Module 4: Introduction to Phylogenetics and Public Health I

Ifeanyi Ezeonwumelu, PhD

PATHOGEN MULTIOMICS AND BIOINFORMATICS III

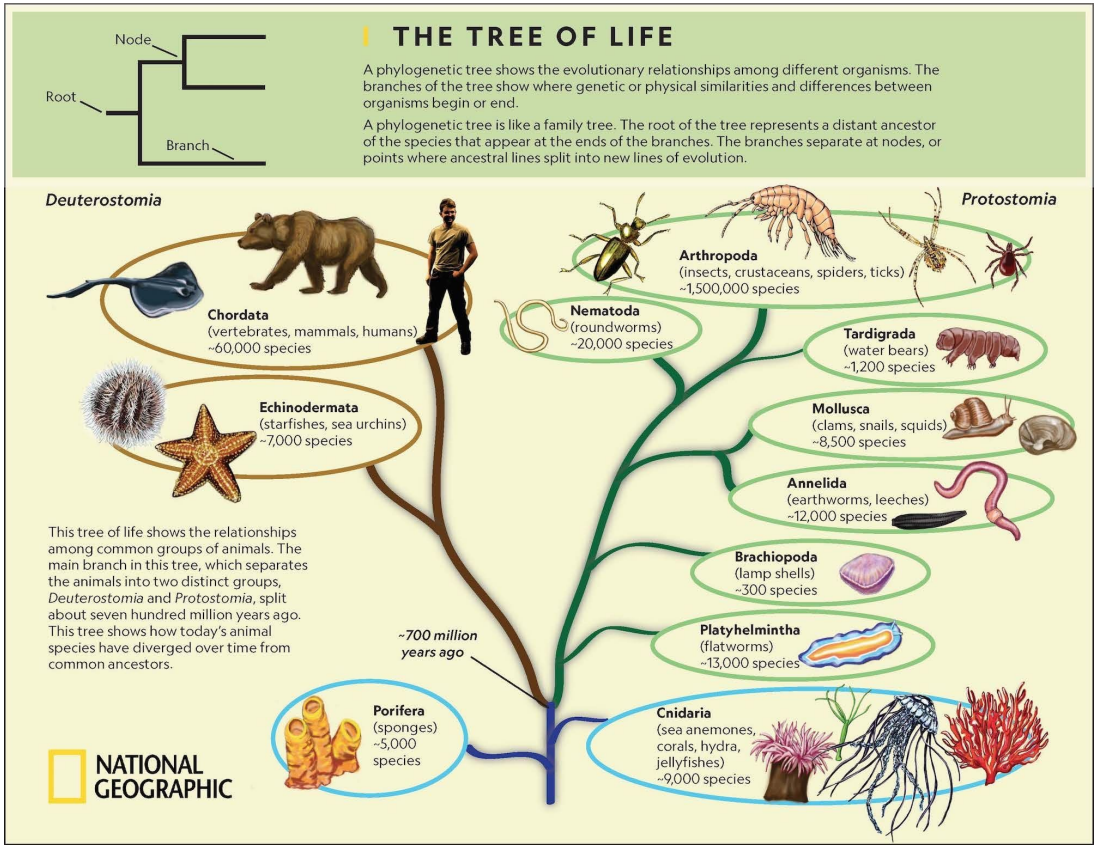
04.07.2023

GLADSTONE INSTITUTES

UCSF
University of California
San Francisco

Phylogenetics

*Phylogenetics pertains to the **study of the evolutionary relationships***



Between what?

- *organisms, e.g., species or strains*
- *genomes*
- *genes*
- *Etc.*

Phylogenetics should refer to how closely the taxa are

and…

its evolutionary history

The Phylogenetic reconstruction approach

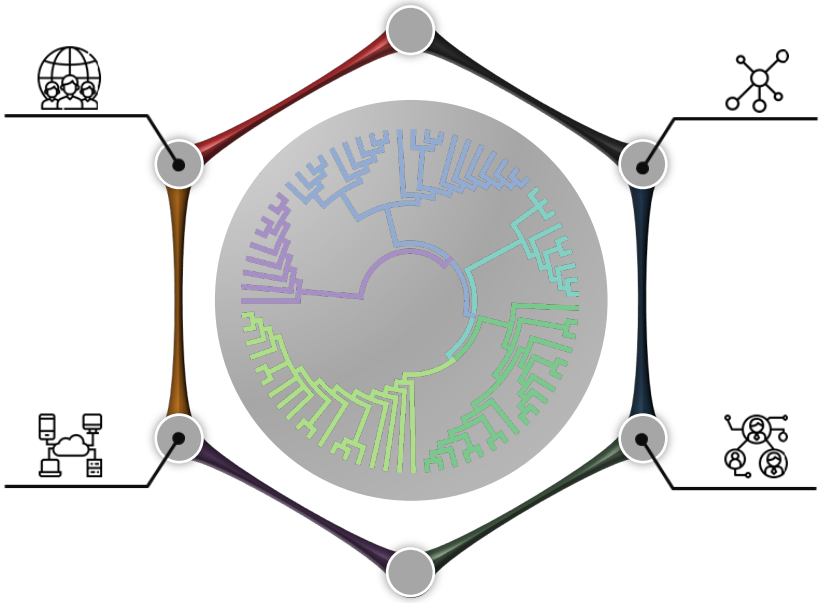
Given a dataset of DNA sequences...

DNA Sequences | Translated Protein Sequences

Species/Abbrv	
1. Rf1b EU781828	TCCGGCTCGTGGCTAAGGGAATGTTTGGGACTGGATATATGCACGGTGTT
2. Rf1b AY587016	TCCGGCTCGTGGCTCAGGGATGTTTGGGACTGGATATATGCACGGTATT
3. Rf1b EF032892	TCCGGCTCGTGGCTGAGGGATGTTTGGGACTGGATATATGCACGGTATT
4. 554M1m CAN 14	TCCGGTTCCTGGCTAAGGGACATCTGGGACTGGATATATGCGAGGTGCT
5. 561M1m CAN 14	TCcGGTTcCTGGCTAAGGGACATCTGGGACTGGATATATGCGAGGTGCT
6. 068M1m CAN 14	TCCGGcTCCTGGCTAAGGGACATCTGGGACTGGGTATATGCGAGGTACT
7. 081M1m ARA 14	TCCRGTTCCTGGWTSAGGGACATCTGGGACTGGATATATGCGAGGTGCT
8. 115M1m PAI 14	TCCGGTTCCTGGCTAAGGGACATCTGGGACTGGATATATGCGAGGTGCT
9. 906M1m MAD 14	TCCGGTTCCTGGCTGAGGGACATCTGGGACTGGATATATGCGAGGTGCT

Controls

DB controls: demographics (case subjects) age, location, social network,
Subtype controls
BLAST, LANL HIV database



Quality control

Hypermutation screening
Drug resistance mutation screening
Recombinant detection
LANL HIV database

Sequence analysis

Multiple sequence alignment
Subtyping
MUSCLE, ClustalW

Phylogenetic analysis

ML and Bay Phylogenetic trees
Hypothesis testing of tree topology
Bayesian evolutionary analysis
PhyML, BEAuti & BEAST

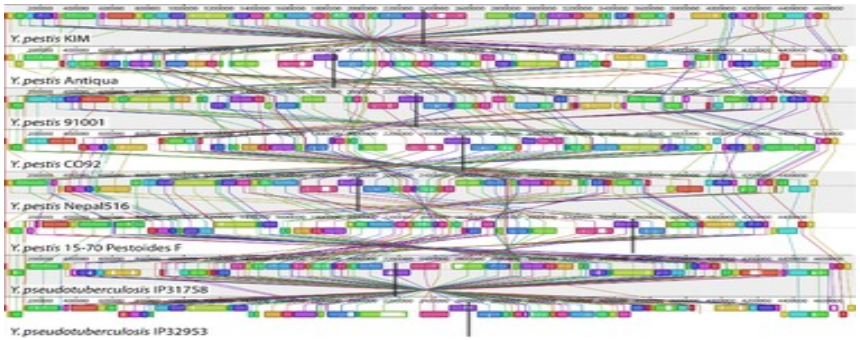
Molecular Phylogenetics

Molecular phylogenetics require a Multiple Sequence Alignment

High amount of information – each column is an evolutionary marker

Possibility of going genome-wide – compare entire genomes

Problem: *alignment of entire genomes consumes large amounts of memory, even more upon tree building*



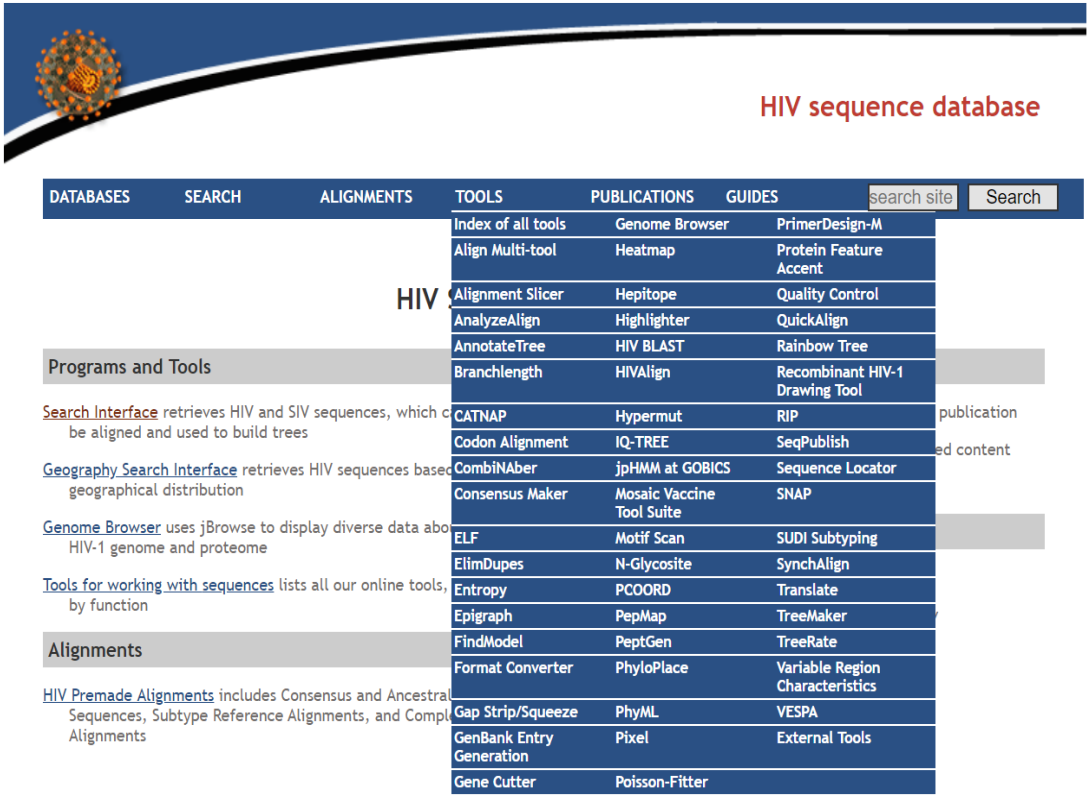
Research Questions & Trade-offs:

- *Sequence length (full length vs gene fragments)*
- *Genetic variation (conserved vs variable regions)*

Control sequences & Quality Control

LANL (Los Alamos National Laboratory, New Mexico, USA) databases: **HIV, SIV** and **HCV**
https://www.hiv.lanl.gov/content/index

A One-Stop Platform for Phylogenetic reconstruction & QC !!!



EMBL (EuropeanMolecularBiology Laboratory) database, maintained at the EMBLEBI (European Bioinformatics Institute, Hinxton, England, UK).

GenBank, maintained at the NCBI (National Center for Biotechnology Information, Bethesda, Maryland, USA).
https://www.ncbi.nlm.nih.gov/genbank/

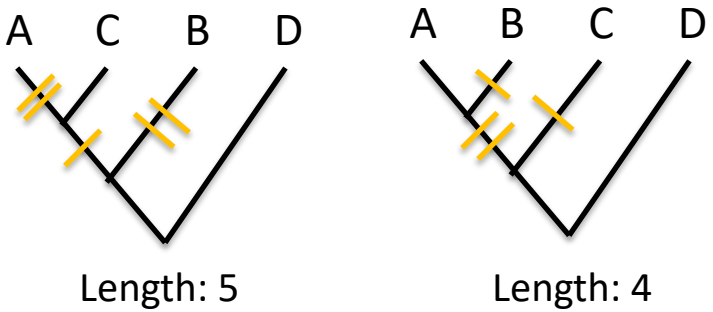
DDBJ (DNA Data Bank of Japan), maintained at the NIG/CIB (National Institute of Genetics, Center for Information Biology, Mishima, Japan).

Methods for Phylogenetic Reconstruction

Maximum Parsimony

- *Simplest possible evolution scenario – the best tree is the shortest tree*
- *The rationale is to have the tree with the least homoplasy*

A: GGG
B: GTG
C: TGT
D: TTT

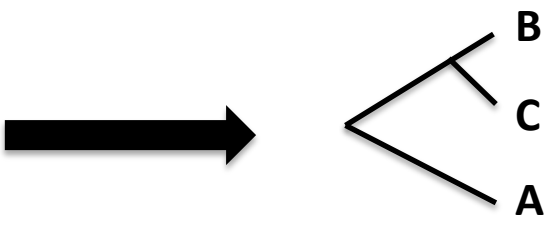


Length: 5 Length: 4

Distance Matrix Methods

- *Start by calculating pairwise distances from sequence data*
- *Tree is constructed from pairwise distances through clustering algorithms (e.g. Neighbour-Joining)*

	A	**B**	**C**
A	-	4	4
B	4	-	2
C	4	2	-



Methods for Phylogenetic Reconstruction

Maximum Likelihood

More robust approach with parameter estimation using a probabilistic model:

- *Tree topology and branch lengths; nucleotide frequencies & substitution rates*
- *Measure how well the model fits the data*

Calculates the likelihood for every column first and for the entire alignment using different permutations, random starting values.

Likelihood (Model) = Probability (Data | Model)

Maximum Likelihood – Best set of parameter values yielding the highest possible likelihood

Software: PhyML [MEGA, Seaview, IQ-TREE, etc]

Bayesian Inference

- Based on the calculation of posterior probabilities given a set of prior parameter values
- Requires starting prior value
- Involves millions of iterations and measures the convergence to parameter values over the iterations

*Bayes Rule: P (Model|Data) = P (Data | M1) * P (M1)/P (Data)*

Software: MrBayes, BEAST

Nucleotide Substitution Models

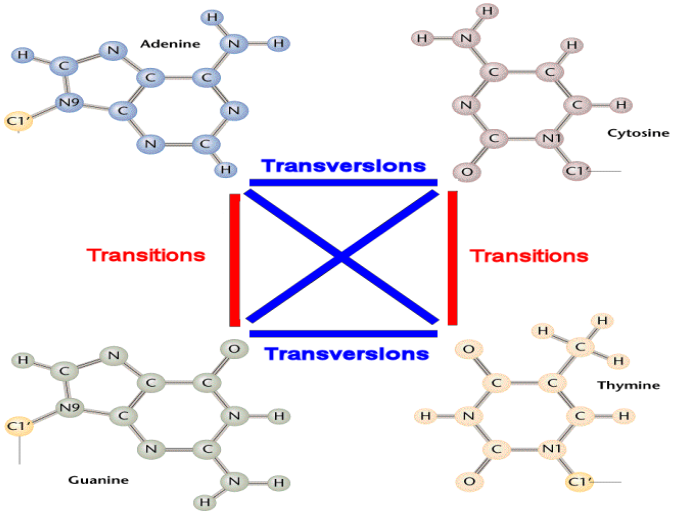
Models of Evolution

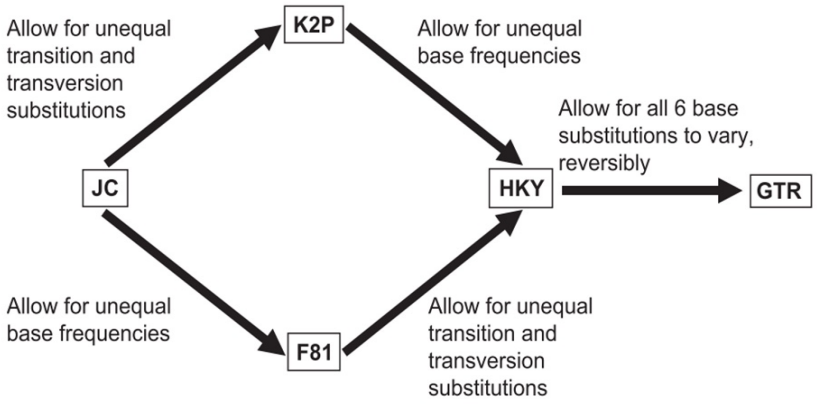
Aim to infer the number of real evolutionary events based on observed events!

Need to know how sequences are evolving – requires a model!!

Possibility of the existence of different susbstitution rates across site (gamma distribution).

The six possible substitution patterns for nucleotide data



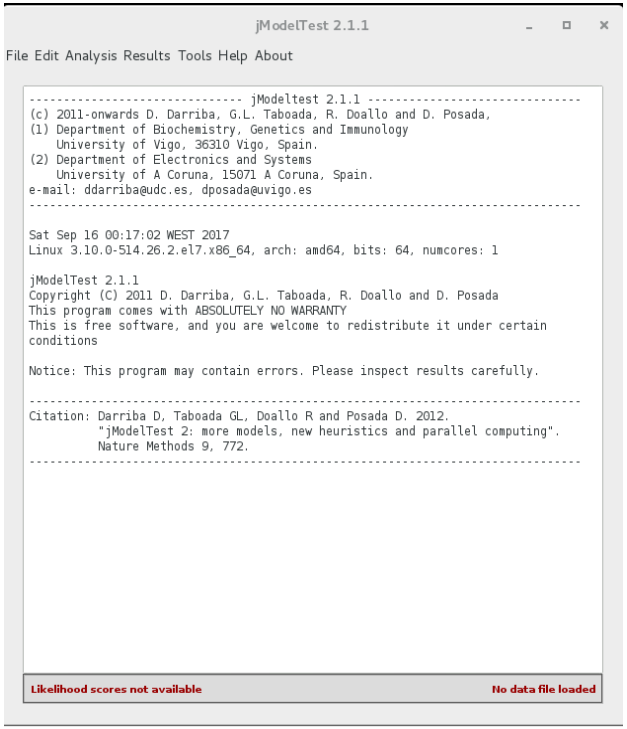


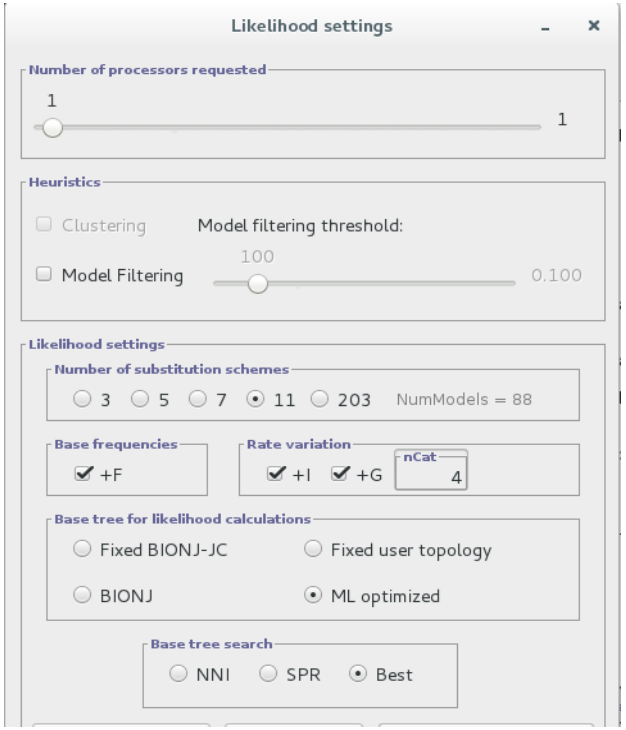
*Transitions occur at higher frequency than Transversions

How to choose the right parameters?

Answer: *test for a set of parameter permutations and choose the parameters that yields the combination that best fits your data*

jModelTest





Also possible: R, phangorn package (model.test function); ModelTest online server
Increasingly integrated in popular Tree building programs: RAxML, IQ-Tree

Tree Searching

Tree space – the number of all possible trees for a given dataset

How many branches are present in a tree with 3 tips? And with 4 tips?

Number of branches in a tree with x tips = 2x-3

3 tips ->	3 branches
4 tips ->	5 branches
5 tips ->	7 branches
10 tips ->	17 branches
20 tips ->	37 branches
40 tips ->	77 branches
100 tips ->	197 branches

But, how many possible trees:

$$\prod_{i=2}^{n-1}(2i-3)$$

For every tree with x tips it is possible to construct 2x-3 derived trees by adding an extra tip

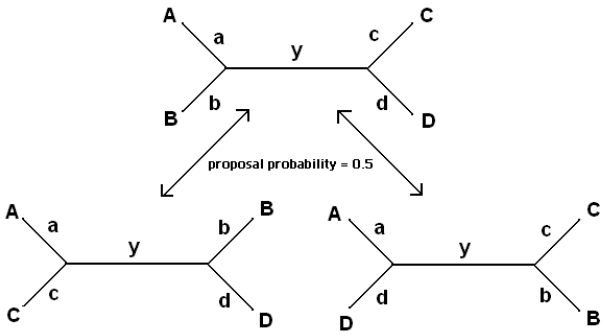
3 tips ->	1 tree
4 tips ->	3 trees
5 tips ->	15 trees
6 tips ->	105 trees
7 tips ->	945 trees
8 tips ->	10395 trees
9 tips ->	135135 trees
10 tips ->	2027025 trees
100 tips ->	1.7×10^{182} trees

It is not possible to exhaustively screen and search all possible trees in present day datasets.

Answer: start with random tree(s) and progress by making alterations and removing less likely pathways

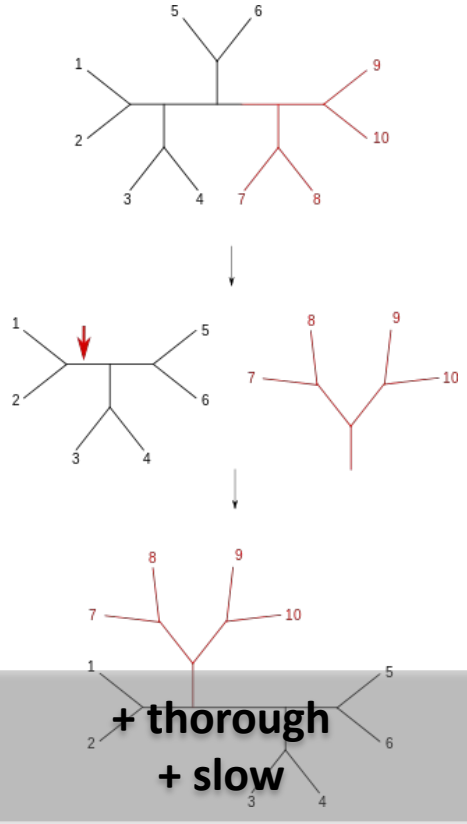
Tree Searching

Nearest Neighbour Interchange (NNI)



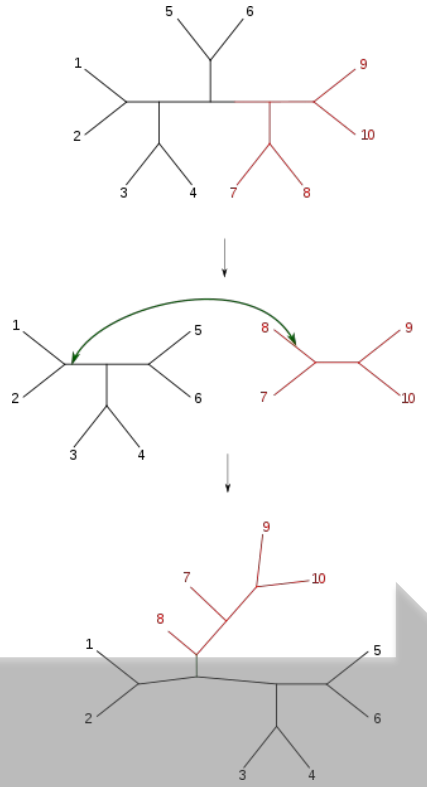
Swap neighbors at every internal branch

Subtree Prunning and Regrafting (SPR)



Cut a subtree at every possible point and regrafts at multiple points

Tree Bisection and Reconnection (TBR)

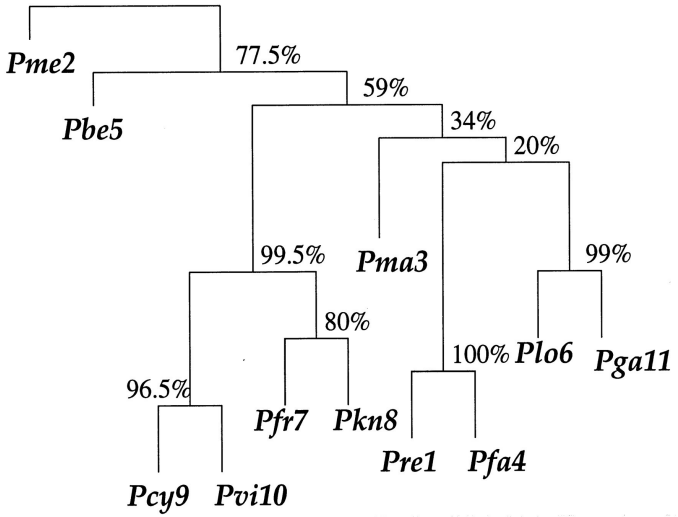


Divides tree in two parts, regrafts using every possible branch of the detached tree at all possible branches of the other tree

Tree Reliability

Bootstrap

- Widely used;
- Random sampling from the alignment (with replacement) until achieving the original length;
- Tree reconstruction n times – for each new alignment;
- Calculation for each branch the occurrence of that same clade in each tree;
- Expressed as a percentage/fraction (0-100%).
- Can be extremely time consuming



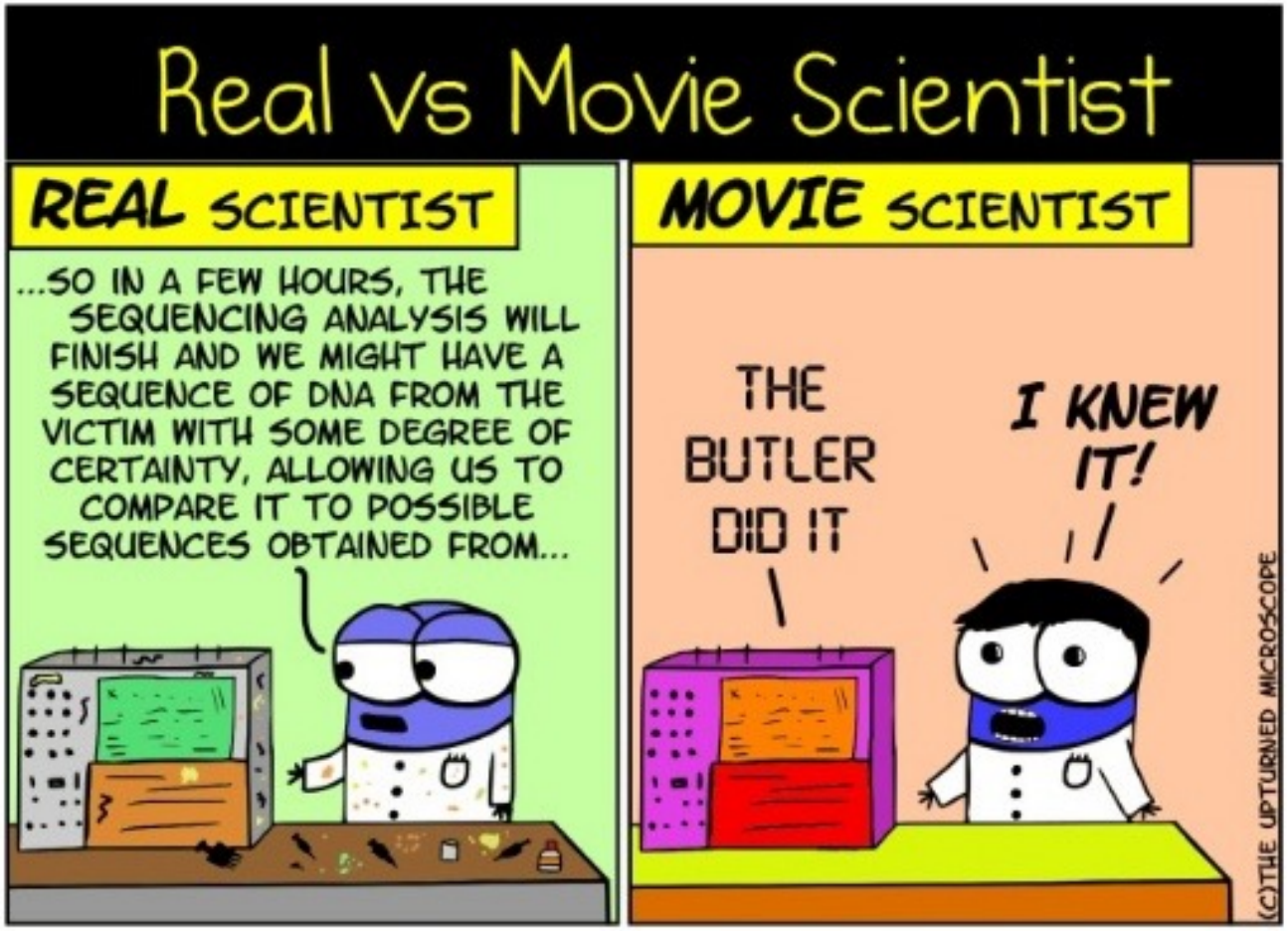
Alternatives:

- Jacknife – removes a position from the alignment (leave one out)

- aLRT – aproximate Likelihood Ratio Test – provides a *p*-value – likelihood gains between having a branch and not having a branch

Applications

DNA sequence analysis



ses;

Applications:
Criminal investigation of HIV-1 transmission

Deliberate transmission of HIV?

Phylogenetic analysis of HIV sequences from suspect and victim

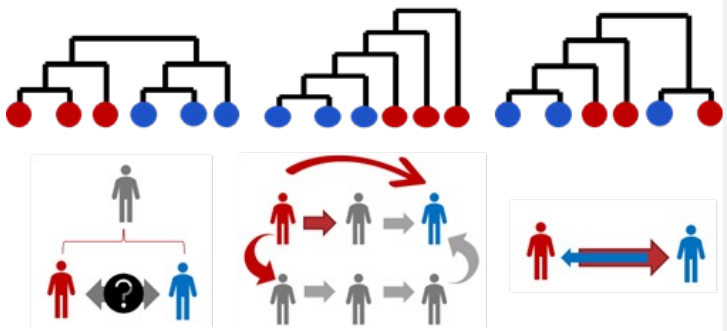
No linkage between sequences (e.g. different subtypes)

Exoneration

Sequences related by descent Direction of transmission?

Tree topology

(MM) (PM) (PP)



When did transmission occur?

Time to most recent common ancestor

Important issues:

- *Who was infected first?*
- *Suspect was aware of infection status?*
- *Suspect informed victim of infection status?*
- *Suspect was taking antiretroviral drugs?*

Suspect/Accused

Intermediary/Common ancestor

Victim

Applications:
Criminal investigation of HIV-1 transmission

EurekAlert! | AAAS

HOME NEWS RELEASES MULTIMEDIA MEETINGS

NEWS RELEASE 27-SEP-2018

Unusual case of father-to-son HIV transmission reported

Peer-Reviewed Publication

MARY ANN LIEBERT, INC./GENETIC ENGINEERING NEWS

LIVESCIENCE

Home > News

Father Transmits HIV to Newborn Son in Rare Case: How Did It Happen?

By Rachael Rettner September 28, 2018

AIDS Research and Human Retroviruses, Vol. 34, No. 10 | Transmission

Free Access

Accidental Father-to-Son HIV-1 Transmission During the Seroconversion Period

Ifeanyi Ezeonwumelu, Inês Bártolo, Francisco Martin, Ana Abecasis, Teresa Campos, Ethan O. Romero-Severson, Thomas Leitner, and Nuno Taveira

Published Online: 12 Oct 2018 | https://doi.org/10.1089/aid.2018.0060

January 2009 – Father (patient CC2) has skin blisters in face, hand and arms which the doctors associated with varicella-zoster virus infection (likely symptoms of acute HIV infection), no diagnosis of HIV infection was done

April 8 2009 – Baby (patient CC1) is born; father (CC2) is prescribed penicillin and skin blister leak out liquid (possibly with infectious HIV)

April 23 2009 – Father (CC2) is diagnosed with Syphilis

May 28 2009 – Father (CC2) is diagnosed with HIV infection

December 2012 – Baby (CC1) is diagnosed with HIV infection

Samples were collected in March 20 2013 and December 12 2013

HIV-1 seronegative Mother

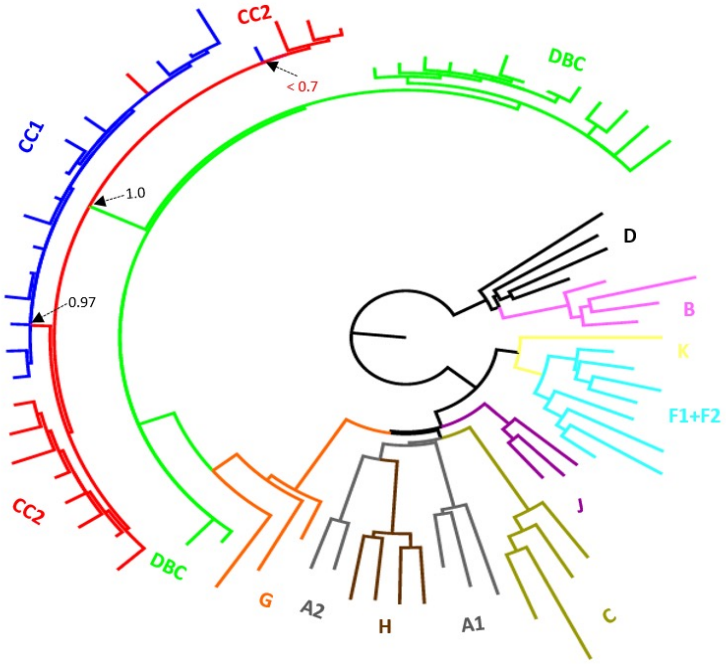
A priori hypothesis

Father-to-son HIV-1 transmission via dermal contact with HIV-1 infectious blister fluid

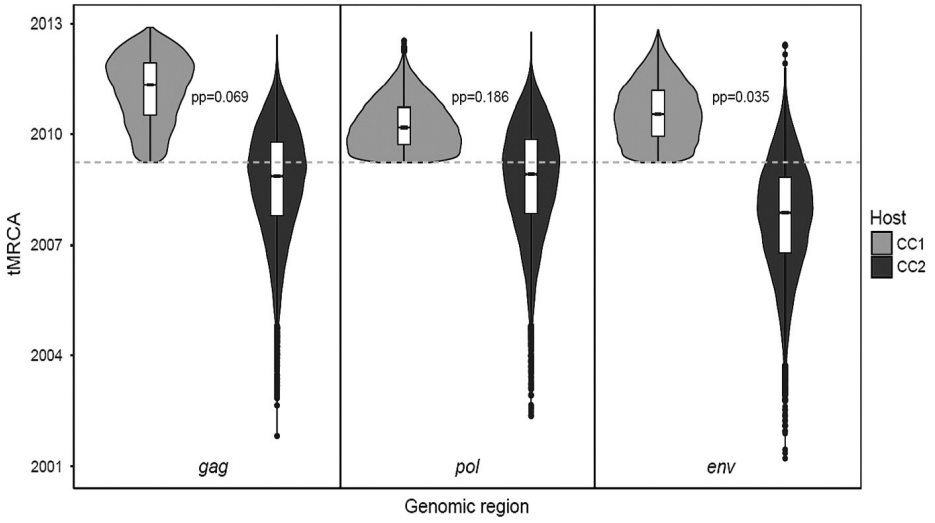
Serial sampling & sequencing: *env*, *gag* & *pol*

Applications:
Criminal investigation of HIV-1 transmission

Direction of transmission:
Father-to-son



Time of transmission:
shortly after birth of the child

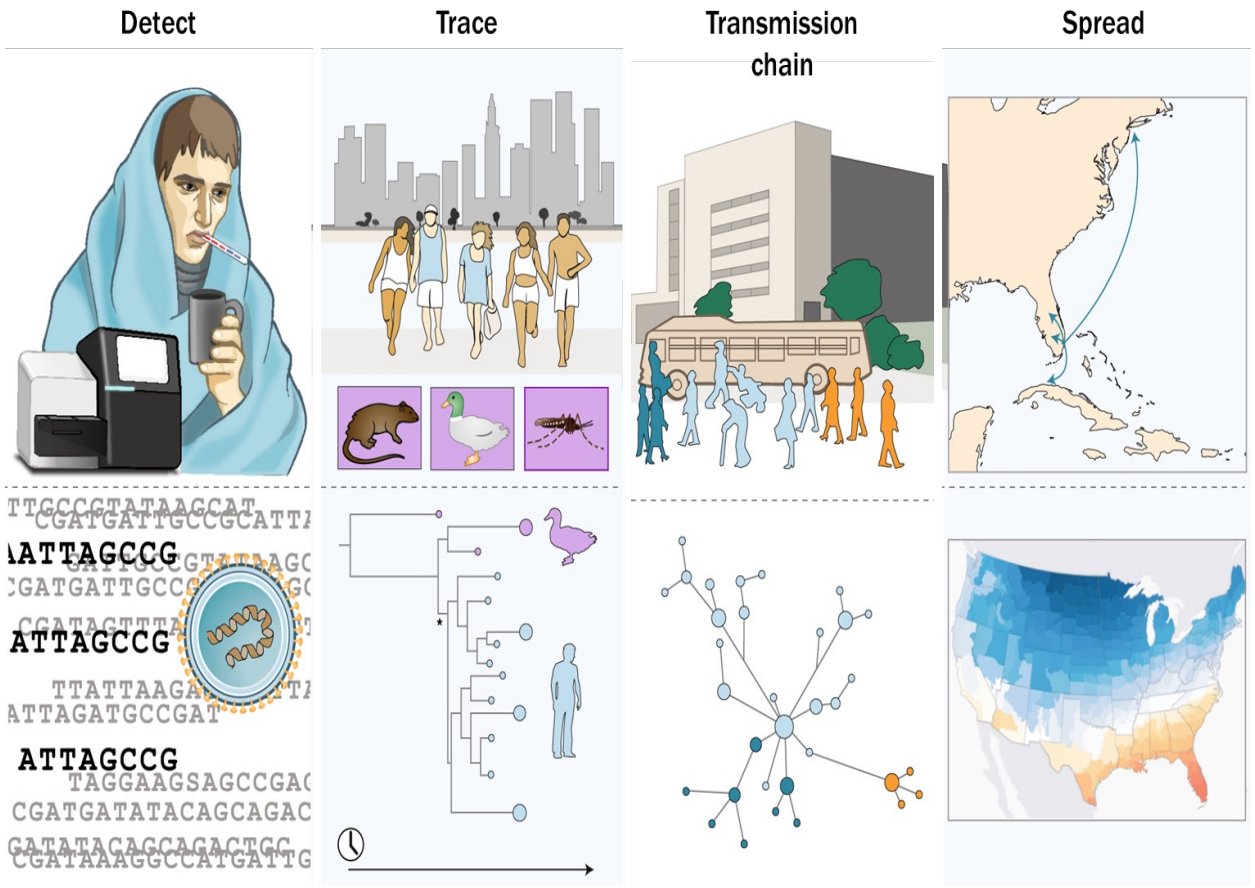


No Charges: Accidental transmission

- Strongly supported transmission cluster (aLRT > 0.70)
- Consistent topological relationship with ML and Bayesian methods
- Paraphyly of CC2 with respect to CC1 (*gag* & *pol*): Father-to-son
- Hypothesis testing of tree topology: Paraphyly (98%) > Monophyly (*gag*)

*Phylogenetic evidence = Circumstantial evidence

Applications:
Phylogeography of infectious diseases



Grubaugh *et al.*, 2019, *Nat Microbiol*

Tracking Virus Outbreaks

Critical questions:

- What: *virus, drivers…*
- Who: *spreaders, "at risk"*
- Where: *hotspots, localized*
- When: *recency, resurgence*
- Why: *vectors, climate change*
- How: *airborne, biological fluids*

Viral Genomic epidemiology

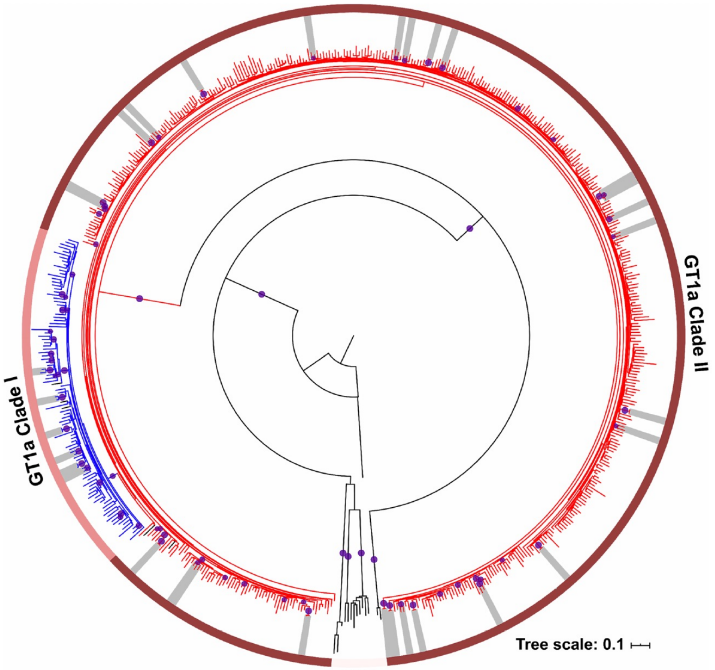
- *Virus evolutionary rates*
- *Sampling time*
- *Sampled cases*
- *Epidemiological data*

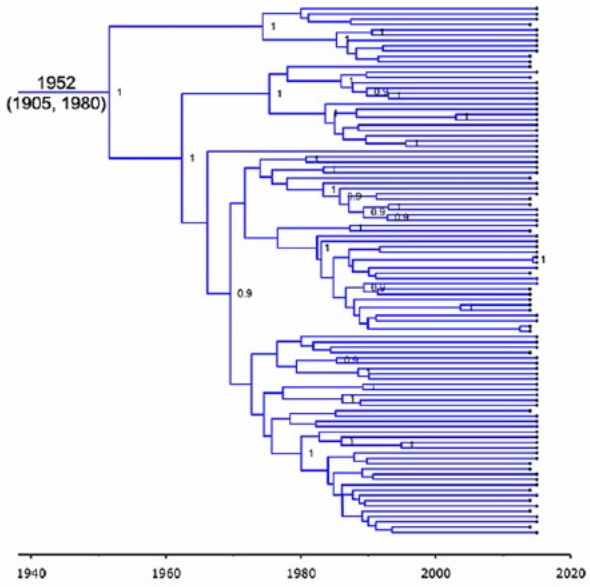
Phylogeography Practice Session: Case study

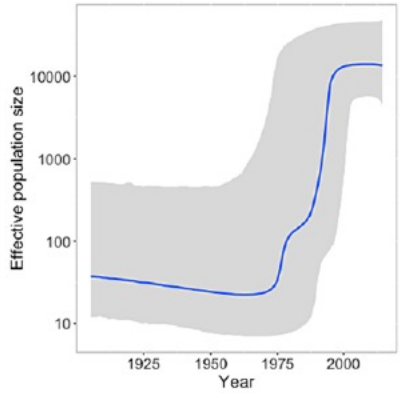
SCIENTIFIC REPORTS
nature research
Check for updates

Epidemic history and baseline resistance to NS5A-specific direct acting drugs of hepatitis C virus in Spain

Claudia Palladino[1,4], Ifeanyi Jude Ezeonwumelu[1,4], Irene Mate-Cano[2], Pedro Borrego[1], Paula Martínez-Román[2], Sonia Arca-Lafuente[2], Salvador Resino[2], Nuno Taveira[1,3] & Verónica Briz[2]







Transmission clusters → Time-scaled Phlogeny → Spatial distribution

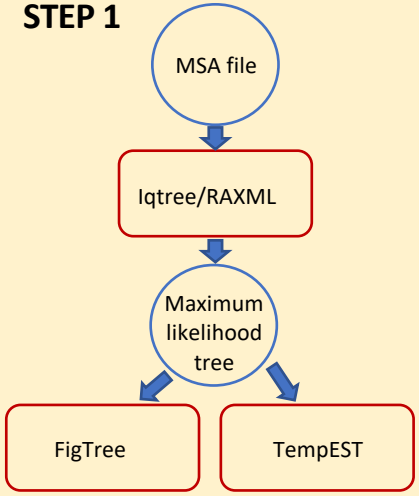
Phylogeography Practice Session: OUTLINE

TempEST v1.5.3; FigTree v1.4.3; IQ-Tree v2; MEGA X

Session I

Session II

STEP 1



IqTree:

- ML tree

FigTree:

- Tree visualization

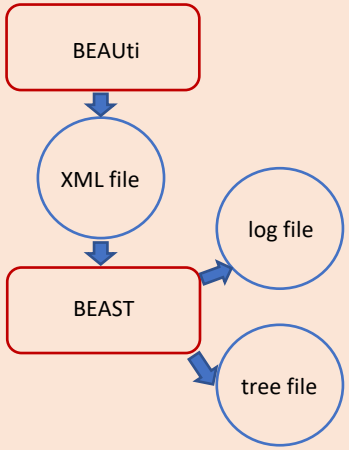
ClusterPicker:

- Transmission Clusters

TempEST:

- Remove outliers
- Determine Molecular clock likelihood

STEP 2



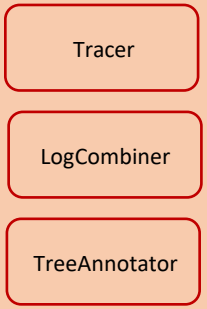
BEAUTi:

- Define data partitions, select models and define priors
- Decision on best Model/prior selection: specify MLE

BEAST:

- Bayesian Inference (MCMC)

STEP 3



Tracer:

- Check for convergence of MCMC, visualize log files, summarize data.

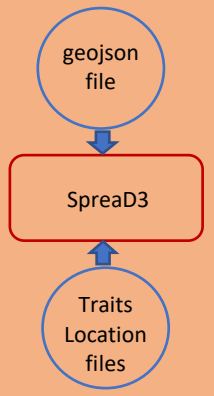
LogCombiner:

- Combine output files

TreeAnnotator:

- Summarize tree files to Maximum clade credibility trees

STEP 4



spreaD3:

- Spatiotemporal visualization
- BF for location transitions

SpreaD3 v0.9.6

BEAST v1.10 package