Computational Practical 8

What are phylogenetic trees?

A phylogeny, also known as phylogenetic tree, depicts estimated evolutionary relationships between taxa - these can be species, strains or even genes.

In the context of infectious diseases epidemiology, phylogenetic trees are commonly used to define evolutionary relationships between strains of the same bacterial species. This is possible because bacteria reproduce clonally.

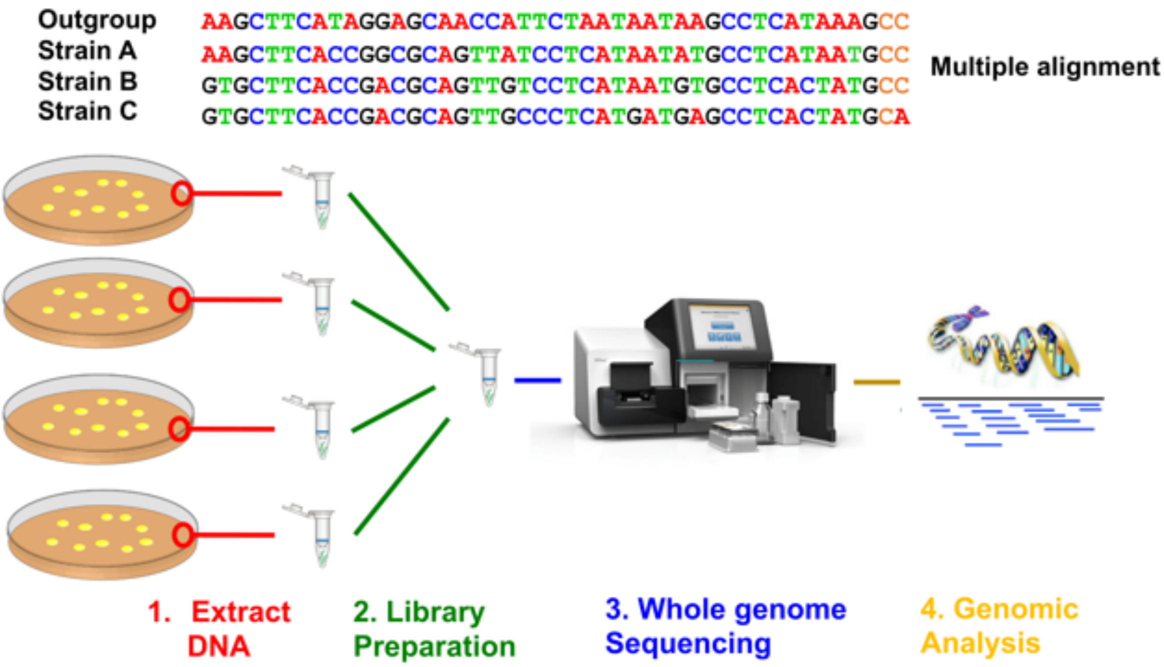
During clonal reproduction, bacterial progenitor cells replicate their DNA at high fidelity. Despite this, random errors in DNA replication may still occur, resulting in a clonal progeny that will inherit these genetic replication 'errors' (i.e. mutations) in their DNA and may not be strictly identical to their progenitor cells. Bacterial strains that have recently originated from the same progenitor cell are thus expected to share identical genomes or have diverged at most by only a few genetic differences (mutations). The number and pattern of shared mutations between bacterial strains can be used to reconstruct their genealogical and evolutionary relationships.

On a phylogenetic tree, isolated bacterial strains are depicted on the tips (or leaves) of the tree (i.e. taxa), whereas the internal nodes of the tree denote their hypothetical ancestors. Nodes and taxa are connected by branches, the length of which represent genetic distances between connected groups. Groups of bacterial strains (taxa) that share the same common ancestor form a monophyletic group (also known as clade). A group of strains that descends from a common ancestor, but does not include all descendants, is called paraphyletic.

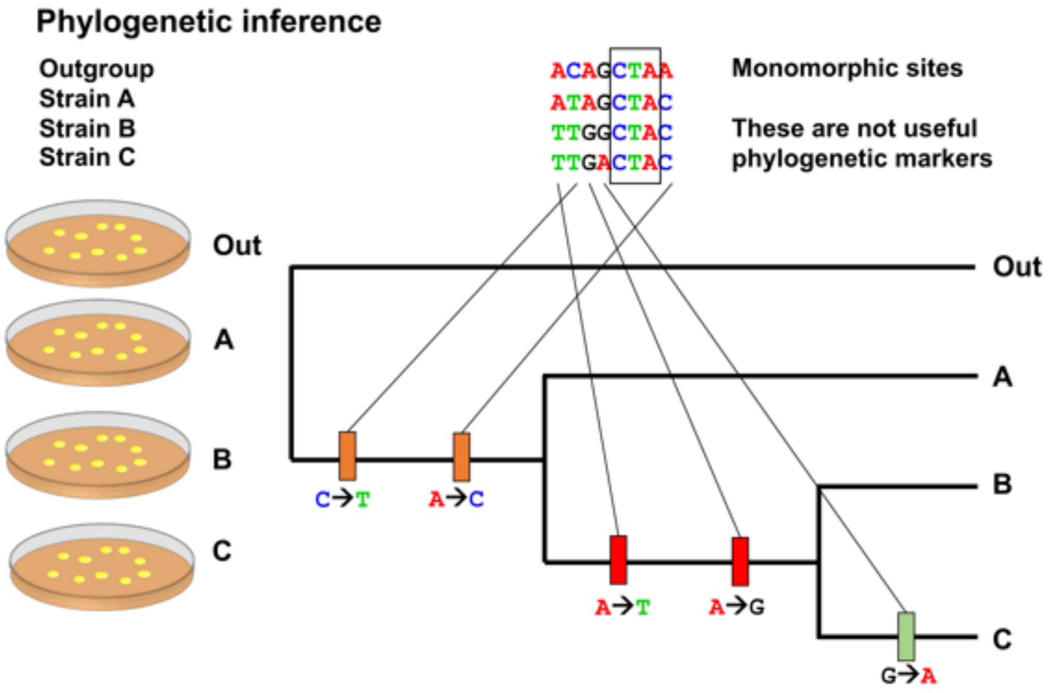
How are phylogenetic trees reconstructed?

Today almost all phylogenetic trees are inferred from molecular sequence data, most often from DNA sequences. This is because DNA is an inherited material; it can easily, reliably and inexpensively be extracted and sequenced; and DNA sequences are highly specific to bacterial species and strains.

The application of whole-genome sequencing now makes it possible to 'read' the DNA sequence of the whole bacterial chromosome, which provides the ultimate level of resolution possible to discriminate between closely related strains. The figure below illustrates the common workflow to generate multiple DNA sequence alignments from a collection of bacterial strains. Generally, bacterial DNA is extracted from a single colony picked from culture plates (referred as to 'isolate'), followed by library preparation and whole-genome sequencing using rapid benchtop sequencers. Raw sequence data generated by sequencers are processed using bioinformatic and genomic pipelines, which generally involve mapping (aligning) the generated short reads to a reference genome to reconstruct the isolate's DNA sequence along the whole bacterial chromosome. Mapping the short reads of multiple sequenced isolates to the same reference genome allows the creation of multiple sequence alignments. Multiple sequence alignments are the first and critical point from which phylogenetic trees can be re-constructed.


Outgroup
Strain A
Strain B
Strain C
AAGCTTCATAGGAGCAACCATTCTAATAATAAGCCTCATAAAGCC
AAGCTTCACCGGCGCAGTTATCCTCATAATATGCCTCATAATGCC
GTGCTTCACCGACGCAGTTGTCCTCATAATGTGCCTCACTATGCC
GTGCTTCACCGACGCAGTTGCCCTCATGATGAGCCTCACTATGCA
Multiple alignment
1. Extract DNA
2. Library Preparation
3. Whole genome Sequencing
4. Genomic Analysis

Polymorphic sites (that is, nucleotide positions that are variable across multiple strains) in multiple alignments are used to infer evolutionary relationships, whereas monomorphic sites (nucleotide positions with the same DNA base) are generally ignored. The figure below shows an example of a simple multiple alignment of eight sites from four strains, which include monomorphic (squared) and polymorphic sites. Genetic changes in the phylogenetic tree are showed as coloured vertical rectangles on the branch where they originated. The identification of genetic changes (alleles) that are unique and common to multiple taxa (strains) are used to group them into monophyletic groups (clades) in a hierarchical manner (see example below) with the goal of constructing the most plausible genealogical relationships between strains and clades.



How are phylogenetic trees interpreted?

The preferred interpretation of a phylogenetic tree is as a depiction of lines of descent. That is, trees communicate the evolutionary relationships among strains and clades. Under this interpretation, internal nodes on a tree are taken to correspond to bacterial strains that existed in the past (ancestral) but could not be sampled.

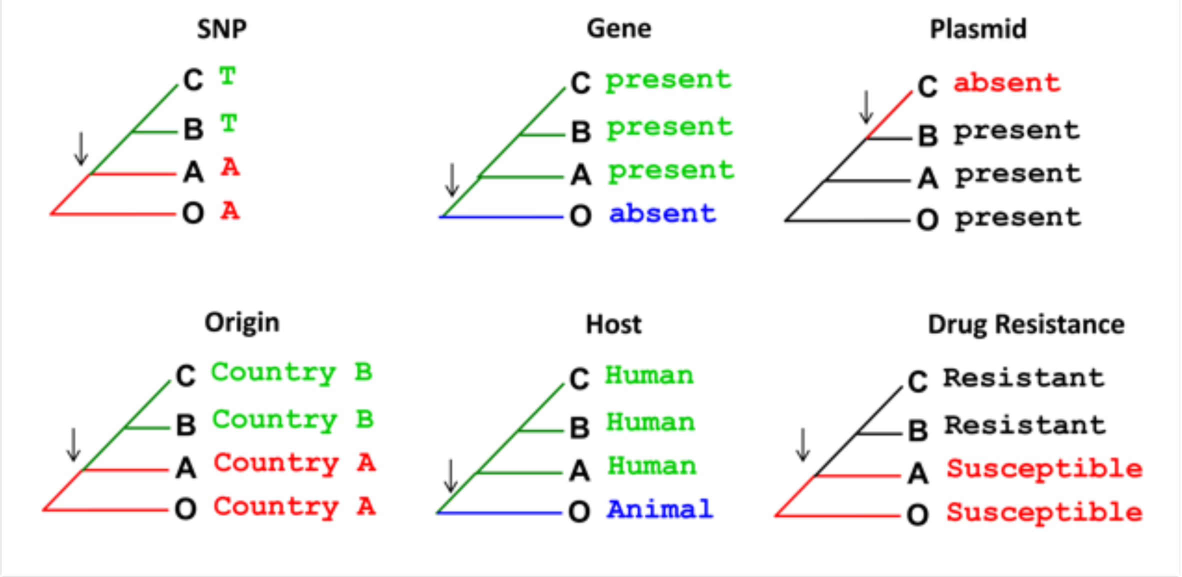
Phylogenies are commonly mis-interpreted when read along the tips. Instead, the correct way to read a tree is as a set of hierarchically nested groups (clades).

In the tree above, strain C is more closely related to strain B than it is to strain A. This is inferred by tracing the ancestor of strains (depicted as internal nodes) using the branch structure (i.e. topology) of the tree. Relatedness should be understood in terms of common ancestry— the more recently strains share a common ancestor, the more closely related they are.

How are phylogenetic trees used in infectious diseases epidemiology?

Phylogenetic trees are commonly used to identify where person-to-person transmission occurs; to identify the sources and study the transmission routes of outbreak and epidemic clones; and to determine whether bacterial clones are restricted to specific hosts and settings or, on the contrary, able to circulate among multiple ones.

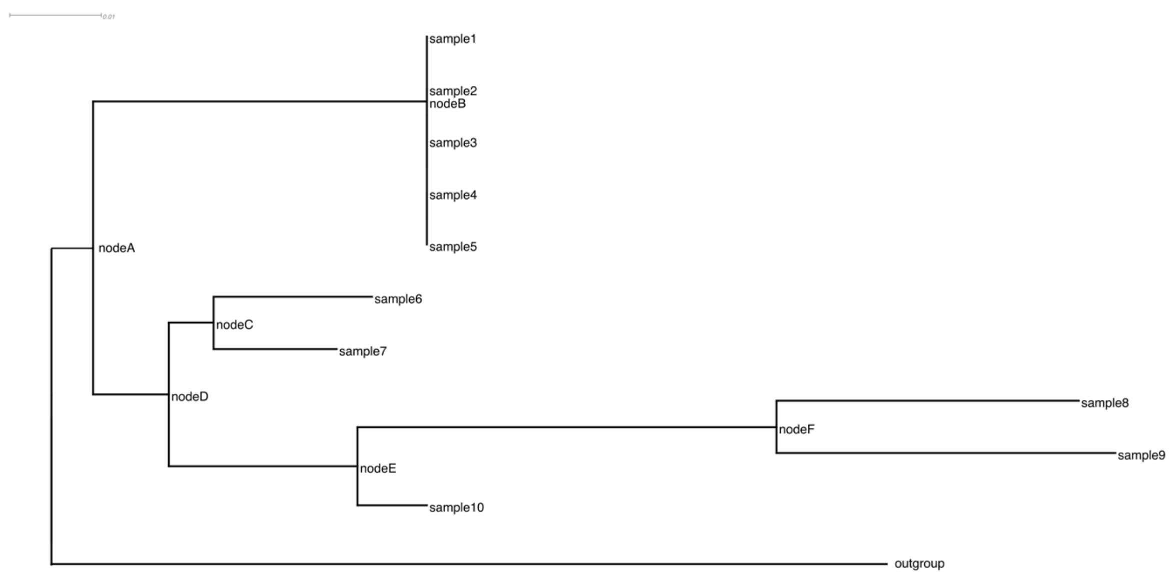
A common phylogenetic method used to study how bacterial characteristics (traits) evolved is ancestral state reconstruction. In the example shown below, strains on the same tree are labelled based on the presence of different traits. Arrows indicate what internal node (ancestor) in the tree most likely changed (lost or gained) such a trait. Bacterial traits we may be interested in reconstructing include: geographical location - to then identify movement between regions (transmission events; colonising or infecting host - to enable us to identify host jumps; and antibiotic susceptibility - to enable us to identify evolution of AMR. The emergence and spread of individual mutations, genes and mobile genetic elements can also be reconstructed in a bacterial phylogeny using this method (see figure below).



© Wellcome Genome Campus Advanced Courses and Scientific Conferences

Exercises on interpreting phylogenetic trees

This section includes questions on inferring genetic relatedness using phylogenetic trees.



Question 1. based on the tree above, what internal node corresponds to the most recent common ancestor of samples 8 and 10:

- Node F
- Node D
- Sample 7
- Node E

Question 2. Based on the tree above, which group of samples are most closely related:Samples 1 to 5

- Samples 6 & 7
- Samples 6 to 10
- Samples 8 & 9

Question 3. Based on the tree above, which of the following statements referring to sample 10 is more accurate:

- Sample 10 is more closely related to sample 7 than to sample 8
- Sample 10 is more closely related to sample 8 than to sample 7
- Sample 10 is equally related to sample 7 and sample 8
- Sample 10 is related to sample 8, but it is not related to sample 7

Question 4. Based on the tree above, which of the following statements referring to sample 7 is more accurate:

- Sample 7 is more closely related to sample 8 than to sample 10
- Sample 7 is more closely related to sample 10 than to sample 8
- Sample 7 is equally related to sample 8 and sample 10
- Sample 7 is related to sample 8, but it is not related to sample 10

Question 5. Based on the country of origin of samples on the tree above, which of the following statements about transmission events is more certain:

- The common ancestor of samples 6 to 10 (node D) most likely circulated in country A first and later on transmitted to country B and C
- The common ancestor of samples 6 to 10 (node D) most likely circulated in country B first and later on transmitted to country C
- The common ancestor of samples 6 to 10 (node D) most likely circulated in country C first and later on transmitted to country B
- The common ancestor of samples 6 to 10 (node D) could have circulated in country A or B

Question 6. Based on the country of origin of samples on the tree above, which of the following statements about transmission events is more certain:

- The common ancestor of samples 1 to 10 (node A) most likely circulated in country A first and later on transmitted to country B and C
- The common ancestor of samples 1 to 10 (node A) most likely circulated in country B first and later on transmitted to country A and C
- The common ancestor of samples 1 to 10 (node A) most likely circulated in country C first and later on transmitted to country A and B
- The common ancestor of samples 1 to 10 (node A) could have circulated in country A or B

Answers to exercises on interpreting phylogenetic trees:

Question 1: ‘Node E’ is the correct answer. ‘Node F’ is an ancestor of sample 8 but not of sample 10. ‘Node D’ is a common ancestor of samples 8 and 10, but it is more ancient common ancestor than ‘node E’. ‘Sample 7’ is a living specimen and is not an ancestor.

Question 2: Samples 1 to 5 is the correct answer. Remember that in a tree represented as a rectangular layout, the length of horizontal lines (branches) represent genetic distances whereas vertical lines are only used to connect horizontal lines. In the tree above, samples 1 to 5 have the shortest branches connecting them to their common ancestor (node B).

Question 3: Sample 10 is more closely related to sample 8 than to sample 7. The most recent common ancestor of samples 10 and 8 is at node E, whereas the most common ancestor of samples 10 and 7 is at node D, which is a deeper (more ancestral) internal node in the tree.

Question 4: Sample 7 is equally related to sample 8 and sample 10. The most recent common ancestor of samples 7 and 8 is at node D, as is the most recent common ancestor of sample 7 and 10. All descendants of node E are equally related to sample 7.

Question 5: The common ancestor of samples 6 to 10 (node D) most likely circulated in country B first and later on transmitted to country C. Country B is the most likely origin of the common ancestor represented by ‘node D’ because its direct descendants (‘node C’ and ‘node E’) both contain samples collected on this country. Later on, one clone, represented by ‘node F’, transmitted from country B to C.

Question 6: The common ancestor of samples 1 to 10 (node A) could have circulated in countries A or B. Unfortunately, information on the country of origin of contextual samples descendants from more ancestral nodes to ‘Node A’ are needed to draw more accurate conclusions.

Exercises on making a phylogeny (manually)

There are several methods to construct a phylogenetic tree from the sequence alignment. To demonstrate this, you will next construct a phylogeny by hand using one of a distance method called unweighted pair-group method with arithmetic Mean (UPGMA).

Step 1 Make the sequence alignment which had been prepared for you below

A: A T C G T G G

B: A T C G T G G

C: A T C C C T T

D: A T C C C T C

E: C C G C A G T

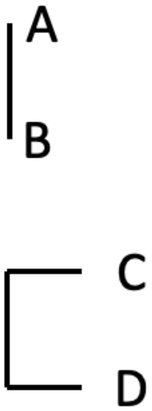
Step 2 Compare pair-wise differences between sequences
For example, there is a single nucleotide difference between sequence A and B. Please count the number of base differences between all possible sequences and fill in the table below.

	A	B	C	D	E
A		0	4	4	6
B			4	4	6
C				1	6
D					6
E					

Next, identify the sequences with fewest differences between them.
These are sequences that are the most closely related. From the table, the most closely related sequences are A and B (0 SNP difference); and C and D (1 SNP difference). The SNP difference is marked in red.

	A	B	C	D	E
A		**0**	4	4	6
B			4	4	6
C				**1**	6
D					6
E					

We can next draw the first grouping on the phylogenetic tree. We can group A and B together to show their close relationship. Similarly, we can group C and D together.



Step 3 Compare average pair-wise differences between grouped sequences
With the first two groupings made on the tree we will next rework the table with the grouped sequences rather than two individual entries in the table.

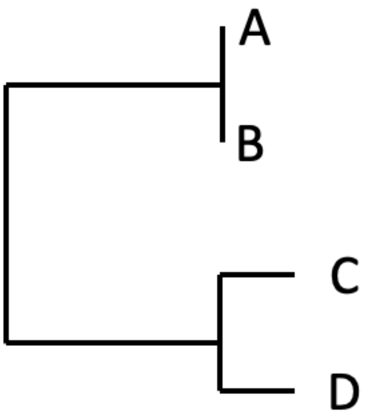
Grouped sequences A and B

	A/B	C	D	E
A/B		4	4	6
C			1	6
D				6
E				

We can next proceed to C and D grouping

	A/B	C/D	E
A/B		**4**	6
C/D			6
E			

We can identify the next closely related group which is A/B and C/D. We can draw this group together.



Finally, we can add E which is equally distant to A/B and C/D as below.