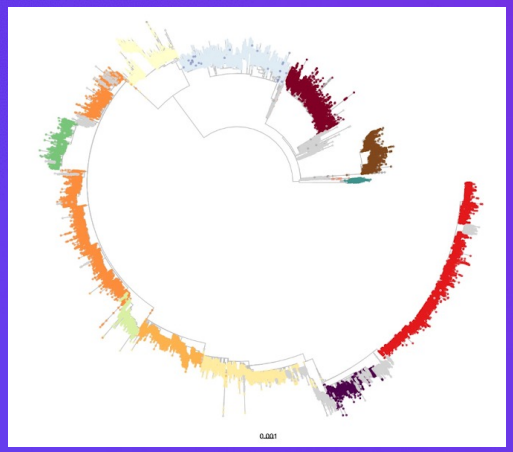
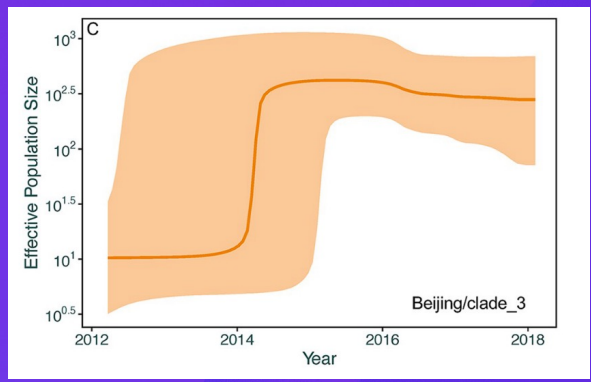
Genomic Analysis and Phylodynamics

Workshop: Simon Fraser University – 5th – 9th February 2024





Instructor: Dr. Ben Sobkowiak
Yale University / University College London

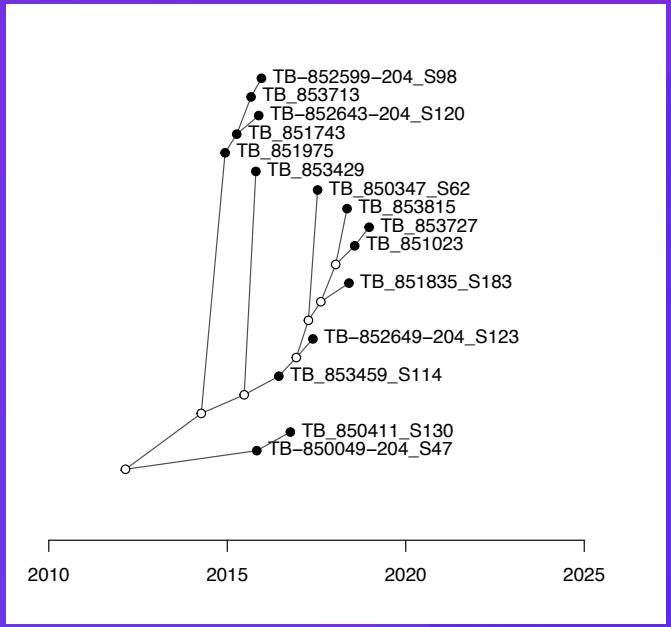
Lecture 2: Variant calling and phylogenetic trees

- What can variation within the genome tell us?
- What are genomic variants?
- How do we detect variation in the genome?
- Linking variation to evolutionary relationships using phylogenetic trees

Lecture 2: Variant calling and phylogenetic trees

What can variation in the genome tell us?

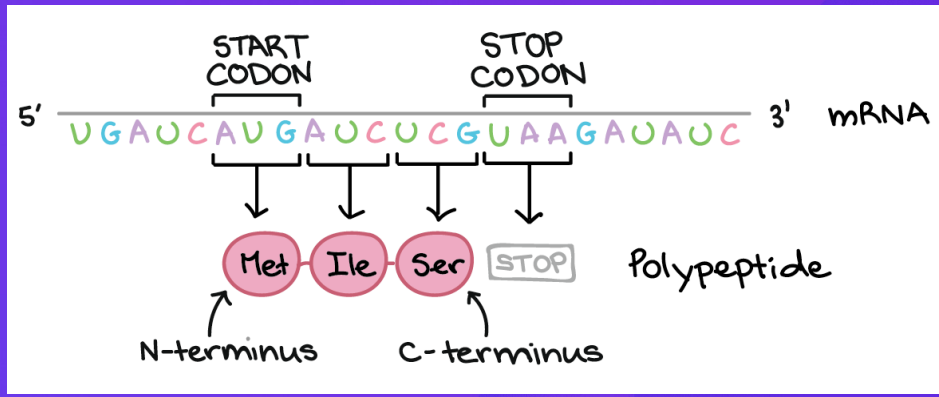
- Insights into the evolution of organisms and selection
  - Variations in the genome contribute to the adaptation of species to their environment
- Specific mutations may alter the protein that is coded for by a gene and change characteristics
- Analysing the amount and patterns of differences can relate to the amount of divergence and common ancestry between individuals
  - Can be linked to evolutionary history and transmission



Lecture 2: Variant calling and phylogenetic trees

How the genome is read

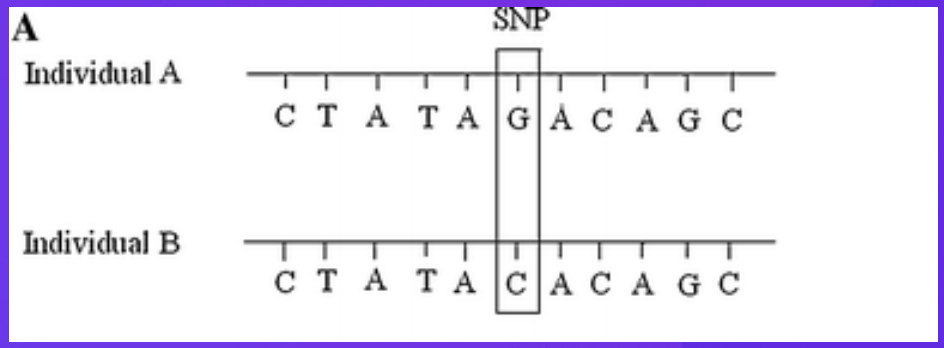
- Process of reading the genome is transcription
- RNA polymerase binds to promotors (upstream of a gene) and then 'reads' the genetic code to make a complimentary RNA strand
- This RNA strand is then translated into proteins by building amino acid blocks coded for by the sequence
- A specific sequence of three nucleotide bases (codon) encodes an amino acid



Lecture 2: Variant calling and phylogenetic trees

Single nucleotide polymorphisms

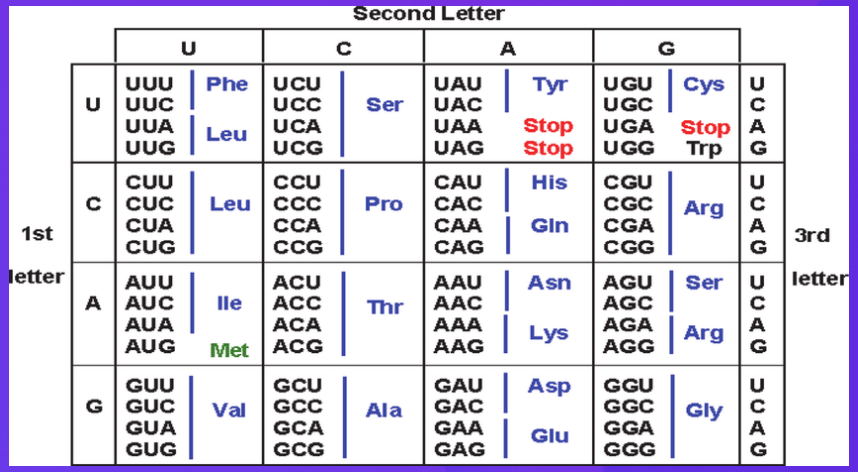
- Single Nucleotide Polymorphisms, or SNPs, are the most common type of genetic variation
- A point mutation where a single nucleotide base (A, C, G, or T) in the DNA sequence is replaced by one of the other three bases at the same position
- Most common form of genomic variation



Lecture 2: Variant calling and phylogenetic trees

Single nucleotide polymorphisms

- SNPs can result in a change in the amino acid that is encoded (non-synonymous) or still code for the same amino acid (synonymous)
- There are 64 codons but only 20 amino acids (+ start and stop codons)

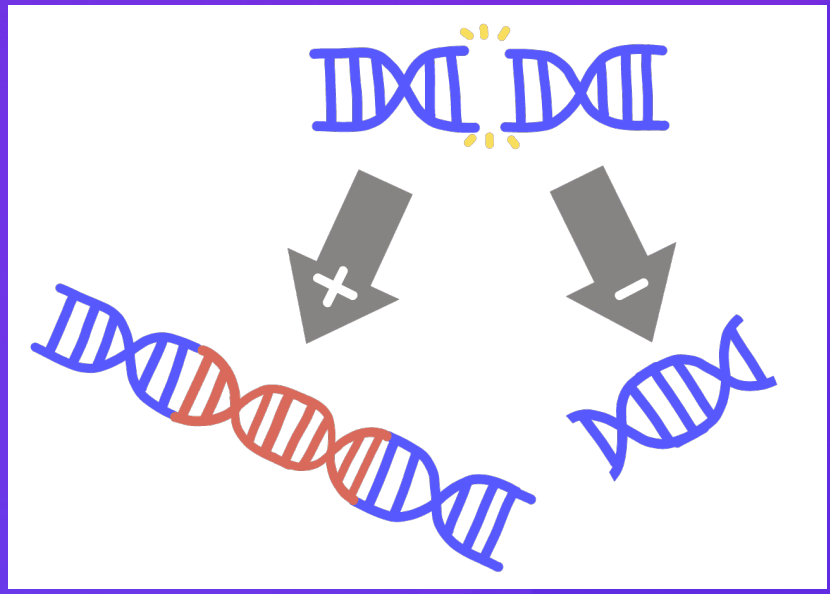


1st letter	Second Letter: U		Second Letter: C		Second Letter: A		Second Letter: G		3rd letter
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

Lecture 2: Variant calling and phylogenetic trees

Insertions and deletions (Indels)

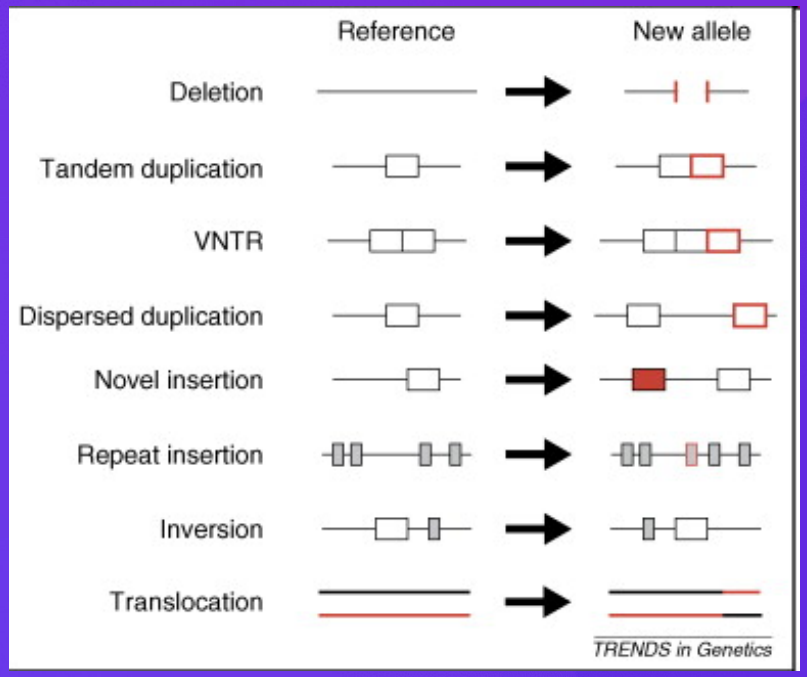
- Often abbreviated as Indels
- The addition or removal of one or more nucleotide bases from a DNA sequence
- Can have significant effects on the structure and function of genes and can contribute to genetic diversity within populations
- Can have most impact when they are ‘frameshift’ indels – will change the codon position



Lecture 2: Variant calling and phylogenetic trees

Other genomic variants

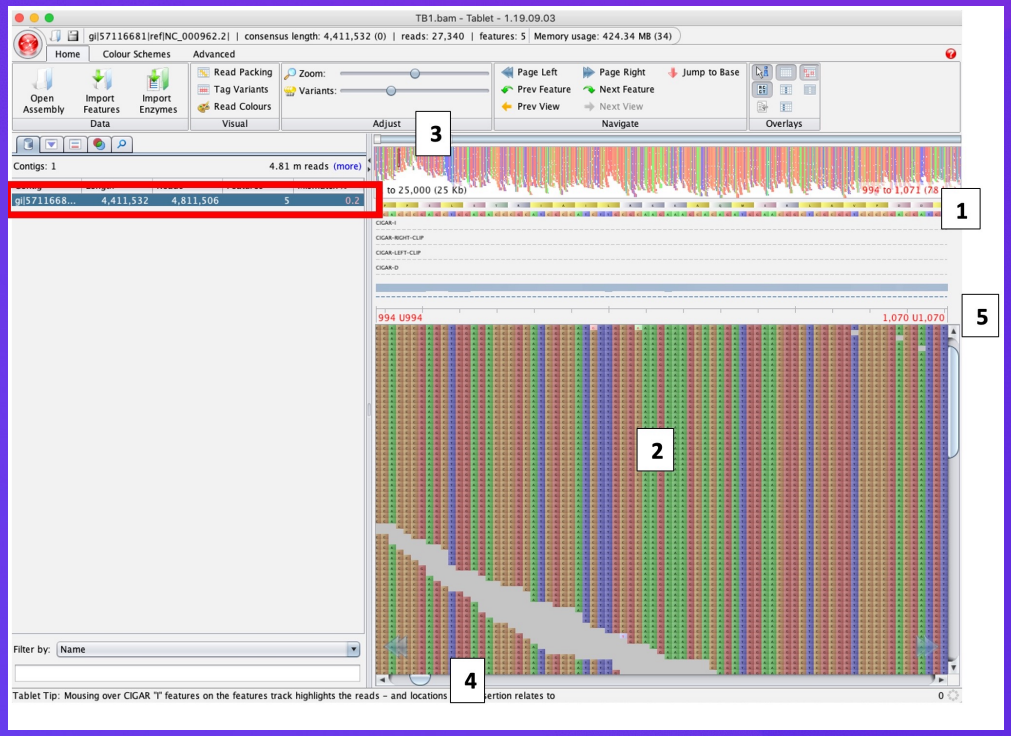
- Gene duplications
- Inversions
- Translocations
- Rarer, and more difficult to detect and analyze



Lecture 2: Variant calling and phylogenetic trees

Assembled/aligned genomes

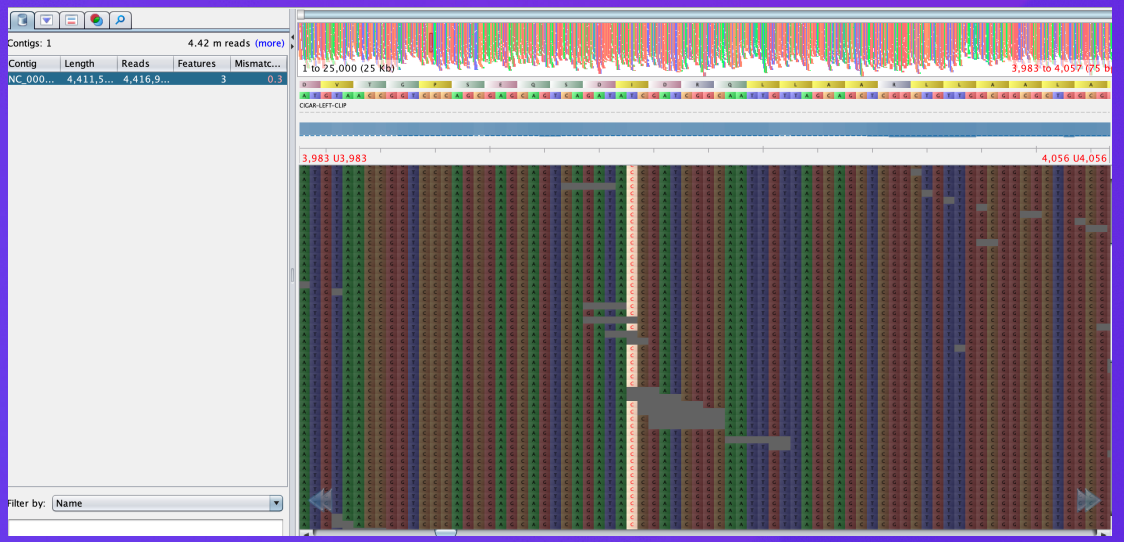
- We now have our assembled or aligned genome
- The file format is called a BAM (or SAM) file
  - A TAB-delimited text format consisting of an optional header and an alignment.
- Minimum format agreed on to report sequencing results, and includes all the data in a *fastq* file



Lecture 2: Variant calling and phylogenetic trees

Assembled/aligned genomes

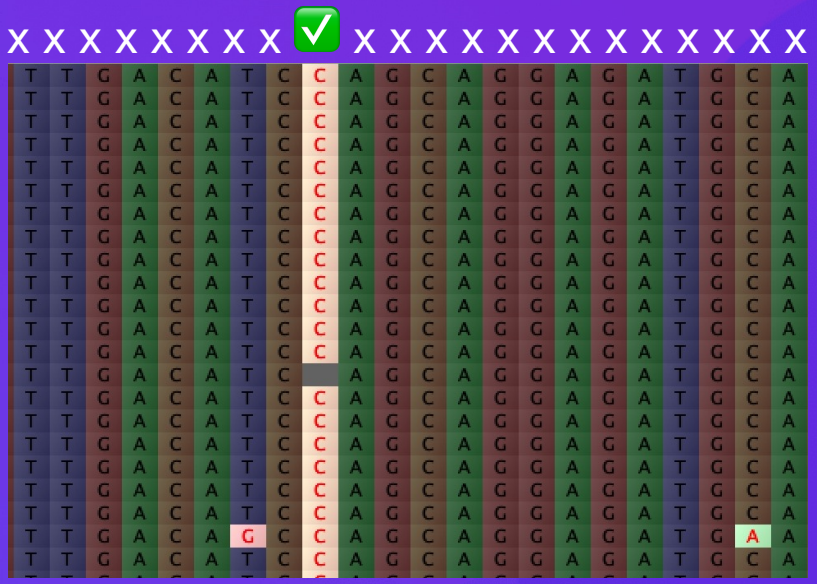
- Could scroll through the alignment file to detect variation
- But, time-consuming and subjective – how do we decide between variation and error?
- We can use tools to read through the BAM file and identify true variation



Lecture 2: Variant calling and phylogenetic trees

Identifying variants

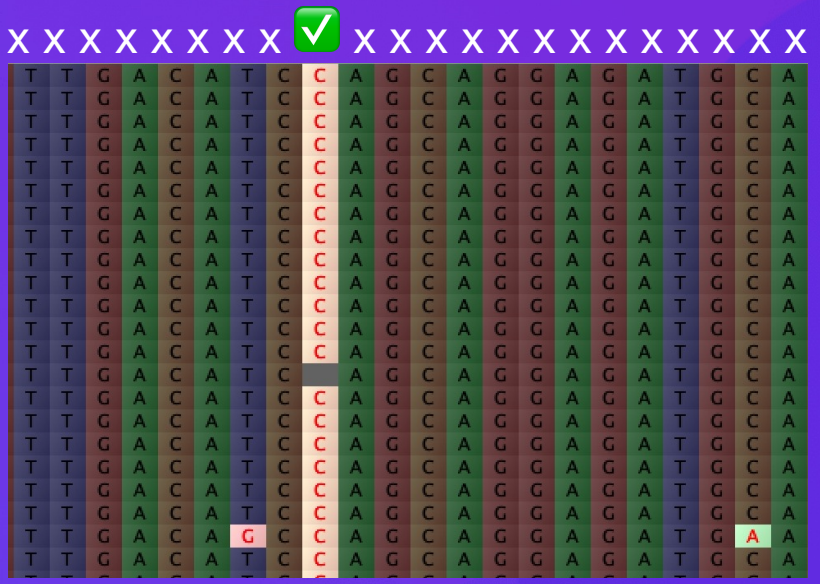
- We use modelling approaches to identify variants between the reference and sequences of samples, and between samples.
- Need to maximize sensitivity/specificity of calls by accounting for sequence error, poor alignment, low quality etc.



Lecture 2: Variant calling and phylogenetic trees

Variant calling software

- Multiple variant calling software are available including:
    - GATK
    - SAM/BCFtools
    - Delly
    - Pilon
    - FreeBayes
    - DeepVariant
- Have different strengths depending on the type of variants to call



Lecture 2: Variant calling and phylogenetic trees

GATK

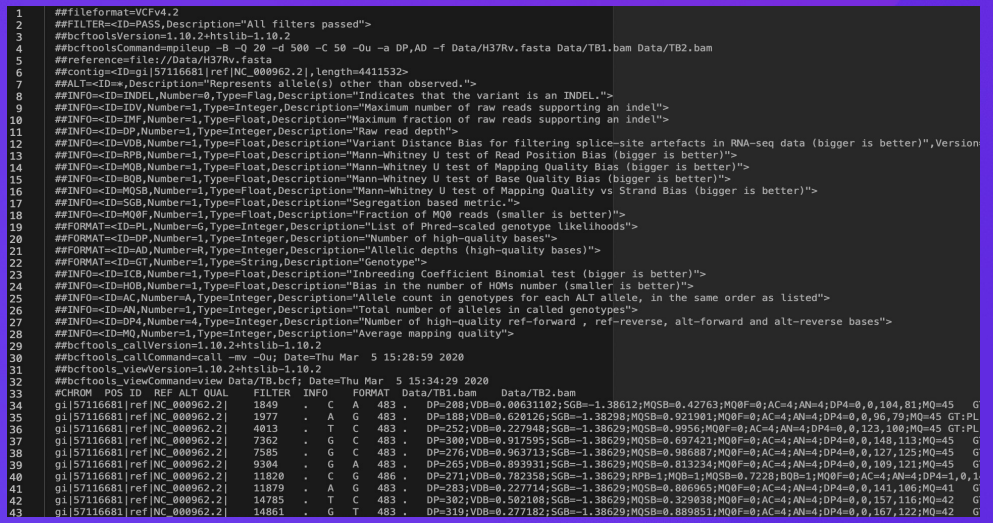
BAM file(s) -> Machine learning model to recalibrate base quality scores -> Bayesian genotype likelihood model to call variants at each genomic position -> Filters variant calls to reduce false positives -> Quality Control and Validation (manual)

Lecture 2: Variant calling and phylogenetic trees

Variant Call Format (VCF) files

- The Variant Call Format (VCF) is a widely used file format in genomics for storing information about genetic variants, such as single nucleotide polymorphisms (SNPs), insertions, deletions, and other types of genetic variation.

- Important part is the Genotype data - the alleles carried by each individual at the variant site.
    - e.g., "0/0" for homozygous reference, "0/1" for heterozygous, "1/1" for homozygous alternate

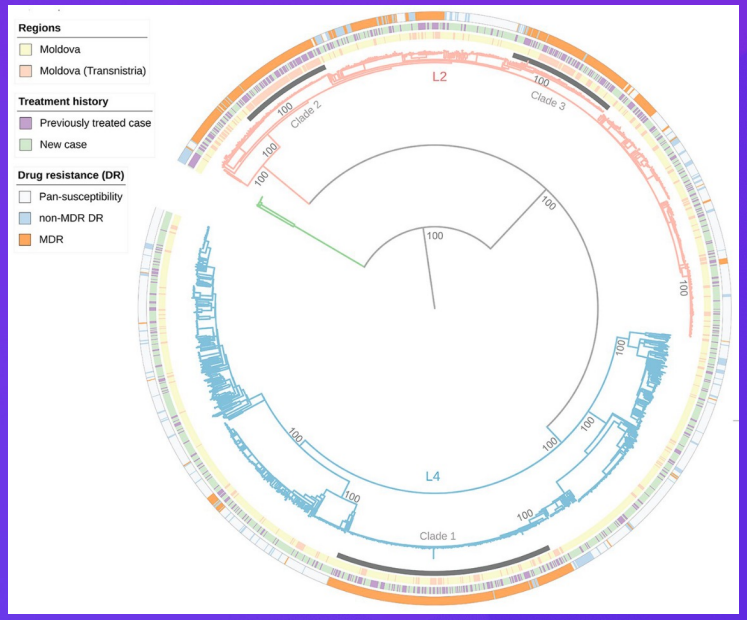


```
##fileformat=VCFv4.2
##FILTER=<ID=PASS,Description="All filters passed">
##bcftoolsVersion=1.10.2+htslib-1.10.2
##bcftoolsCommand=mpileup -B -Q 20 -d 500 -C 50 -Ou -a DP,AD -f Data/H37Rv.fasta Data/TB1.bam Data/TB2.bam
##reference=file://Data/H37Rv.fasta
##contig=<ID=gi|57116681|ref|NC_000962.2|,length=4411532>
##ALT=<ID=*,Description="Represents allele(s) other than observed.">
##INFO=<ID=INDEL,Number=0,Type=Flag,Description="Indicates that the variant is an INDEL.">
##INFO=<ID=IDV,Number=1,Type=Integer,Description="Maximum number of raw reads supporting an indel">
##INFO=<ID=IMF,Number=1,Type=Float,Description="Maximum fraction of raw reads supporting an indel">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO=<ID=VDB,Number=1,Type=Float,Description="Variant Distance Bias for filtering splice-site artefacts in RNA-seq data (bigger is better)",Version
##INFO=<ID=RPB,Number=1,Type=Float,Description="Mann-Whitney U test of Read Position Bias (bigger is better)">
##INFO=<ID=MQB,Number=1,Type=Float,Description="Mann-Whitney U test of Mapping Quality Bias (bigger is better)">
##INFO=<ID=BQB,Number=1,Type=Float,Description="Mann-Whitney U test of Base Quality Bias (bigger is better)">
##INFO=<ID=MQSB,Number=1,Type=Float,Description="Mann-Whitney U test of Mapping Quality vs Strand Bias (bigger is better)">
##INFO=<ID=SGB,Number=1,Type=Float,Description="Segregation based metric.">
##INFO=<ID=MQ0F,Number=1,Type=Float,Description="Fraction of MQ0 reads (smaller is better)">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="List of Phred-scaled genotype likelihoods">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Number of high-quality bases">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths (high-quality bases)">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##INFO=<ID=ICB,Number=1,Type=Float,Description="Inbreeding Coefficient Binomial test (bigger is better)">
##INFO=<ID=HOB,Number=1,Type=Float,Description="Bias in the number of HOMs number (smaller is better)">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes for each ALT allele, in the same order as listed">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=DP4,Number=4,Type=Integer,Description="Number of high-quality ref-forward , ref-reverse, alt-forward and alt-reverse bases">
##INFO=<ID=MQ,Number=1,Type=Integer,Description="Average mapping quality">
##bcftools_callVersion=1.10.2+htslib-1.10.2
##bcftools_callCommand=call -mv -Ou; Date=Thu Mar  5 15:28:59 2020
##bcftools_viewVersion=1.10.2+htslib-1.10.2
##bcftools_viewCommand=view Data/TB.bcf; Date=Thu Mar  5 15:34:29 2020
#CHROM  POS ID  REF ALT QUAL    FILTER  INFO    FORMAT  Data/TB1.bam    Data/TB2.bam
gi|57116681|ref|NC_000962.2|    1849    .   C   A   483 .   DP=208;VDB=0.00631102;SGB=-1.38612;MQSB=0.42763;MQ0F=0;AC=4;AN=4;DP4=0,0,104,81;MQ=45   G
gi|57116681|ref|NC_000962.2|    1977    .   A   G   483 .   DP=188;VDB=0.620126;SGB=-1.38298;MQSB=0.921901;MQ0F=0;AC=4;AN=4;DP4=0,0,96,79;MQ=45 GT:PL
gi|57116681|ref|NC_000962.2|    4013    .   T   C   483 .   DP=252;VDB=0.227948;SGB=-1.38629;MQSB=0.9956;MQ0F=0;AC=4;AN=4;DP4=0,0,123,100;MQ=45 GT:PL
gi|57116681|ref|NC_000962.2|    7362    .   G   C   483 .   DP=300;VDB=0.917595;SGB=-1.38629;MQSB=0.697421;MQ0F=0;AC=4;AN=4;DP4=0,0,148,113;MQ=45   G
gi|57116681|ref|NC_000962.2|    7585    .   G   C   483 .   DP=276;VDB=0.963713;SGB=-1.38629;MQSB=0.986887;MQ0F=0;AC=4;AN=4;DP4=0,0,127,125;MQ=45   G
gi|57116681|ref|NC_000962.2|    9304    .   G   A   483 .   DP=265;VDB=0.893931;SGB=-1.38629;MQSB=0.813234;MQ0F=0;AC=4;AN=4;DP4=0,0,109,121;MQ=45   G
gi|57116681|ref|NC_000962.2|    11820   .   C   G   486 .   DP=271;VDB=0.782358;SGB=-1.38629;RPB=1;MQB=1;MQSB=0.7228;BQB=1;MQ0F=0;AC=4;AN=4;DP4=1,0,1
gi|57116681|ref|NC_000962.2|    11879   .   A   G   483 .   DP=283;VDB=0.227714;SGB=-1.38629;MQSB=0.806965;MQ0F=0;AC=4;AN=4;DP4=0,0,141,106;MQ=41   G
gi|57116681|ref|NC_000962.2|    14785   .   T   C   483 .   DP=302;VDB=0.502108;SGB=-1.38629;MQSB=0.329038;MQ0F=0;AC=4;AN=4;DP4=0,0,157,116;MQ=42   G
gi|57116681|ref|NC_000962.2|    14861   .   G   T   483 .   DP=319;VDB=0.277182;SGB=-1.38629;MQSB=0.889851;MQ0F=0;AC=4;AN=4;DP4=0,0,167,122;MQ=42   G
```

Lecture 2: Variant calling and phylogenetic trees

Phylogenetic trees

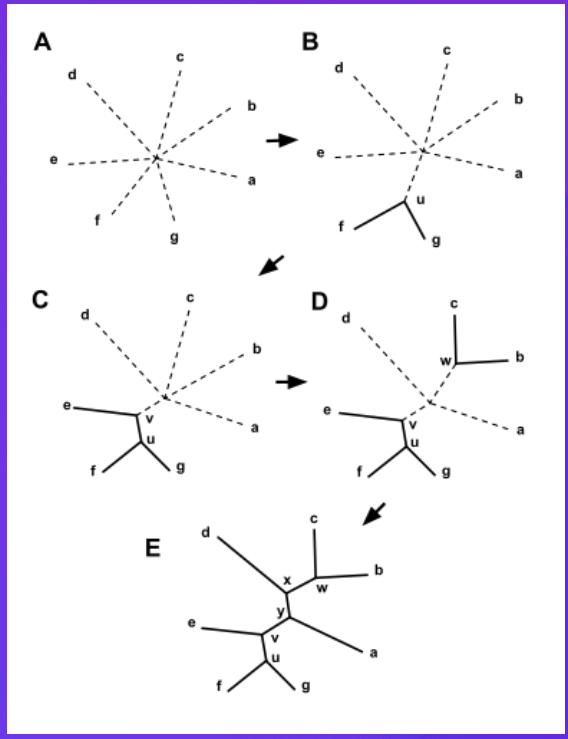
- A phylogenetic tree is a diagram that represents evolutionary relationships among organisms
- They illustrate the ancestral lineage and divergence of species, genes, or other taxonomic units.
- These trees help in understanding evolutionary history, inferring patterns of descent, and clarifying the timing of evolutionary events.



Lecture 2: Variant calling and phylogenetic trees

Simple methods (Neighbour-Joining)

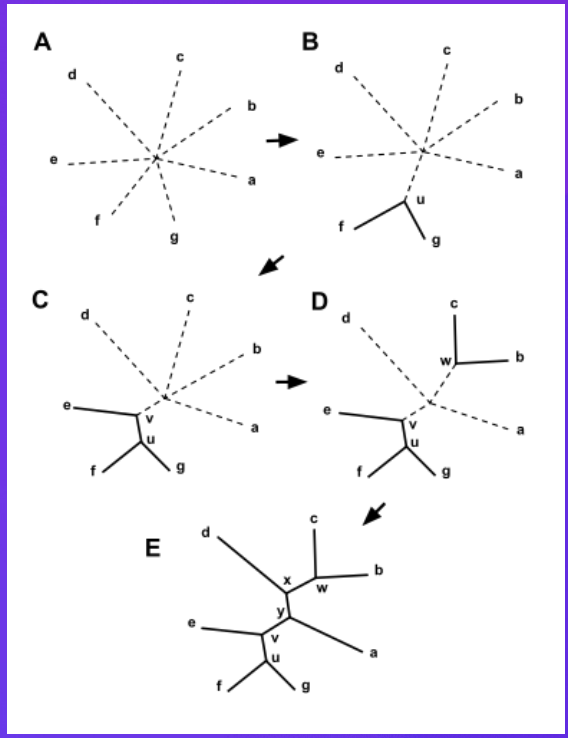
- Neighbor-Joining (NJ) is a distance-based method used to construct phylogenetic trees
- Starts with a star-like tree, where all entities are connected to a central node (A)
- The Neighbor-Joining algorithm iteratively joins entities (nodes) in the tree while minimizing the total branch length



Lecture 2: Variant calling and phylogenetic trees

Simple methods (Neighbour-Joining)

- Neighbor-Joining is a relatively efficient and versatile method for constructing phylogenetic trees
- However, it does not explicitly model the underlying evolutionary processes, such as substitutions, insertions, deletions, or other events
- This limitation makes it less suitable for analyzing complex evolutionary scenarios



Lecture 2: Variant calling and phylogenetic trees

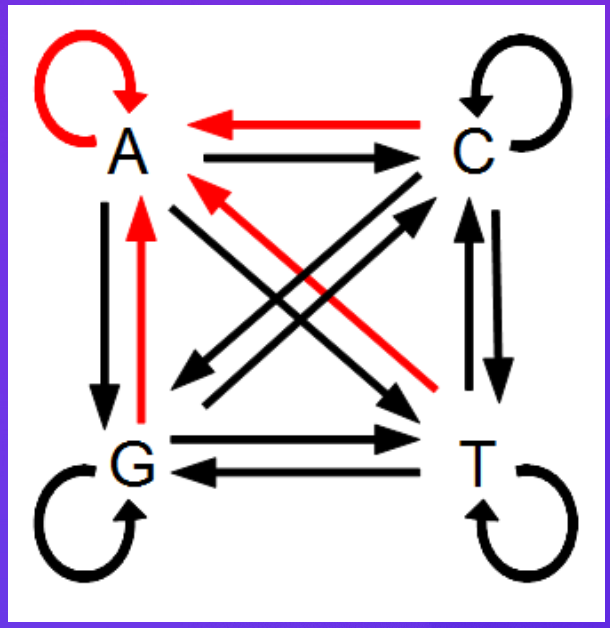
Evolutionary models

- More sophisticated tree building methods can include models:
    - Nucleotide substitution models
    - Molecular clock models
    - Population models
    - Coalescent models

Lecture 2: Variant calling and phylogenetic trees

Nucleotide substitution models

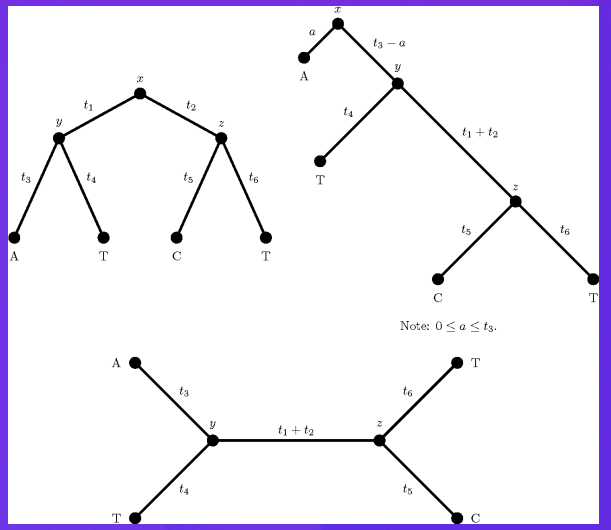
- Describe the rates at which nucleotides in DNA sequences change over time
- Provide a framework for estimating the likelihood of observed DNA sequence data on a phylogenetic tree
- Common models include:
  - Jukes-Cantor (JC) Model
  - Hasegawa-Kishino-Yano (HKY) Model
  - General Time Reversible (GTR) Model



Lecture 2: Variant calling and phylogenetic trees

More complex methods (Maximum Likelihood)

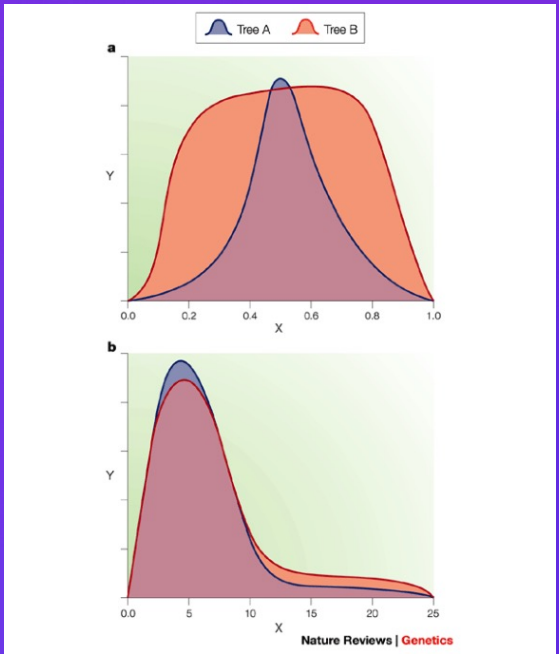
- ML tree construction aims to find the tree topology and branch lengths that maximize the likelihood of the observed sequence data under a specified evolutionary model
- ML is powerful for its statistical rigor and ability to handle complex models of evolution
- Generally considered less flexible in handling complex models compared to Bayesian methods



Lecture 2: Variant calling and phylogenetic trees

Bayesian phylogenies

- Uses Bayesian statistics to estimate the posterior probabilities of different phylogenetic trees, incorporating the observed data and prior knowledge
- Calculates the probabilities of different trees by combining the likelihood of the observed data with the prior probability distributions over tree space, based on certain models of evolution
- Require careful selection of priors and are computationally demanding.

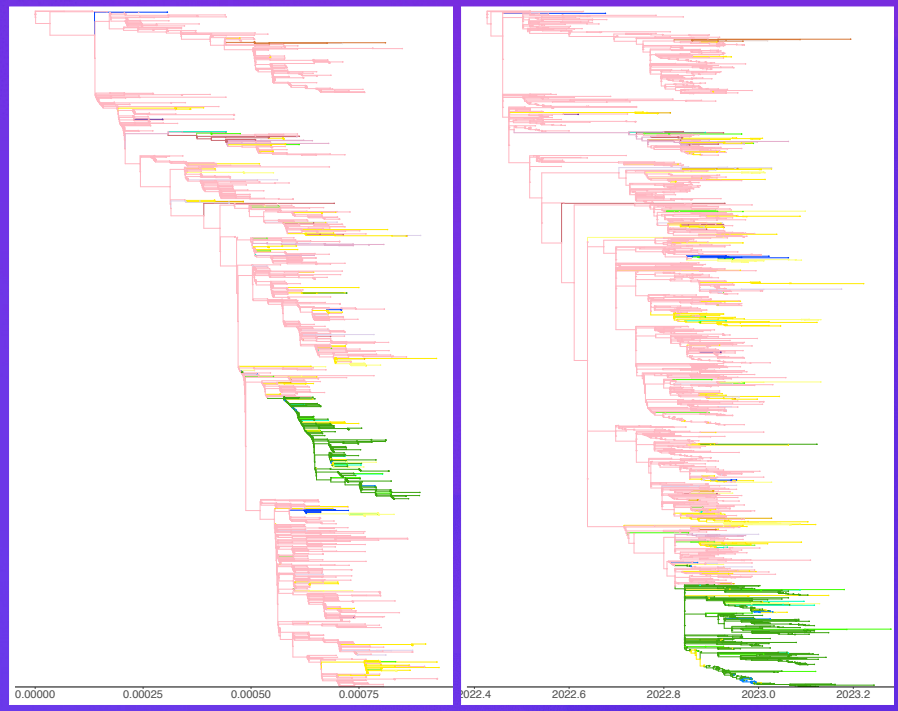


From Holder & Lewis, 2003

Lecture 2: Variant calling and phylogenetic trees

Timed vs untimed trees

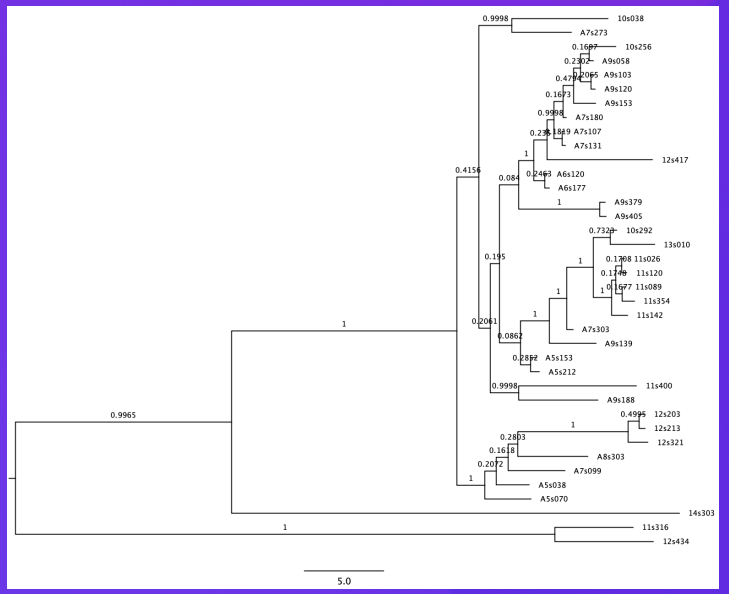
- In untimed trees, branch lengths typically represent genetic or sequence divergence
- Help to understand the evolutionary ancestry and genetic relatedness among taxa
- Timed trees incorporate the estimated timing of evolutionary events, branches are scaled to unit time
- Useful for estimating divergence times, studying temporal changes in evolutionary processes, and reconstructing the evolutionary history of lineages



Lecture 2: Variant calling and phylogenetic trees

Assessing phylogenetic trees

- Bootstrapping is a resampling technique used in phylogenetics to assess the robustness of the inferred phylogenetic tree topology
- It can estimate the reliability of the branching patterns in a phylogenetic tree
- Resamples data and builds multiple trees to assign support values to branches and nodes
- High bootstrap support values indicate the inferred relationships are likely to be accurate



Practical 2: Variant calling and maximum likelihood trees

1. Variant calling and the VCF format

2. Filter variants and construct consensus sequences and SNP matrices

3. Align consensus sequences

4. Maximum Likelihood phylogenetic trees