTTACC...
Sequence D 2004: ...CATTGGTTTACT...
```



BEAUTi (model) input:

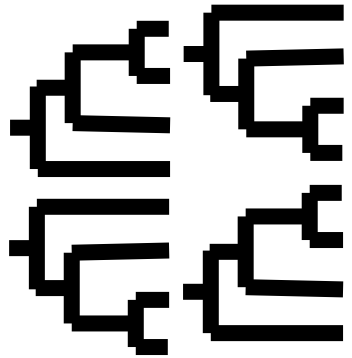
- Recombination filtered alignment
- Sampling dates
- Model **prior** distributions
  - Clock
  - Demographic
  - Substitution
  - …

BEAUTi xml file

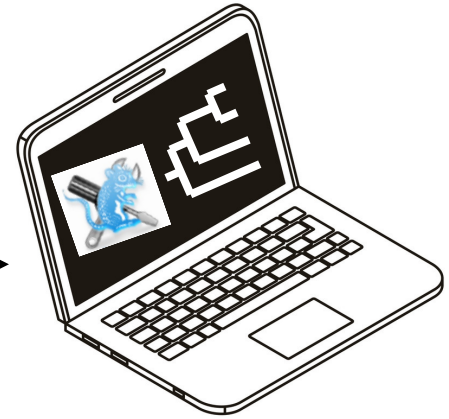
BEAST2
Runs model from xml file using MCMC

Log file
(sampled from MCMC)

Examine posterior distributions in Tracer

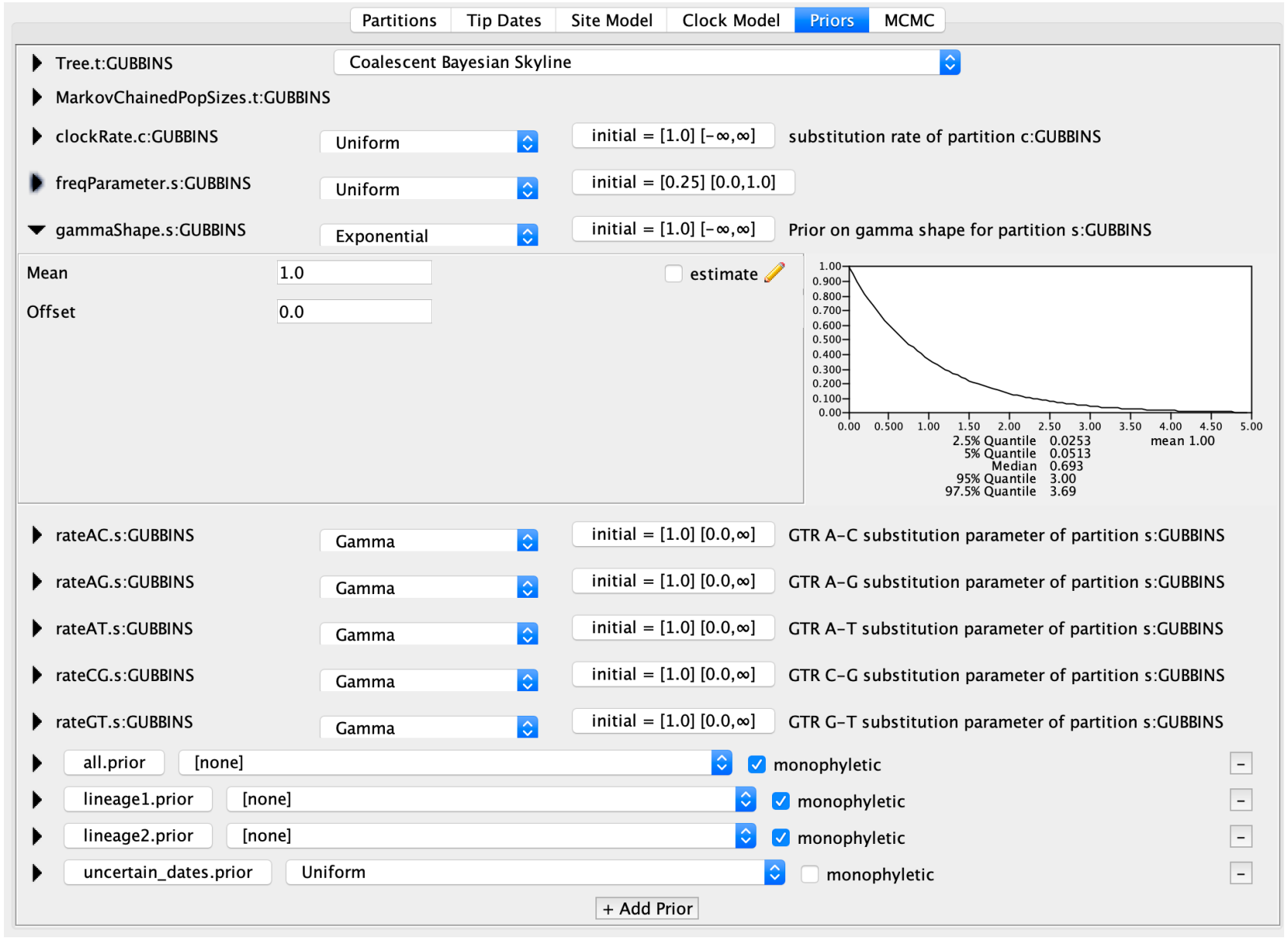


Trees file containing 100s-1000s of trees



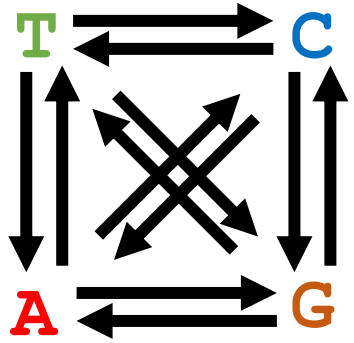
Infer Maximum Clade Credibility (MCC) tree with TreeAnnotator

BEAUTi: BEAST2 model setup (create xml file)



BEAST2 workflow for inferring dated phylogenies

```
Sequence A 2001: ...CGTTAGTACACT...
Sequence B 2002: ...CGATAGTTCACT...
Sequence C 2003: ...CGTTAGTTTACC...
Sequence D 2004: ...CATTGGTTTACT...
```



BEAUTi (model) input:

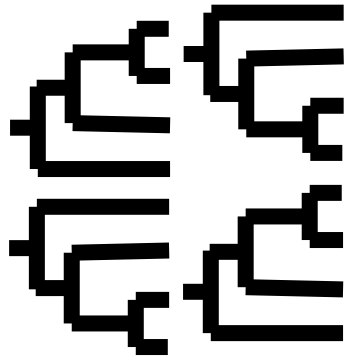
- Recombination filtered alignment
- Sampling dates
- Model **prior** distributions
  - Clock
  - Demographic
  - Substitution
  - …

BEAUTi xml file

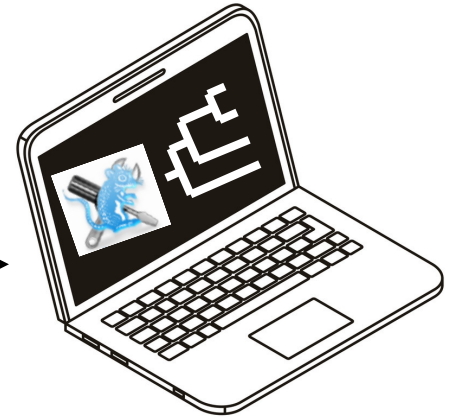
BEAST2
Runs model from xml file using MCMC

Log file
(sampled from MCMC)

Examine posterior distributions in Tracer



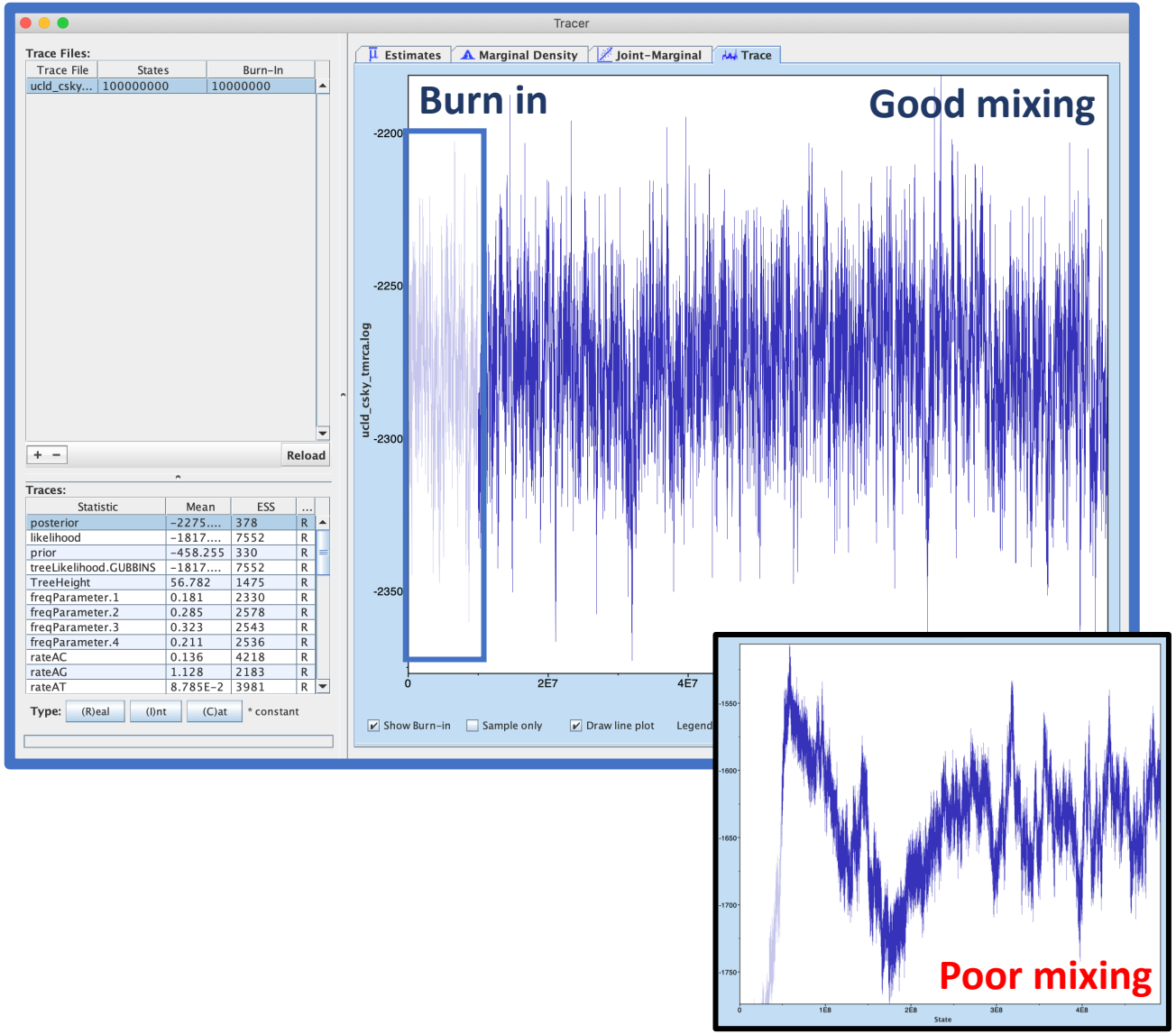
Trees file containing 100s-1000s of trees



Infer Maximum Clade Credibility (MCC) tree with TreeAnnotator

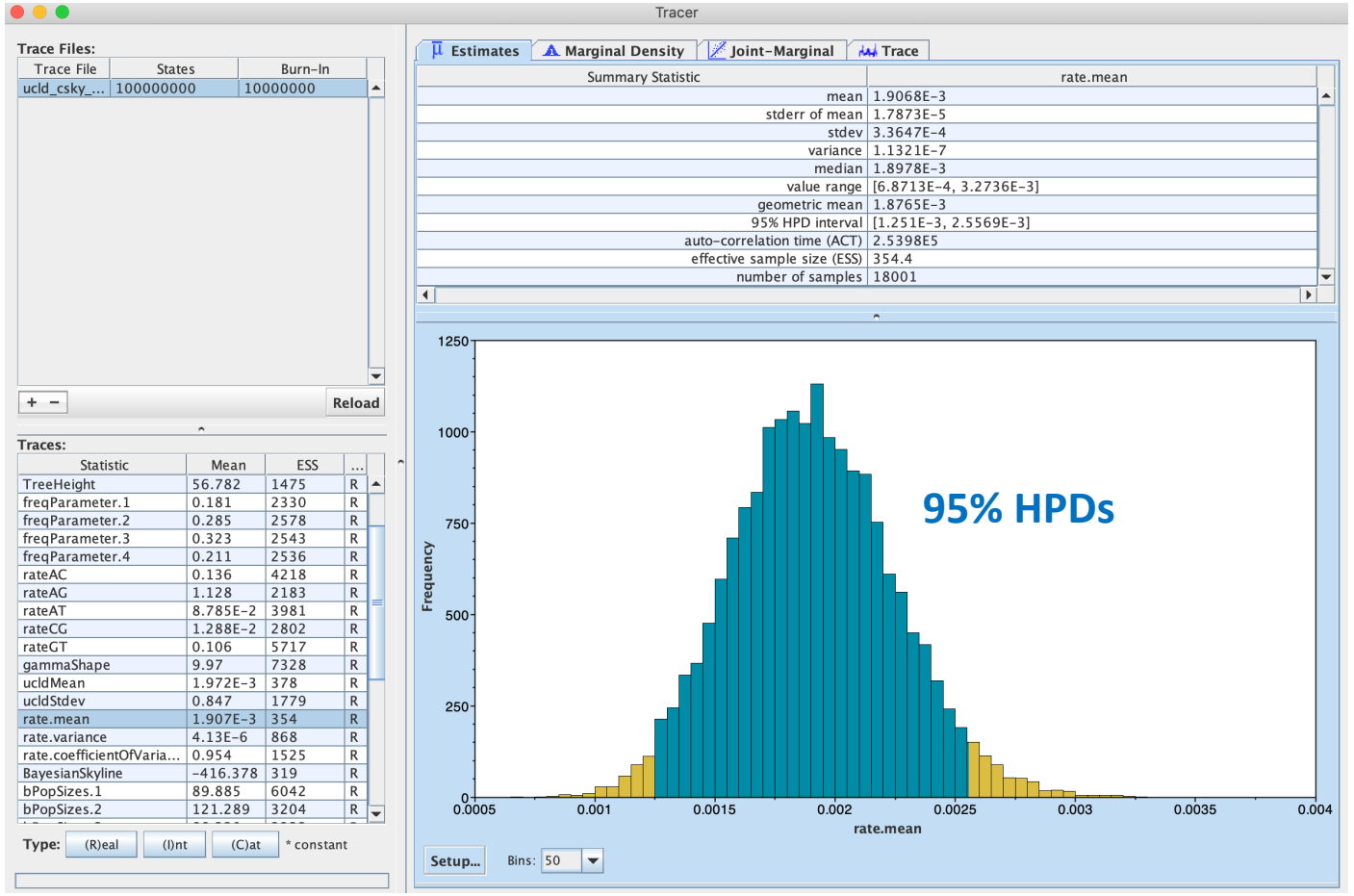
Tracer: Examine BEAST2 run posterior distribution

- **Tracer** in the BEAST2 package is used to view the MCMC trace
- The start of this (usually 10%) is removed as '**burn in**' as the trace won't have stabilized
- Ideally, the trace should look like a '**hairy caterpillar**' which is indicative of **good mixing** & **convergence**



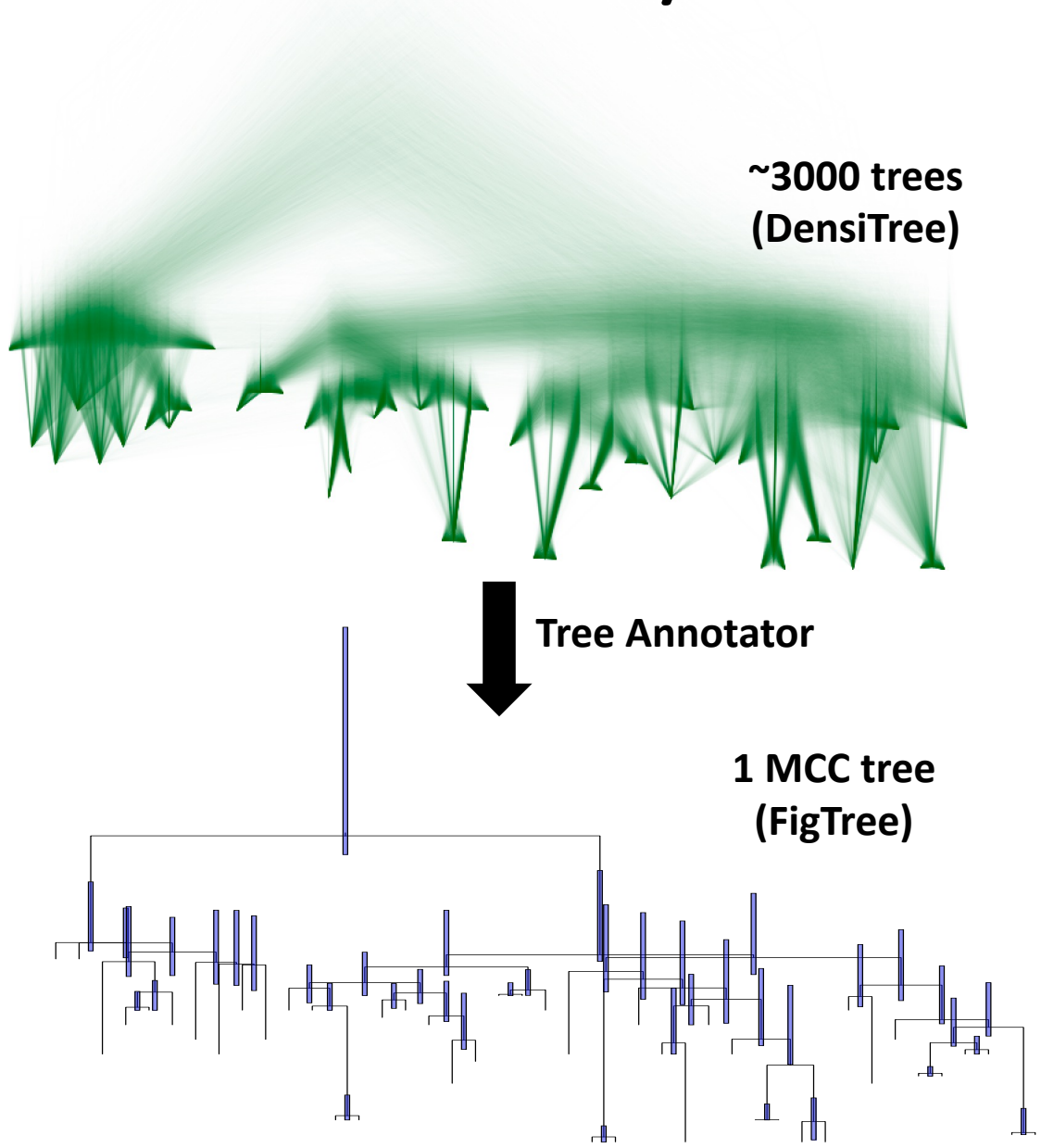
Tracer: BEAST2 95% HPDs for estimates

- **Tracer** can also be used to look at the estimates of different parameters
- Mean estimates are provided along with 95% Highest Posterior Densities (95% HPDs)
- **95% HPDs** are the smallest interval containing 95% of the posterior probability i.e. a credible interval
- **Effective Sample Sizes** (ESS) should be at least 200 for relevant parameters



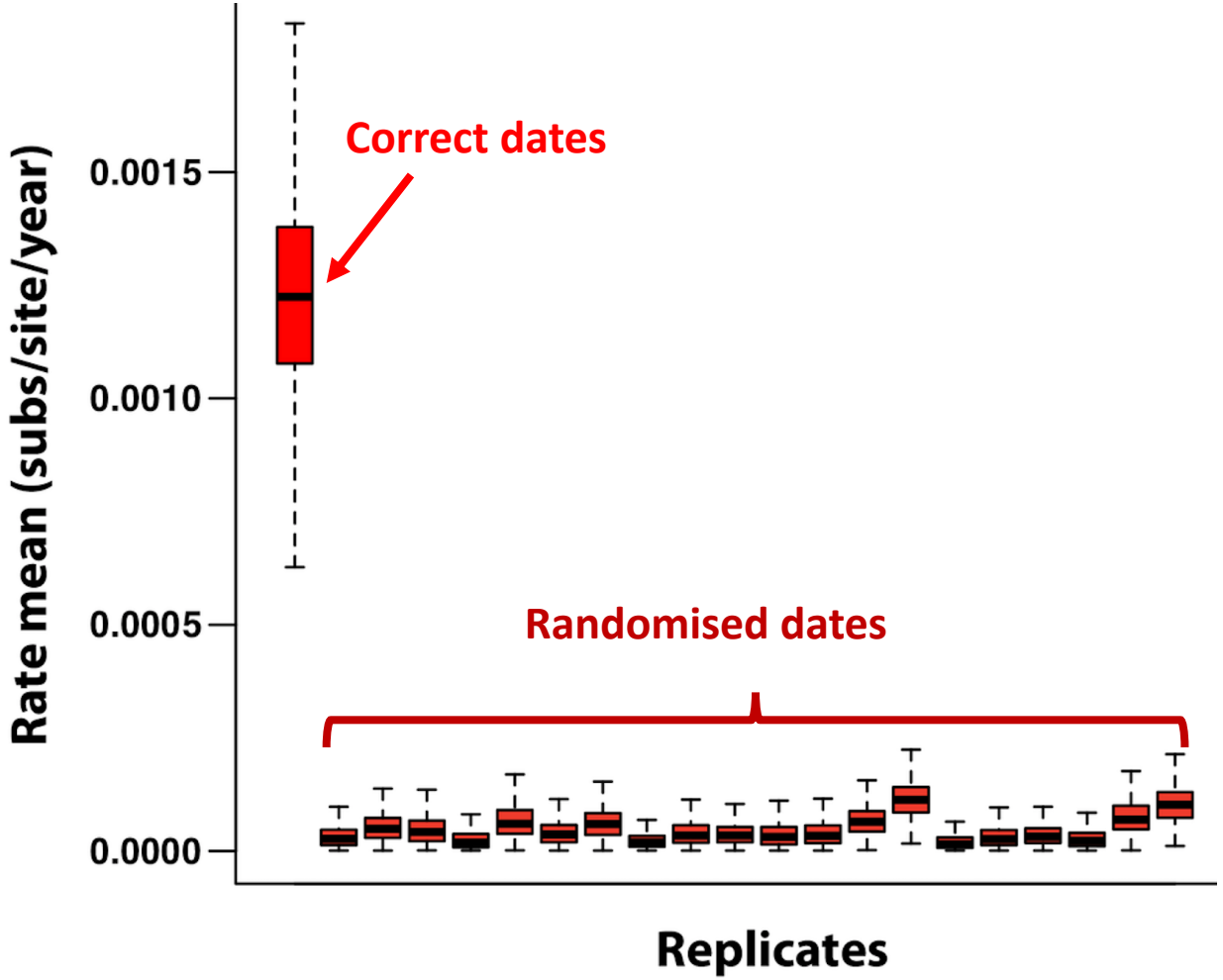
Tree Annotator: Maximum Clade Credibility tree

- **BEAST2** runs produce many trees, sometimes thousands
- TreeAnnotator allows us to summarise many BEAST2 trees into a single **Maximum Clade Credibility (MCC)**
- The MCC tree is a tree within the set of inferred trees, often the **median tree**
- Estimates and associated 95% HPDs for date estimates are summarized across all trees and annotated onto the MCC tree



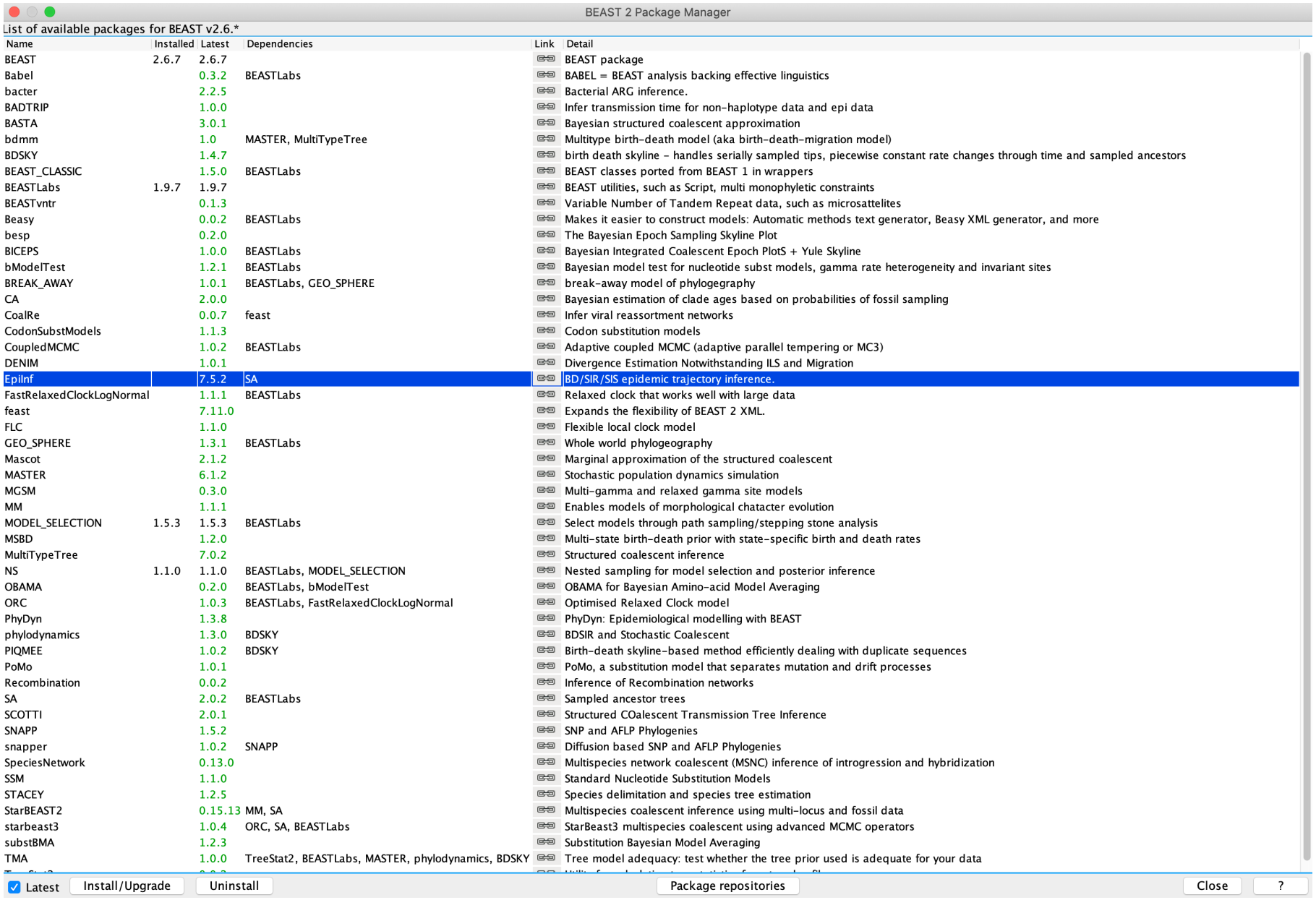
Date-randomization test to confirm temporal signal

- Randomise dates for each taxa
- Re-run analysis multiple times (usually 10-20 times) with the dates randomized differently each time
- Compare **mean substitution rate estimates** for the run with the correct dates to those with randomly assigned dates
- If **sufficient temporal signal** exists, the rate estimates for the run with the correct dates should be non-overlapping with runs with randomized dates



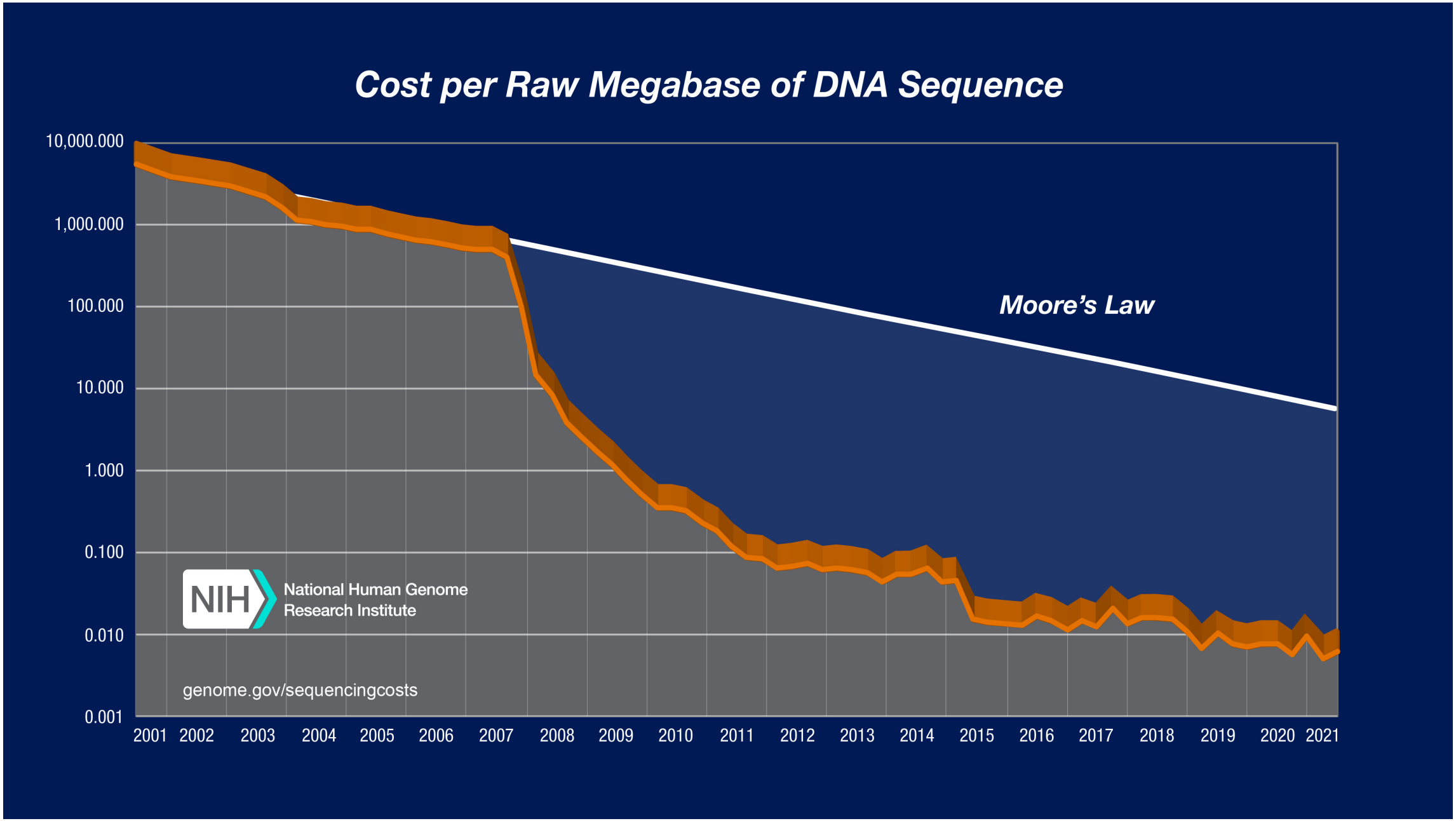
Britto/Dyson *et al.* 2018, PLoS NTDs
Duchene *et al*. 2016, Microb Genom

Many packages extending BEAST/BEAST2 exist

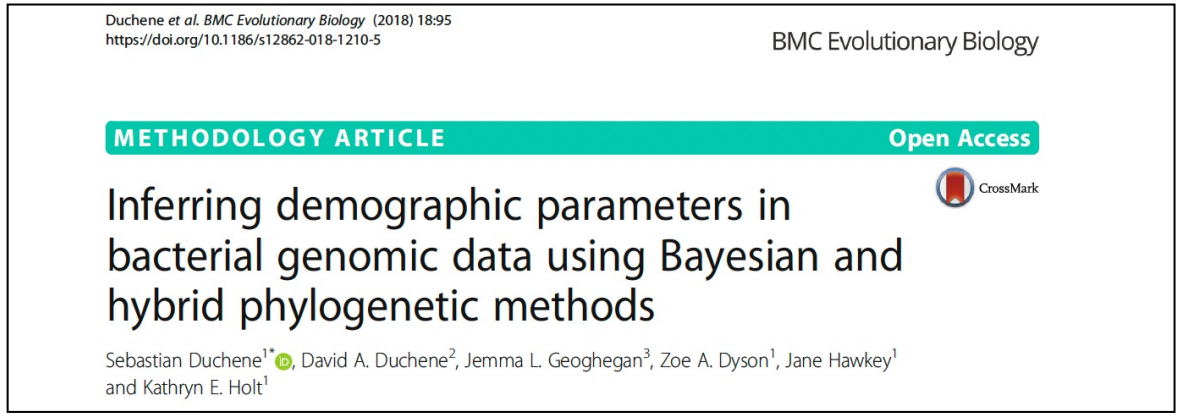


Name	Installed	Latest	Dependencies	Link	Detail
BEAST	2.6.7	2.6.7			BEAST package
Babel		0.3.2	BEASTLabs		BABEL = BEAST analysis backing effective linguistics
bacter		2.2.5			Bacterial ARG inference.
BADTRIP		1.0.0			Infer transmission time for non-haplotype data and epi data
BASTA		3.0.1			Bayesian structured coalescent approximation
bdmm		1.0	MASTER, MultiTypeTree		Multitype birth-death model (aka birth-death-migration model)
BDSKY		1.4.7			birth death skyline - handles serially sampled tips, piecewise constant rate changes through time and sampled ancestors
BEAST_CLASSIC		1.5.0	BEASTLabs		BEAST classes ported from BEAST 1 in wrappers
BEASTLabs	1.9.7	1.9.7			BEAST utilities, such as Script, multi monophyletic constraints
BEASTvntr		0.1.3			Variable Number of Tandem Repeat data, such as microsattelites
Beasy		0.0.2	BEASTLabs		Makes it easier to construct models: Automatic methods text generator, Beasy XML generator, and more
besp		0.2.0			The Bayesian Epoch Sampling Skyline Plot
BICEPS		1.0.0	BEASTLabs		Bayesian Integrated Coalescent Epoch PlotS + Yule Skyline
bModelTest		1.2.1	BEASTLabs		Bayesian model test for nucleotide subst models, gamma rate heterogeneity and invariant sites
BREAK_AWAY		1.0.1	BEASTLabs, GEO_SPHERE		break-away model of phylogegraphy
CA		2.0.0			Bayesian estimation of clade ages based on probabilities of fossil sampling
CoalRe		0.0.7	feast		Infer viral reassortment networks
CodonSubstModels		1.1.3			Codon substitution models
CoupledMCMC		1.0.2	BEASTLabs		Adaptive coupled MCMC (adaptive parallel tempering or MC3)
DENIM		1.0.1			Divergence Estimation Notwithstanding ILS and Migration
EpiInf		7.5.2	SA		BD/SIR/SIS epidemic trajectory inference.
FastRelaxedClockLogNormal		1.1.1	BEASTLabs		Relaxed clock that works well with large data
feast		7.11.0			Expands the flexibility of BEAST 2 XML.
FLC		1.1.0			Flexible local clock model
GEO_SPHERE		1.3.1	BEASTLabs		Whole world phylogeography
Mascot		2.1.2			Marginal approximation of the structured coalescent
MASTER		6.1.2			Stochastic population dynamics simulation
MGSM		0.3.0			Multi-gamma and relaxed gamma site models
MM		1.1.1			Enables models of morphological chatacter evolution
MODEL_SELECTION	1.5.3	1.5.3	BEASTLabs		Select models through path sampling/stepping stone analysis
MSBD		1.2.0			Multi-state birth-death prior with state-specific birth and death rates
MultiTypeTree		7.0.2			Structured coalescent inference
NS	1.1.0	1.1.0	BEASTLabs, MODEL_SELECTION		Nested sampling for model selection and posterior inference
OBAMA		0.2.0	BEASTLabs, bModelTest		OBAMA for Bayesian Amino-acid Model Averaging
ORC		1.0.3	BEASTLabs, FastRelaxedClockLogNormal		Optimised Relaxed Clock model
PhyDyn		1.3.8			PhyDyn: Epidemiological modelling with BEAST
phylodynamics		1.3.0	BDSKY		BDSIR and Stochastic Coalescent
PIQMEE		1.0.2	BDSKY		Birth-death skyline-based method efficiently dealing with duplicate sequences
PoMo		1.0.1			PoMo, a substitution model that separates mutation and drift processes
Recombination		0.0.2			Inference of Recombination networks
SA		2.0.2	BEASTLabs		Sampled ancestor trees
SCOTTI		2.0.1			Structured COalescent Transmission Tree Inference
SNAPP		1.5.2			SNP and AFLP Phylogenies
snapper		1.0.2	SNAPP		Diffusion based SNP and AFLP Phylogenies
SpeciesNetwork		0.13.0			Multispecies network coalescent (MSNC) inference of introgression and hybridization
SSM		1.1.0			Standard Nucleotide Substitution Models
STACEY		1.2.5			Species delimitation and species tree estimation
StarBEAST2		0.15.13	MM, SA		Multispecies coalescent inference using multi-locus and fossil data
starbeast3		1.0.4	ORC, SA, BEASTLabs		StarBeast3 multispecies coalescent using advanced MCMC operators
substBMA		1.2.3			Substitution Bayesian Model Averaging
TMA		1.0.0	TreeStat2, BEASTLabs, MASTER, phylodynamics, BDSKY		Tree model adequacy: test whether the tree prior used is adequate for your data

Dramatic increase in the use of genome sequencing



Molecular dating of large datasets



Duchene *et al. BMC Evolutionary Biology* (2018) 18:95
https://doi.org/10.1186/s12862-018-1210-5

BMC Evolutionary Biology

METHODOLOGY ARTICLE Open Access

Inferring demographic parameters in bacterial genomic data using Bayesian and hybrid phylogenetic methods

Sebastian Duchene[1*], David A. Duchene[2], Jemma L. Geoghegan[3], Zoe A. Dyson[1], Jane Hawkey[1] and Kathryn E. Holt[1]

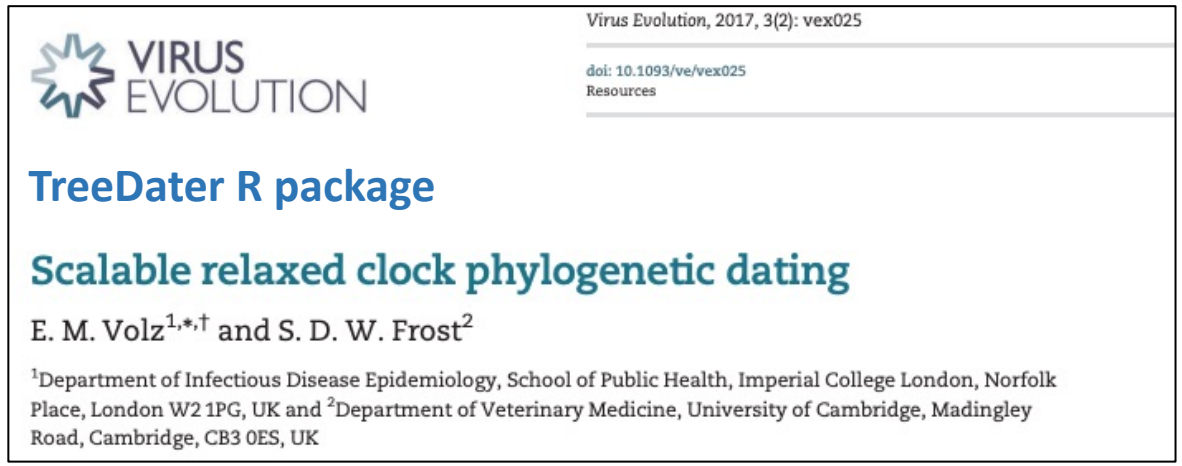
Published online 3 September 2018 *Nucleic Acids Research, 2018, Vol. 46, No. 22 e134*
doi: 10.1093/nar/gky783

BactDating R package

Bayesian inference of ancestral dates on bacterial phylogenetic trees

Xavier Didelot[1,*], Nicholas J. Croucher[1], Stephen D. Bentley[2], Simon R. Harris[2] and Daniel J. Wilson[3]

[1]Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, London, UK, [2]The Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK and [3]Big Data Institute, Nuffield Department of Population Health, University of Oxford, Oxford, UK



VIRUS EVOLUTION

Virus Evolution, 2017, 3(2): vex025

doi: 10.1093/ve/vex025
Resources

TreeDater R package

Scalable relaxed clock phylogenetic dating

E. M. Volz[1,*,†] and S. D. W. Frost[2]

[1]Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, Norfolk Place, London W2 1PG, UK and [2]Department of Veterinary Medicine, University of Cambridge, Madingley Road, Cambridge, CB3 0ES, UK

- BEAST/BEAST2 fail to converge with more than ~300 samples
- Tree inference is the computationally intensive component
- **Hybrid methods proposed** – ML tree (fast) + BEAST dating
- Recent methods developed in R can **infer a dated phylogeny from a user supplied ML tree and sampling dates**, allowing rapid analysis of thousands of genomes
- R packages allow for easy installation
- However, these ML tree-based methods often allow **fewer priors**

Molecular dating for large datasets

VIRUS EVOLUTION

Virus Evolution, 2017, 3(2): vex025

doi: 10.1093/ve/vex025
Resources

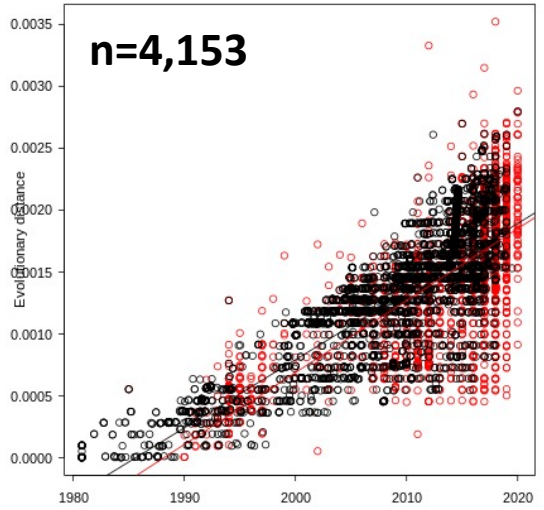
TreeDater R package

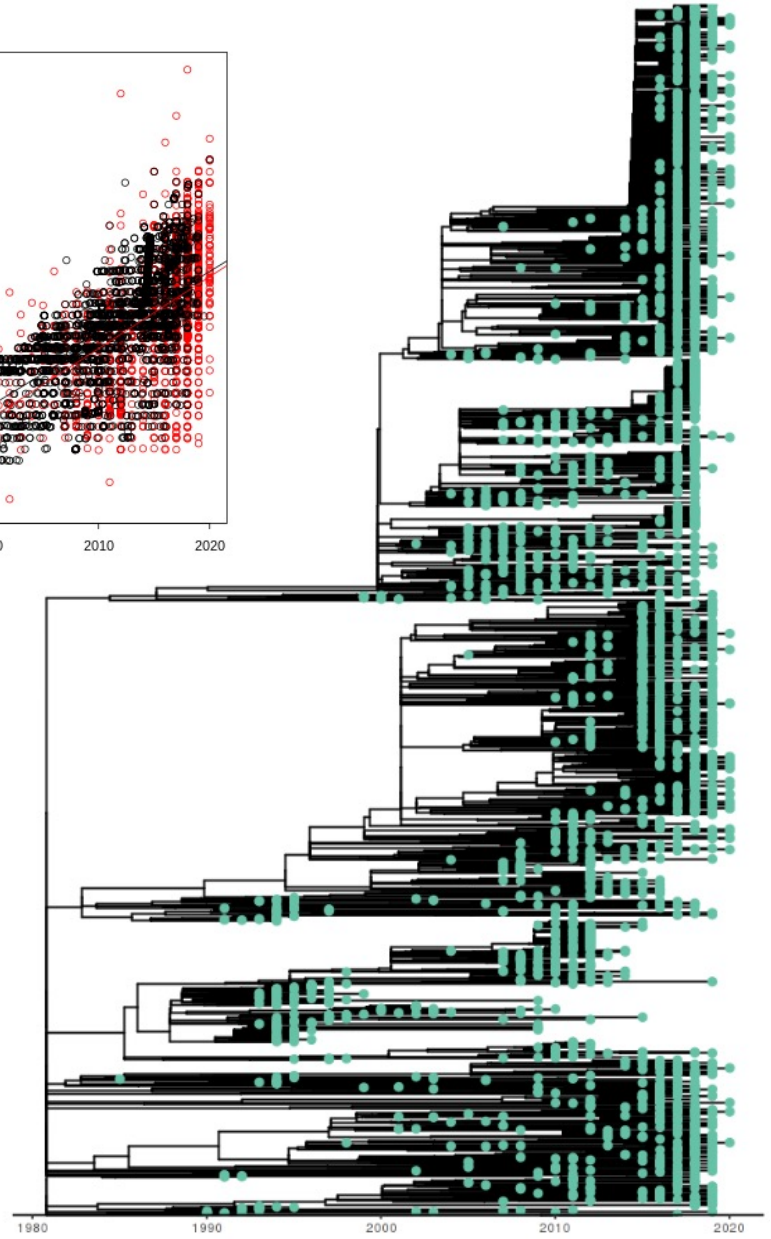
Scalable relaxed clock phylogenetic dating

E. M. Volz[1,*,†] and S. D. W. Frost[2]

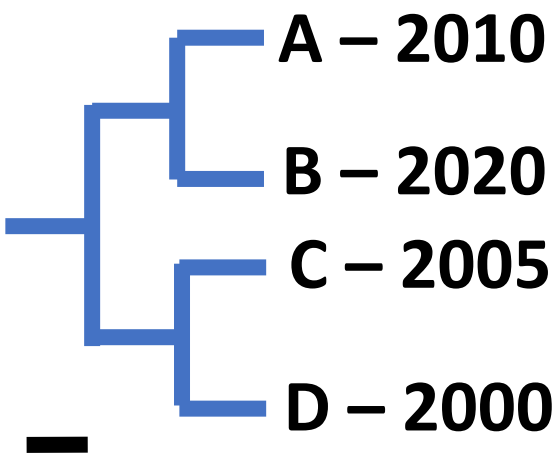
[1]Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, Norfolk Place, London W2 1PG, UK and [2]Department of Veterinary Medicine, University of Cambridge, Madingley Road, Cambridge, CB3 0ES, UK

*Corresponding author: E-mail: e.volz@imperial.ac.uk
†http://orcid.org/0000-0001-6268-8937

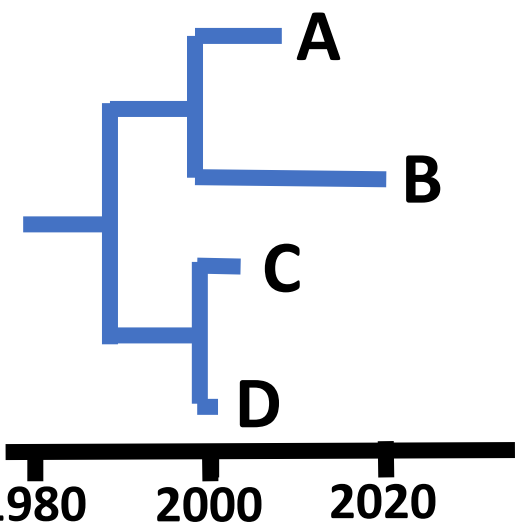




ML tree + dates



Dated phylogeny



Dyson *et al.* 2023, unpublished

Extended functionality via R packages…

BactDating R package

Syst. Biol. 67(4):719–728, 2018

DOI:10.1093/sysbio/syy007
Advance Access publication February 7, 2018

Modeling the Growth and Decline of Pathogen Effective Population Size Provides Insight into Epidemic Dynamics and Drivers of Antimicrobial Resistance

ERIK M. VOLZ* AND XAVIER DIDELOT

Department of Infectious Disease Epidemiology, Imperial College London, Norfolk Place, W
**Correspondence to be sent to: Department of Infectious Disease Epidemiology, Imperial College London, N*
Email: e.volz@imperial.ac.uk.

Received 19 October 2017; reviews returned 1 February 2018; accepted 4 February 2
Associate Editor: Jeffrey Townsend

Transphylo R package

Genomic Infectious Disease Epidemiology in Partially Sampled and Ongoing Outbreaks

Xavier Didelot,*[,1] Christophe Fraser,[1,2] Jennifer Gardy,[3,4] and Caroline Colijn[5]

[1]Department of Infectious Disease Epidemiology, Imperial College London, Norfolk Place, London, United Kingdom

[2]Oxford Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom

[3]Communicable Disease Prevention and Control Services, British Columbia Centre for Disease Control, Vancouver, British Columbia,

nd Public Health, University of British Columbia, Vancouver, British Columbia, Canada

natics, Imperial College, London, United Kingdom

Outbreaker2 R package

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Bayesian inference of transmission chains using timing of symptoms, pathogen genomes and contact data

Finlay Campbell[1]*, Anne Cori[1], Neil Ferguson[1], Thibaut Jombart[1,2,3]*

1 MRC Centre for Global Infectious Disease Analysis, Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, United Kingdom, **2** Department of Infectious Disease Epidemiology, London School of Hygiene & Tropical Medicine, London, United Kingdom, **3** UK Public Health

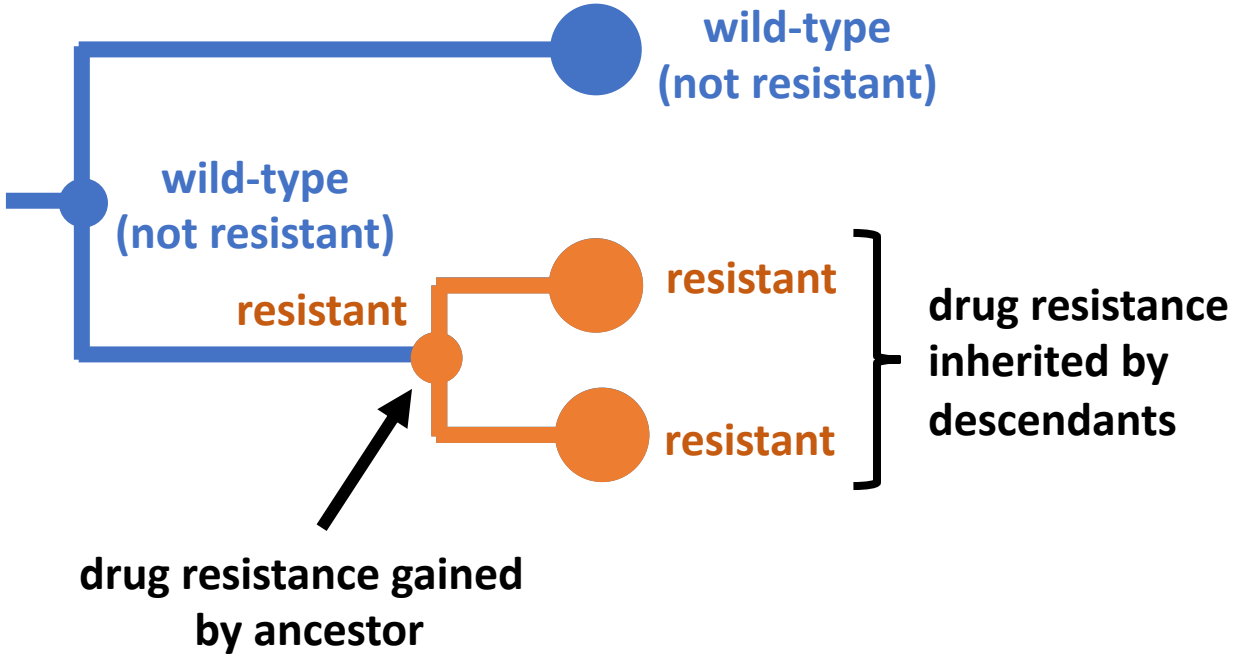
Ancestral state reconstruction

- **Ancestral state reconstruction** combines information about evolutionary relationships from a phylogenetic tree with the observed state of a trait for each tip

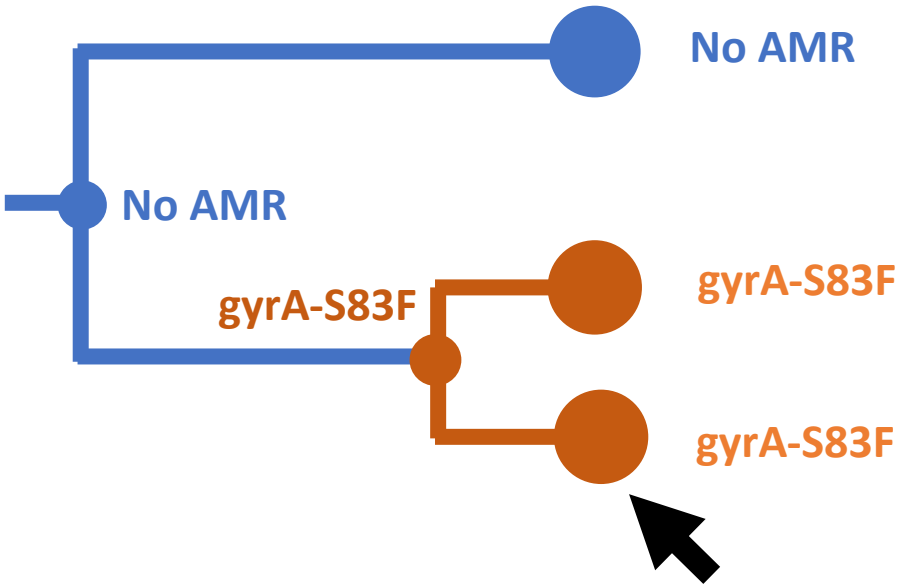
- Given a tree, and a set of traits, these algorithms will determine the **most likely trait state for ancestral (internal) nodes** allowing us to identify the point in the evolution of a pathogen population **when a trait was acquired, or lost** (e.g. drug resistance)

- We can extend this to other traits such as geography (referred to as phylogeography) to **understand pathogen spread**

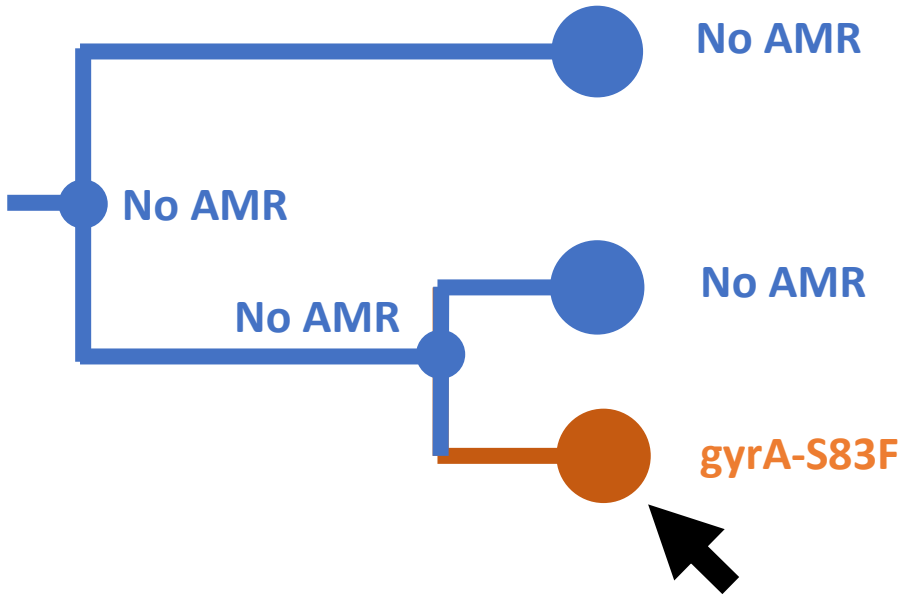
- Traits may be **discrete** (e.g. country of origin), or **continuous** (e.g. GPS coordinates)



Ancestral state reconstruction

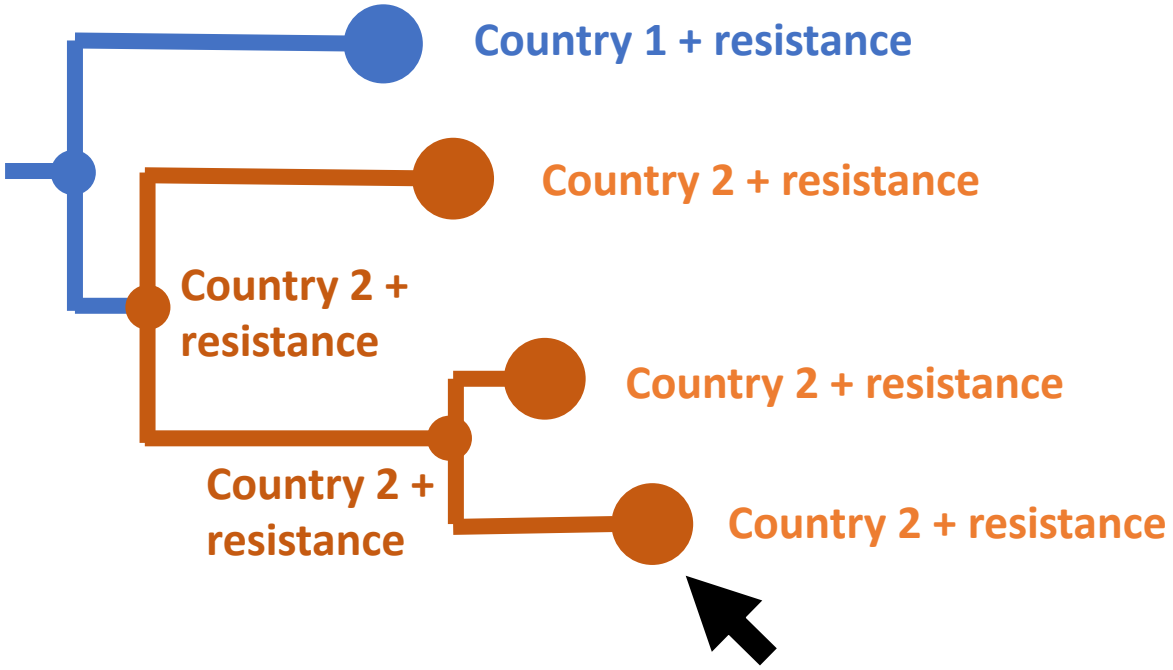


Inherited/transmitted

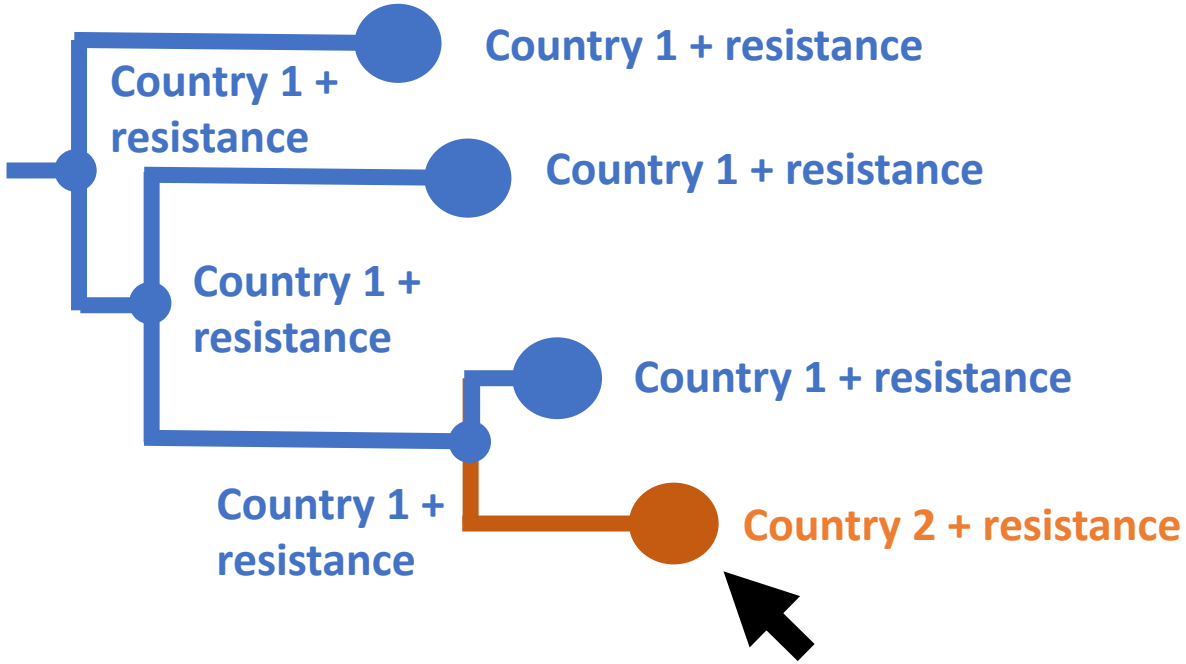


de novo acquisition (gain)

Ancestral state reconstruction (resistance + country)

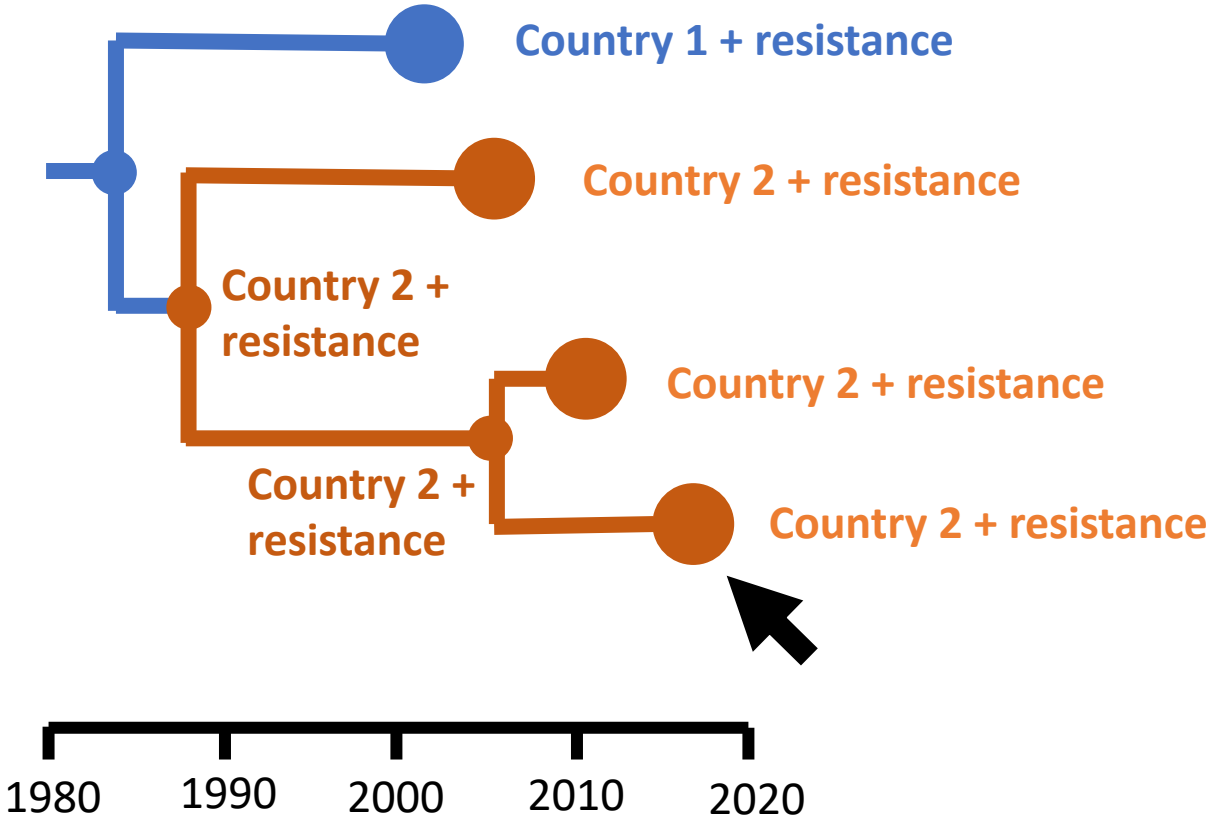


Transmission (local)

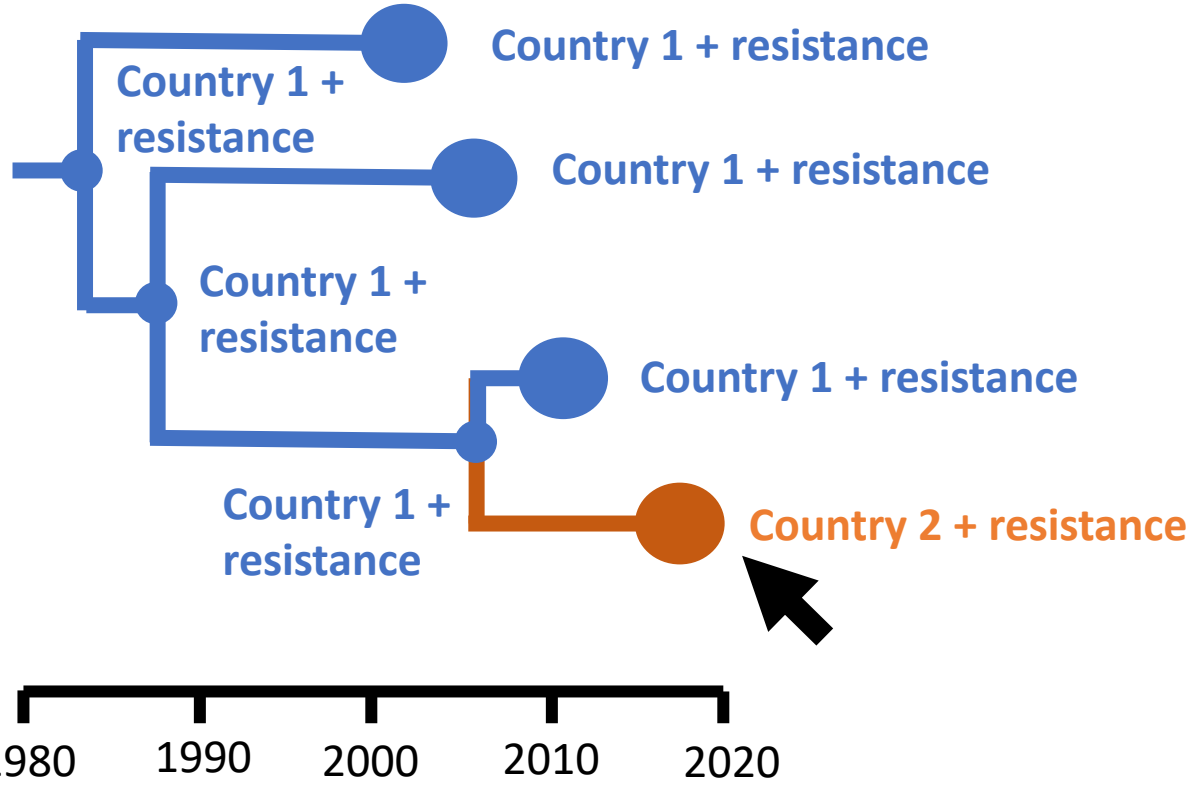


Transmission (international/external)

Ancestral state reconstruction (resistance + country + time)



Transmission (local)



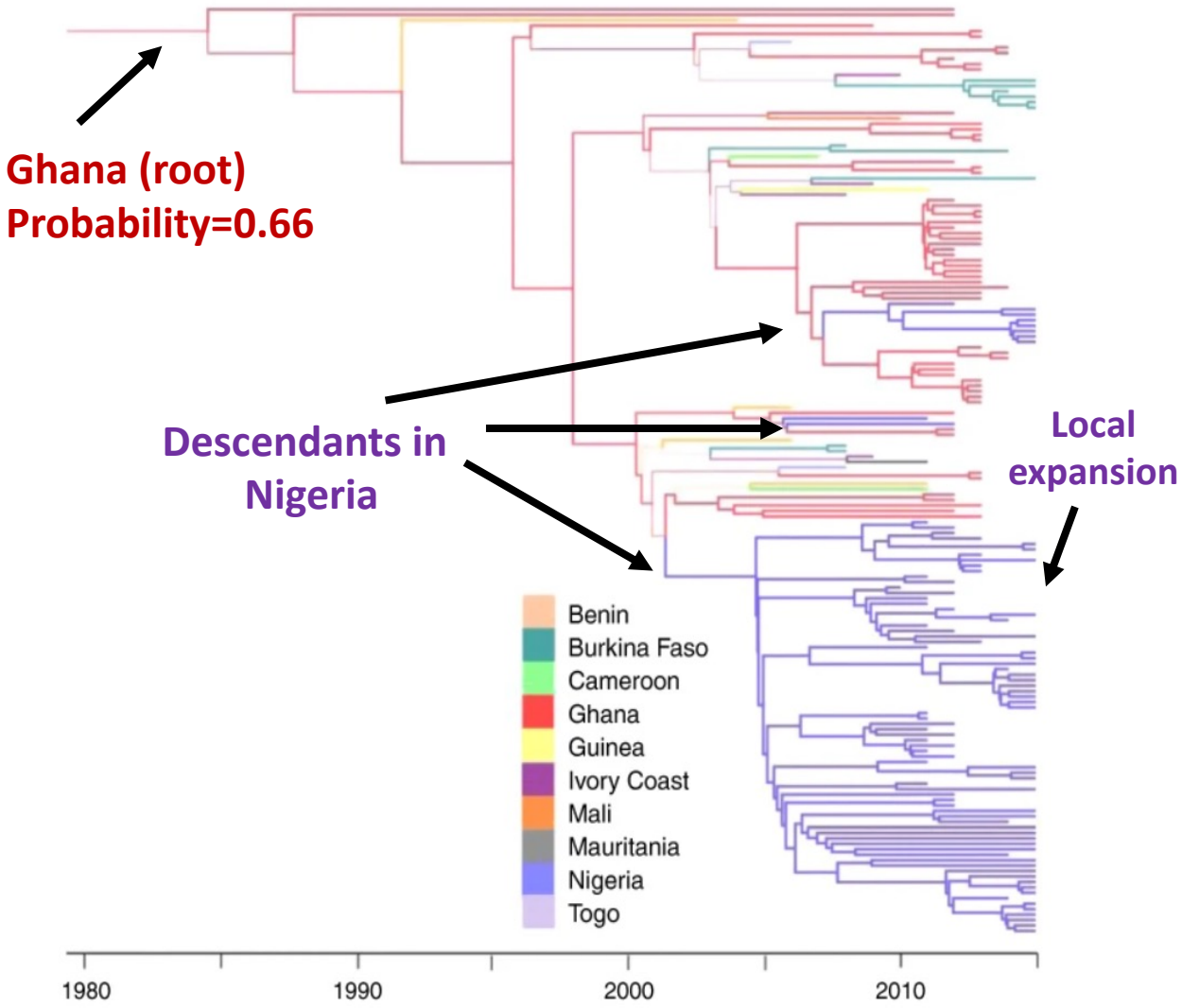
Transmission (international/external)

Extending phylodynamics: phylogeography

- **Phylogeography** is the study of the historical processes that have contributed to the present geographic distribution of lineages

- **BEAST2** can be used to model geography either:
  - **discretely** e.g. by country
  - **continuously** e.g. using GPS coordinates

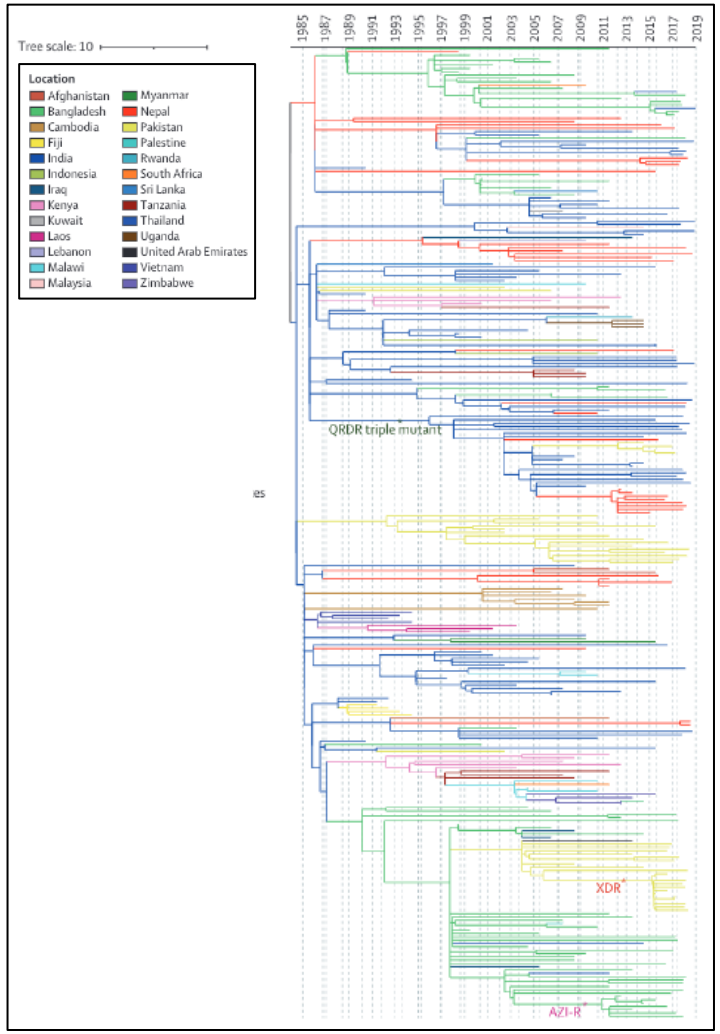
- When **discrete trait mapping** is carried out, locations are mapped to the tips of the tree and the location for ancestors inferred at internal nodes. This allows for the identification of the point in time when a pathogen has spread from one location to another.

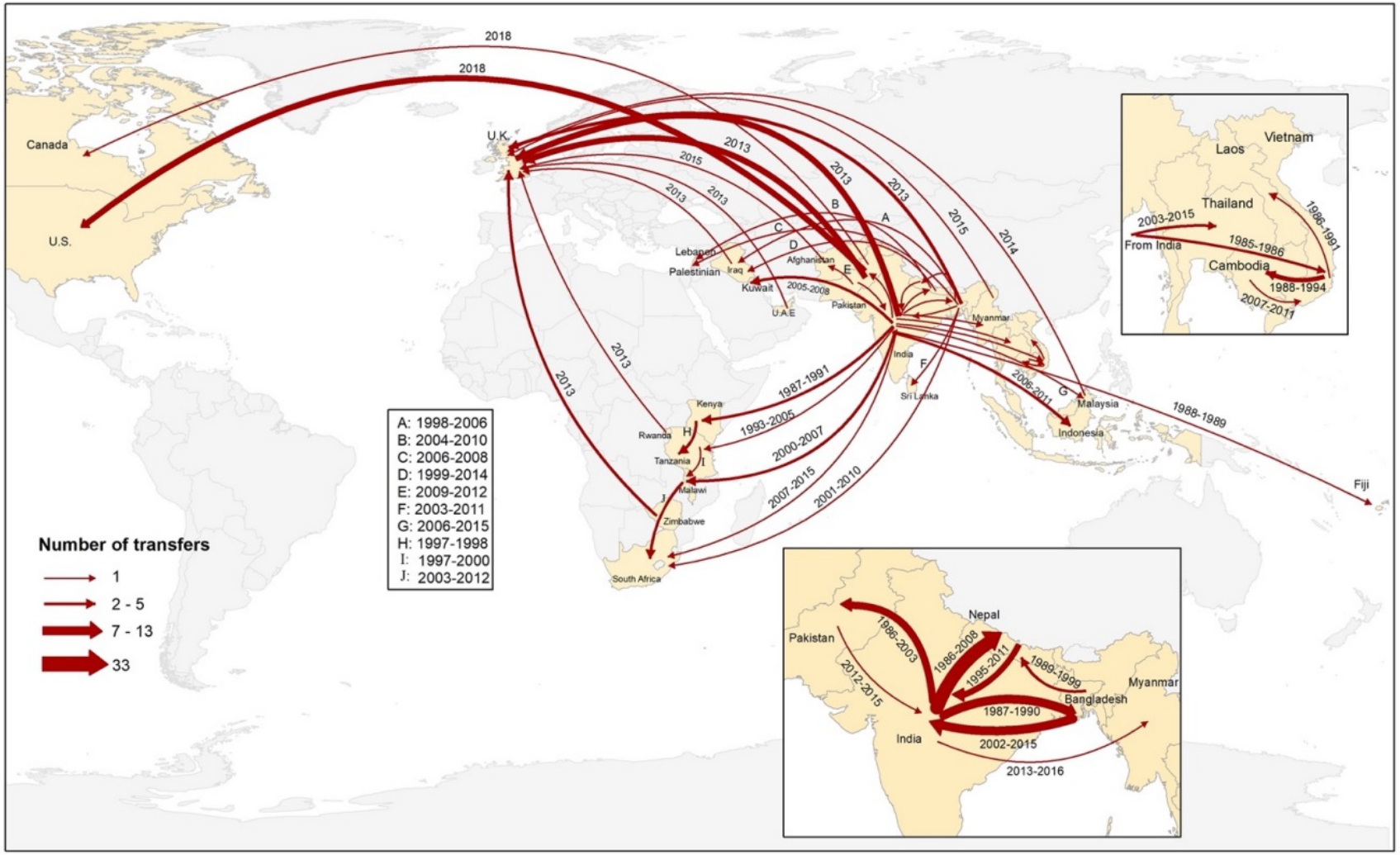


Spread of *Salmonella* Typhi from Ghana to Nigeria

Park *et al.* 2018, Nat Commun

Examples of extending phylodynamics: phylogeography

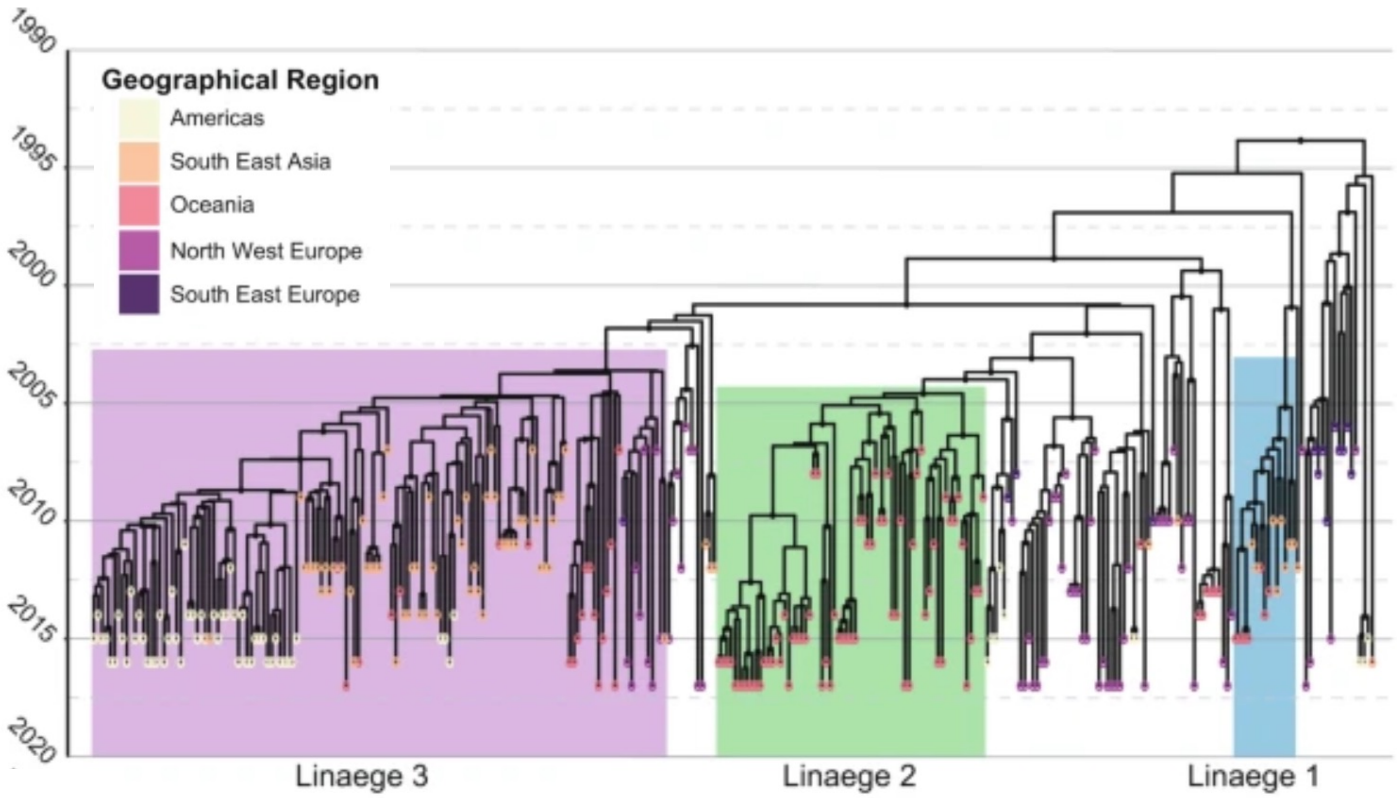


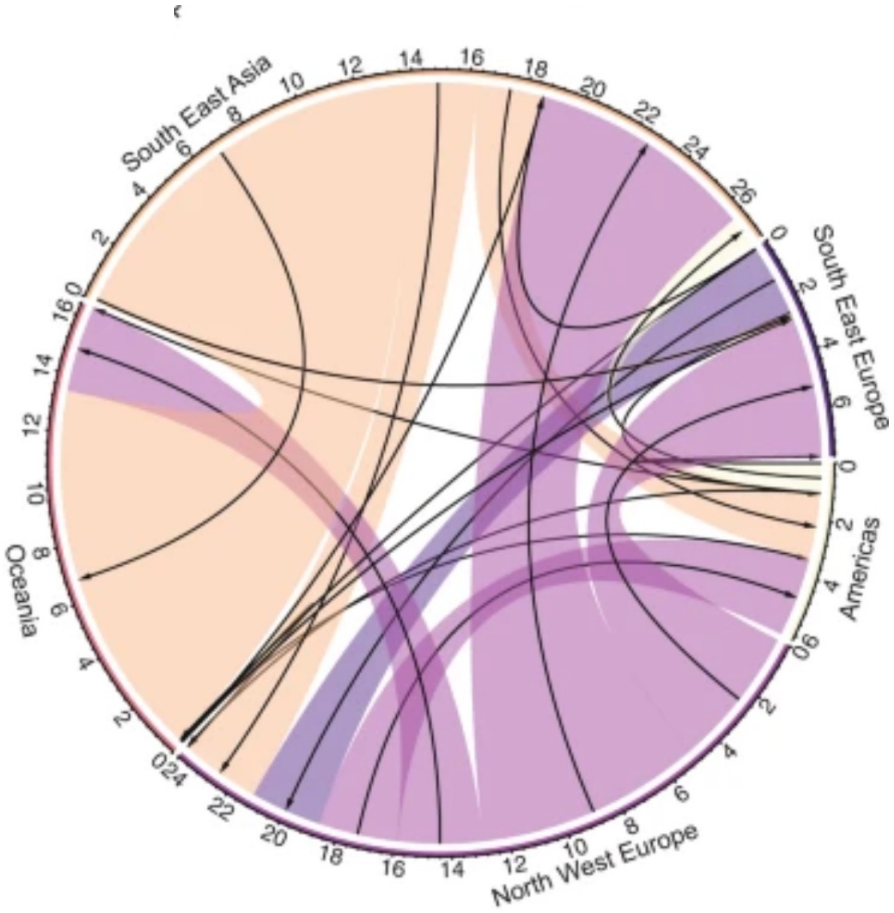


Global dissemination of *Salmonella* Typhi genotype 4.3.1
n=4761, BactDating R package

da Silva *et al.* 2022, *Lancet Microbe*

Continuous Time Markov Chain: regional transmission





Markov jumps *are transitions between different states (regions) along phylogenetic branches*

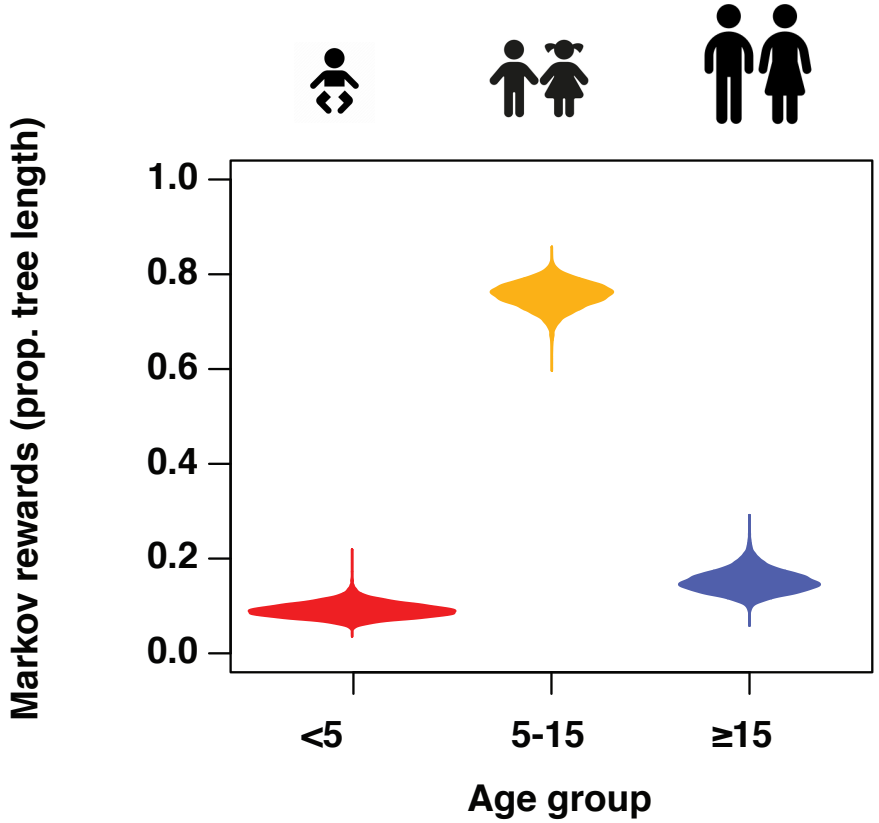
Ingle *et al.* 2019, Nat Commun

Continuous Time Markov Chain: age group transmission

Markov jumps

From \ To	Age	<5	5-15	≥15	TOTAL
<5	<5	-	0 (0-0.04)	0 .02 (0-0.09)	**0.02 (0-0.09)**
5-15	5-15	0.3 (0.3-0.4)	-	0.6 (0.5-0.7)	**0.9 (0.3-0.7)**
≥15	≥15	0 (0-0.06)	0 (0-0.02)	-	**0 (0-0.06)**

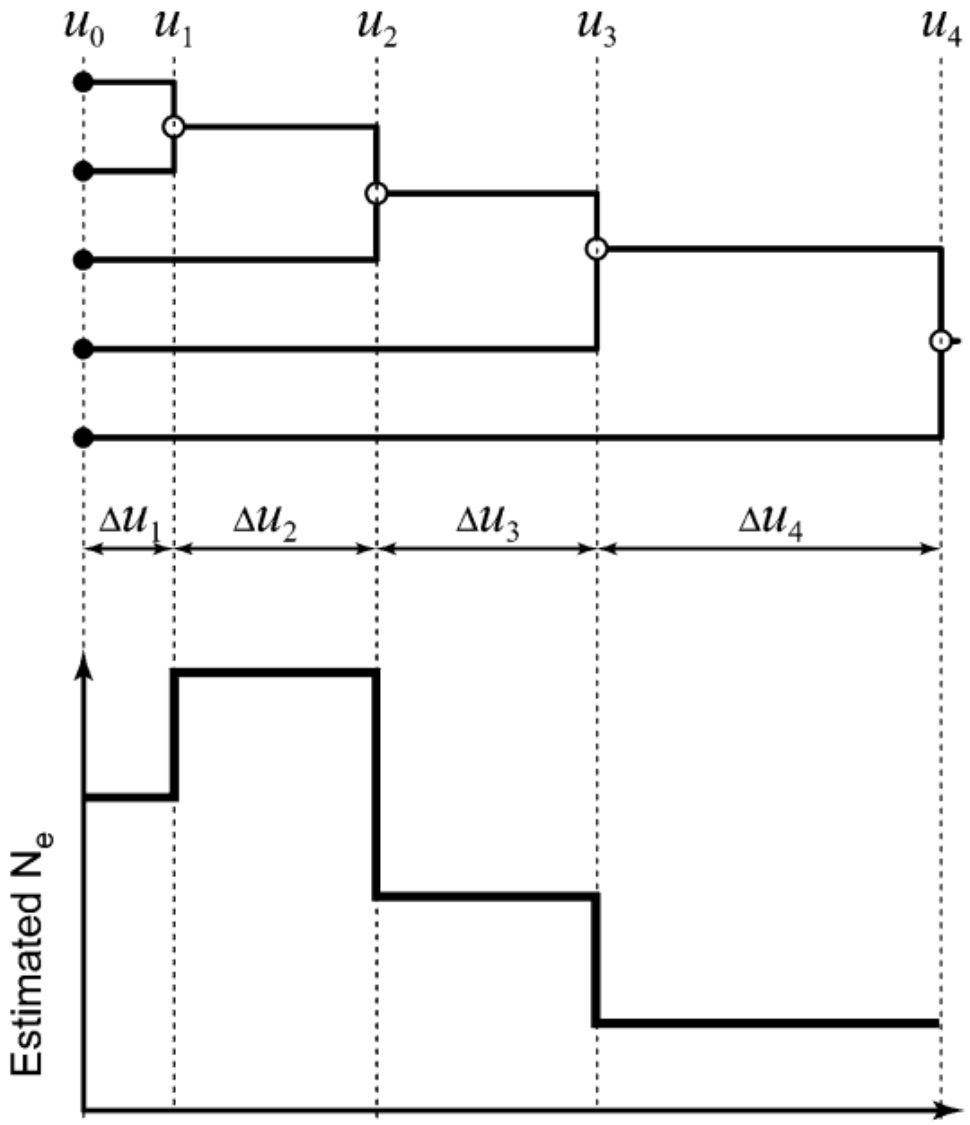
Markov rewards



Markov jumps *are transitions between different states (age groups) along phylogenetic branches*
Markov rewards *can be considered as the time spent in the states between two transitions*

Examples of extending phylodynamics: N_e

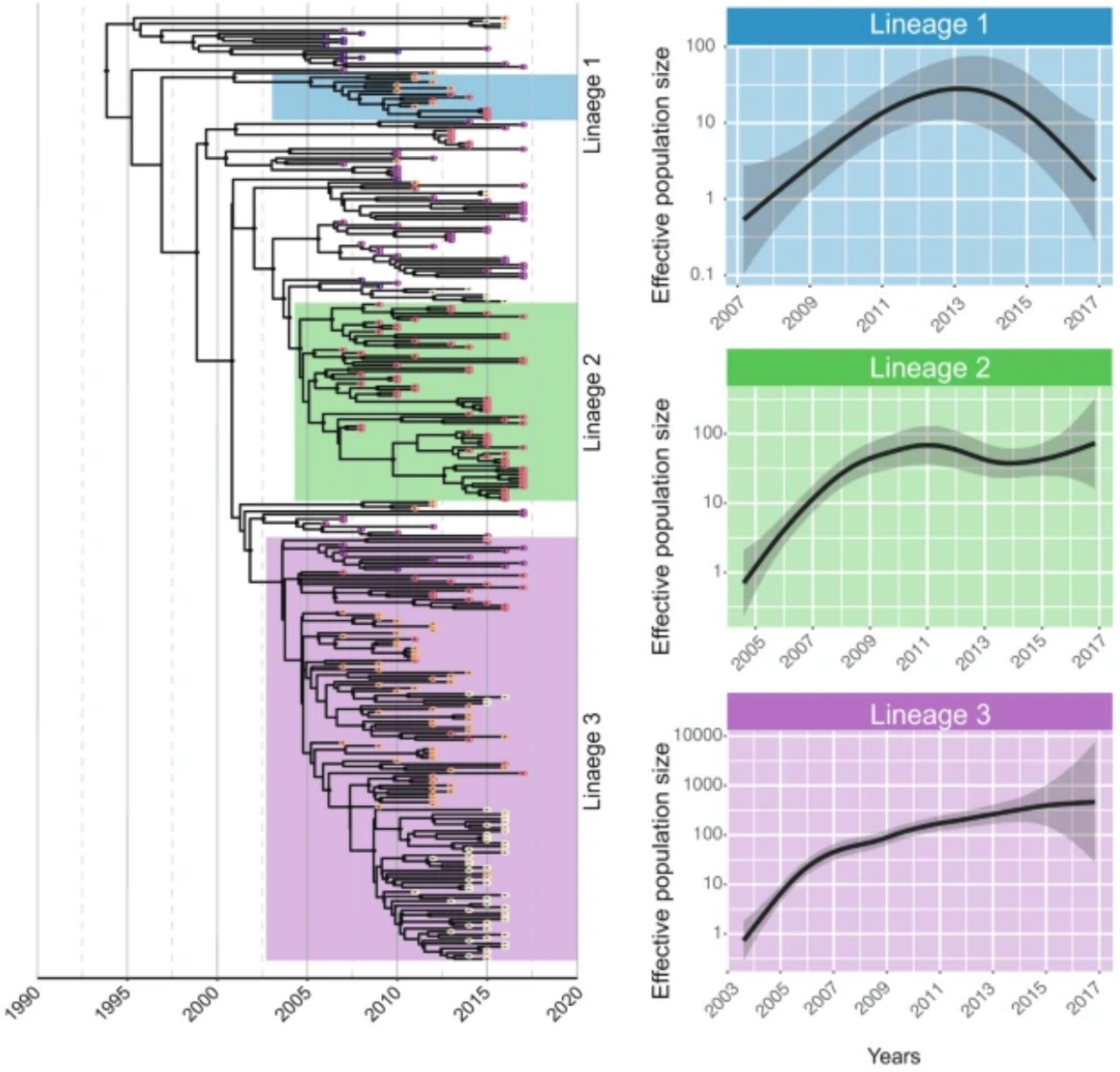
- Phylodynamic techniques can also be used to model changes in the **effective population size (N_e)** over time
- **N_e** is the size of an idealized population showing the same rate of change in genetic diversity as the real population under study
- N_e estimates are usually plotted in a skyline or skygrowth plot & are **sensitive to sampling**!



Drummond *et al.* 2005, *Mol Biol Evol*

Examples of extending phylodynamics: N_e

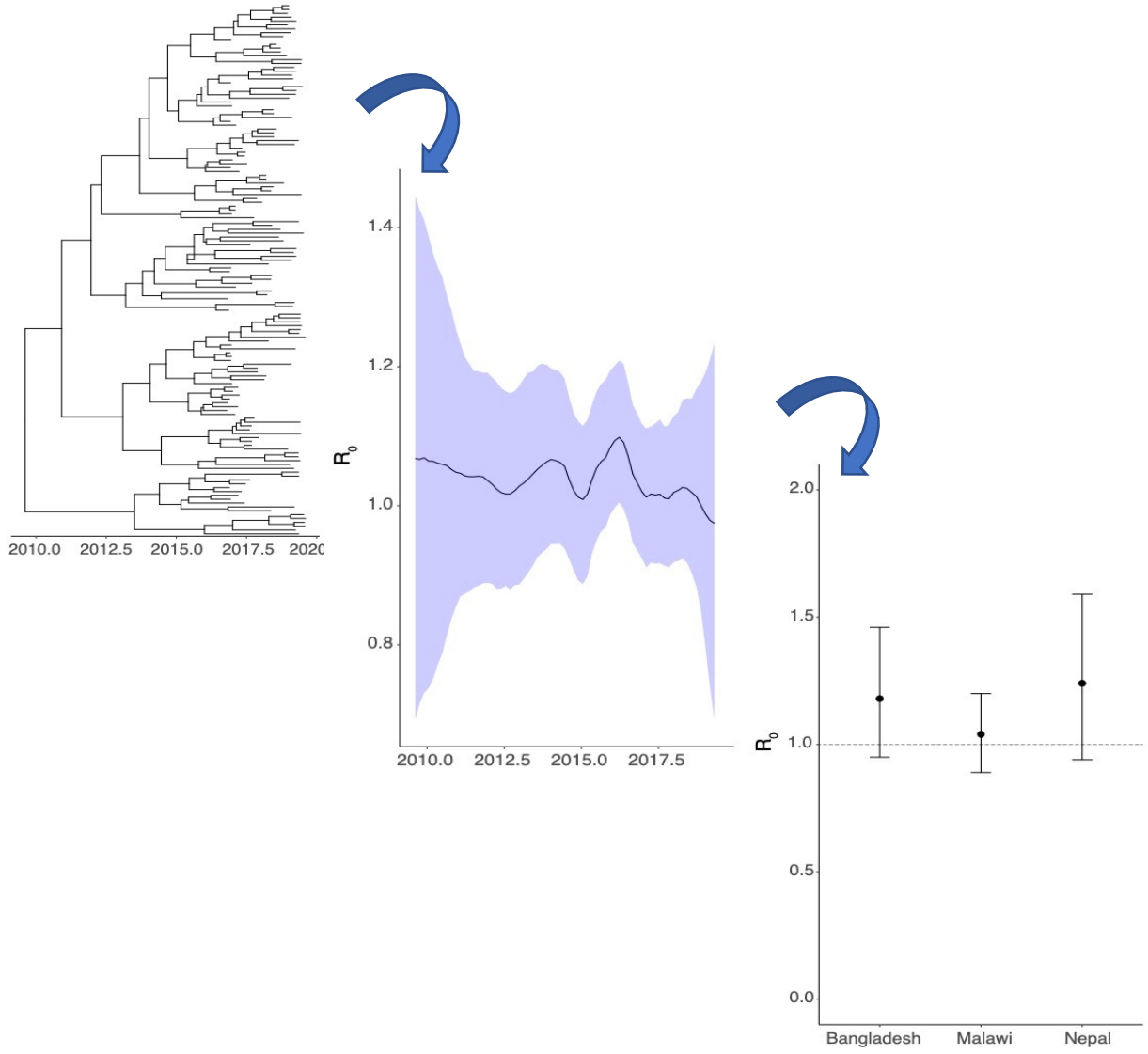
- Phylodynamic techniques can also be used to model changes in the **effective population size (N_e)** over time
- **N_e** is the size of an idealized population showing the same rate of change in genetic diversity as the real population under study
- N_e estimates are usually plotted in a skyline or skygrowth plot & are **sensitive to sampling**!



Ingle *et al.* 2019, Nat Commun

Examples of extending phylodynamics: N_e & R_0

- Phylodynamic techniques can also be used to model changes in the **effective population size (N_e)** over time
- **N_e** is the size of an idealized population showing the same rate of change in genetic diversity as the real population under study
- N_e estimates are usually plotted in a skyline or skygrowth plot & are **sensitive to sampling**!
- It is also possible to model the **basic reproduction number (R_0)**, a measure of how contagious a pathogen is i.e. how many cases are derived from one case in a sensitive population. This can by modelled using the phylogeny & additional information e.g. the infectious period



Dyson et al. 2023, medrxiv, generated with skygrowth

Join at menti.com | use code 8692 7095

Mentimeter

What data could you use in a phylodynamic analysis?

Intended learning outcomes

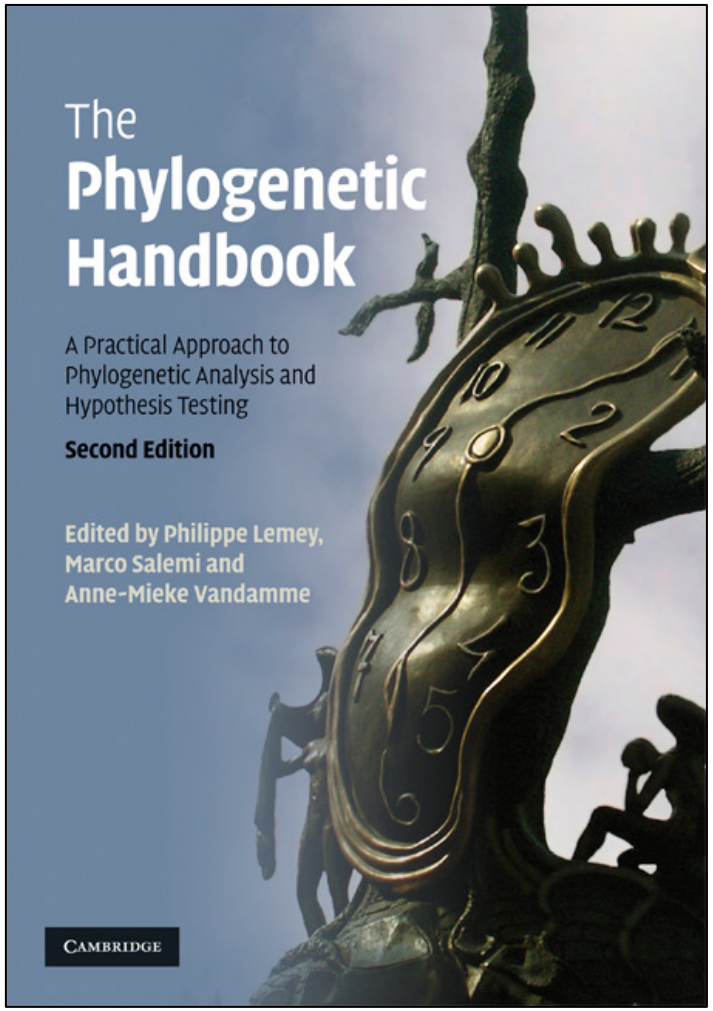
1. Recognise the basic principles of phylogenetics
2. Interpret data on a phylogenetic tree
3. Explain the methods used to infer a phylogenetic tree from bacterial pathogen whole genome sequencing data
4. Explain core concepts related to phylodynamics and how these can provide insights into pathogen evolution and epidemiology

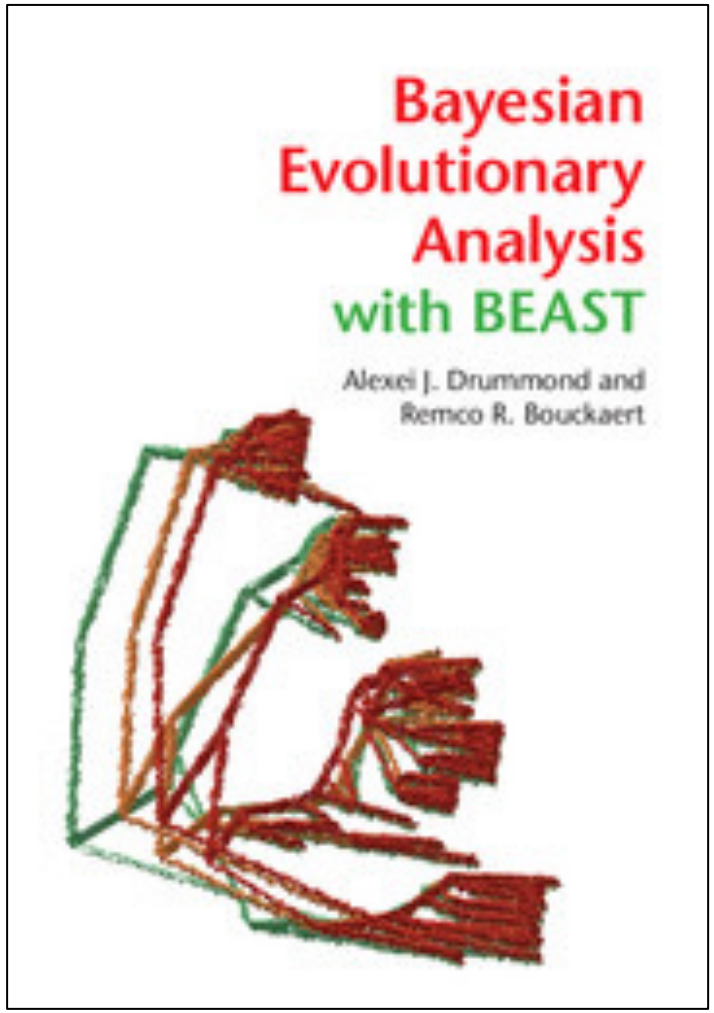
If you would like to learn more about phylogenomics

Bayesian dated trees:
Taming the BEAST: https://taming-the-beast.org/
TempEst: http://tree.bio.ed.ac.uk/software/tempest/
DensiTree:
https://www.cs.auckland.ac.nz/~remco/DensiTree/

ML Trees:
RAxML online: http://www.phylo.org/index.php/

Other tools:
FigTree: http://tree.bio.ed.ac.uk/software/figtree/
Phandango:
https://jameshadfield.github.io/phandango/#/





Dr Zoe Anne Dyson
Assistant Professor
Department of Infection Biology
London School of Hygiene & Tropical Medicine
zoe.dyson@lshtm.ac.uk