The Case That Went Viral: Using Phylogenetic Modeling to Identify HIV-Transfer in a Louisiana Criminal Trial

Varun Subramaniam | PUBH 6860 (Principles of Bioinformatics) | November 9th 2022

1 Introduction

In a startling case in 1996, Richard J. Schmidt, a gastroenterologist from Lafayette, Louisiana was accused of purposefully injecting his ex-girlfriend, Janice Trahan, with a vial of HIV-infected blood (Metzker *et al.*, 2002). Schmidt maintained his innocence, arguing that Trahan, who tested positive for HIV-1 immediately after the incident, had contracted the virus from the doctor's office in which she worked as a nurse. Conversely, Trahan alleged that Schmidt, who stored refrigerated vials of HIV-infected blood in his office from patients under his care, had both motive and means to execute this crime. Due to several examples of Schmidt's previous threats to harm Trahan and the unusual nature of this case, the Louisiana court, for the first time in United States' criminal trial history, turned to experts in phylogenetic analysis for help in determining the true source of Trahan's HIV-1 infection (Dye, 2002).

Researchers first collected several sample DNA sequences from the viral envelope region of Trahan's HIV-1 strain, to compare with potential source strains. However, since HIV mutates very rapidly, no two strains are genetically identical. Consequently, conventional DNA testing techniques, which seek to identify *identical* sequences, were unlikely to help determine the source of Trahan's infection. In order for Trahan's claims to be evidentially supported, prosecutors needed to prove that her HIV-1 strain was very closely related to a strain from the patient's vials in Schmidt's office (Science News Staff, 1997). They identified a single likely candidate from Schmidt's patient vials due to unconventional collection methods, suspicious documentation, and his failure to reveal the corresponding record to the court. Based on the prosecution's *a priori* hypothesis of suspected transmission from this candidate patient's vial to Trahan, researchers then collected several DNA sequences from the same envelope region of the patient's HIV sample. Lastly, they sequenced envelope regions from a random, representative sample of HIV-positive individuals in metropolitan Lafayette (Metzker *et al.*, 2002) as controls. These 132 total sequences from the envelope region were compiled into a single FASTA file ("Lab3sequences.fasta") and were each labeled as "P," "V," or "LA," denoting "Patient," "Victim," and "Louisiana" respectively.

In this paper, we use the sequence data in Lab3sequences.fasta to construct a phylogenetic tree representing the evolutionary history of the envelope region of Trahan's HIV-1 strain. We use the results from this tree, as well as those from concurrent studies, to test the prosecution's hypothesis of patient-to-victim transmission of HIV by doctor's injection, as well as alternative hypotheses of other transmission routes.

2 Materials and methods

2.1 Generating Multiple Sequence Alignment using MUSCLE

We used the MEGA11 software on Mac OS to generate a multiple sequence alignment (MSA) using the MUSCLE method. None of the default parameters for gap penalties,

memory/iterations, or advanced options (including cluster methods) could be changed in MEGA11 (Tamura *et al.*, 2021). For this MSA, we therefore retained the default penalties of 400.00 for "Gap Open" and 0.00 for "Gap Extend." The heavy penalty for new gaps, coupled with unpenalized extensions, yield alignments with fewer but longer gaps. Essentially, these parameters lend themselves to alignments with longer conserved regions and with gaps concentrated around sites. The parameters for this MUSCLE MSA are shown below:

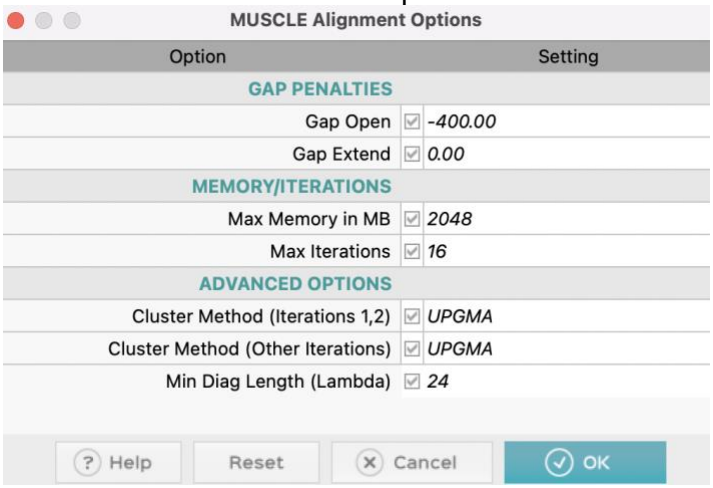

Fig. 1. Parameters for MUSCLE MSA in MEGA11

2.2 Generating Alternative MSA using MAFFT

We generated an alternative MSA for the sequences in Lab3sequences.fasta using the MAFFT method on the European Bioinformatics Institute's publicly available online tool (Madeira *et al.*, 2022). We once again retained all default parameters, as shown below. These parameters penalize gaps to a far lower extent than the MUSCLE MSA (gap penalties of 1.53 versus 400.00 respectively). However, unlike the MUSCLE MSA, the MAFFT alignment also penalizes *extensions* of gaps. Consequently, the latter lends itself to more-fragmented gaps, distributed across sites through the sequence, rather than concentrated regions of gaps or conserved nucleotides.

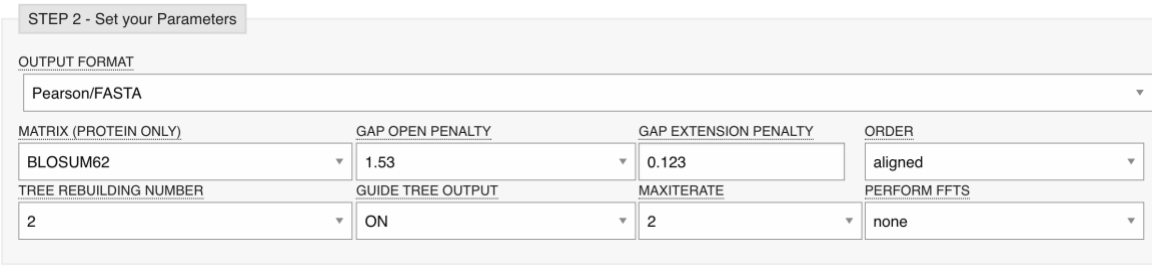

Fig. 2. Parameters for MAFFT MSA in MEGA11

2.3 Choosing Between MSAs

We used the "Compute Overall Mean Distance" function in MEGA11, once again retaining all default parameters, to yield a distance score for each alignment. The distance value produced from this method represents the average number of substitutions per site across all sequences. Higher distance values indicate that more substitutions were necessary, inferred to be

a poorer alignment (Tamura *et al.*, 2021). In this case, the distance values for both alignments were identical (*D = 0.08)*. Therefore, both alignments are equally strong by distance evaluation techniques.

We chose to proceed with the MAFFT MSA as it was a better indicator of overall homology between the sequences, rather than of conserved regions. Though comparably accurate and computationally effective, the MUSCLE algorithm prioritizes alignments with gapped and conserved *regions*. However, HIV mutates randomly across *several* sites which can result in more distributed gaps; therefore, the MAFFT-aligned sequences are more suitable for this study, which aims to analyze overall similarity with candidate sequences. MUSCLE alignments of Lab3sequences.fasta might be more applicable for studies on gene conservation between related strains of HIV-1.

2.4 Finding Best-Fit Model for MAFFT MSA

We used ModelTest 0.1.6, the latest available version of the software with an integrated OS-compatible graphical user interface, to find the best-fit model for the MAFFT MSA via a "Maximum Likelihood" tree. ModelTest was preferable to software such as MEGA11 for this step, as it tests alignments against several, more-complex models (Darriba *et al.*, 2020). The top ten models from ModelTest are shown below. Lower BIC scores (in the score column) indicate evolutionary models that better fit the input MSA (Stecher *et al.*, 2020).

In this case the best fit model is TVM + I + G4 or the transversion model, which assumes variable base frequencies and transversion mutation rates with constant transition mutation rates (Woodhams *et al.*, 2015). However, this model is not available in MEGA11 for subsequent tree construction. Therefore, the best-fit model for the purpose of this investigation, with the second-lowest BIC score, is GTR + I + G4: the "Generalized Time-Reversible" model, proposed by Tavaré et al. in 1986, which assumes variable nucleotide frequencies and substitution rates. This model also includes a proportion of invariable sites (I) and a gamma parameter (G4), referring to the rate of variation across sites (Arenas, 2015). The Generalized Time-Reversible model is more complex and accepts more parameters than other evolutionary models, such as the TPM3u model: a derivation of the three-parameter model, which assumes unequal base frequencies, equal base-substitution rates, and includes just five parameters (Posada, 2008).

	Model	K	lnL	score	delta	weight	cum weight
1	TVM+I+G4	9	-9318.8192	20489.1672	0.0000	0.6529	0.6529
2	GTR+I+G4	10	-9316.6577	20491.7017	2.5345	0.1839	0.8367
3	TPM3uf+G4	6	-9330.5417	20492.0397	2.8725	0.1553	0.9920
4	TPM3uf+I+G4	7	-9330.0782	20497.9701	8.8029	0.0080	1.0000
5	TIM3+G4	7	-9338.3763	20514.5663	25.3991	0.0000	1.0000
6	TIM1+G4	7	-9342.3670	20522.5477	33.3805	0.0000	1.0000
7	TPM1uf+I+G4	7	-9344.4980	20526.8097	37.6425	0.0000	1.0000
8	TPM1uf+G4	6	-9348.2024	20527.3611	38.1939	0.0000	1.0000
9	TIM1+I+G4	8	-9342.9641	20530.5994	41.4322	0.0000	1.0000
10	TIM3+I+G4	8	-9343.4655	20531.6022	42.4350	0.0000	1.0000

Fig. 3. ModelTest output for MAFFT MSA sorted by BIC Score

2.5 Creating a Maximum Likelihood Tree for the MAFFT-Aligned Sequences

We used the "Construct/Test Maximum Likelihood Tree" function on MEGA11 to generate a phylogenetic tree for the evolutionary trajectories of the sequences from the MAFFT MSA (Tamura *et al.*, 2021). We input "General Time Reversible model" into "Substitutions Type," as per the output from ModelTest described above. We also applied four gamma categories as the exact gamma value in the model was G4. The parameters for this maximum likelihood (ML) tree are shown below. We reiterate that the GTR + I + G4 model was only used for creation of an ML tree on MEGA11 due to limited model availability on the software.

We used a MacBook Air with an M2 processing chip for this investigation, which dramatically increased processing time for bootstrapping in MEGA11. For example, 1,000 level bootstrapping was estimated to take over 500 hours and was beyond the scope and resources of our investigation. Implementing this method in future replications of this study–specifically, by using parallel computing or other online tools–can improve confidence in the ML tree. The ML tree from MEGA11 was then annotated manually and inserted into the Interactive Tree of Life (iTOL) online tool for further optimized visualizations (Letunic and Bork, 2007).

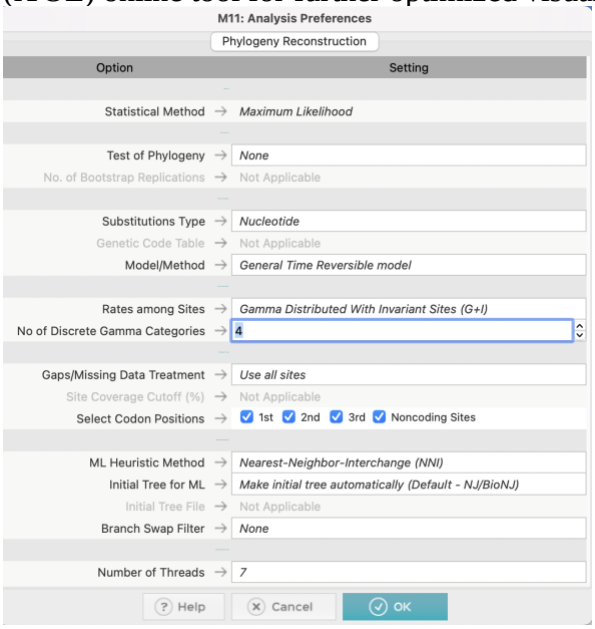

Fig. 4. Parameters for Maximum Likelihood Tree in MEGA11

3 Results

3.1 Visualizing Rectangular ML Tree for MAFFT-Aligned Data using MEGA11

Below is the ML Tree for the MAFFT MSA using MEGA11, described in Section 2.5. We identified three distinct clades for patient (P), victim (V), and control (LA) sequences. Each P sequence was more similar to all V sequences than to any LA sequences. Notably, within this envelope region of DNA, no V sequences were nested within the clade of P sequences.

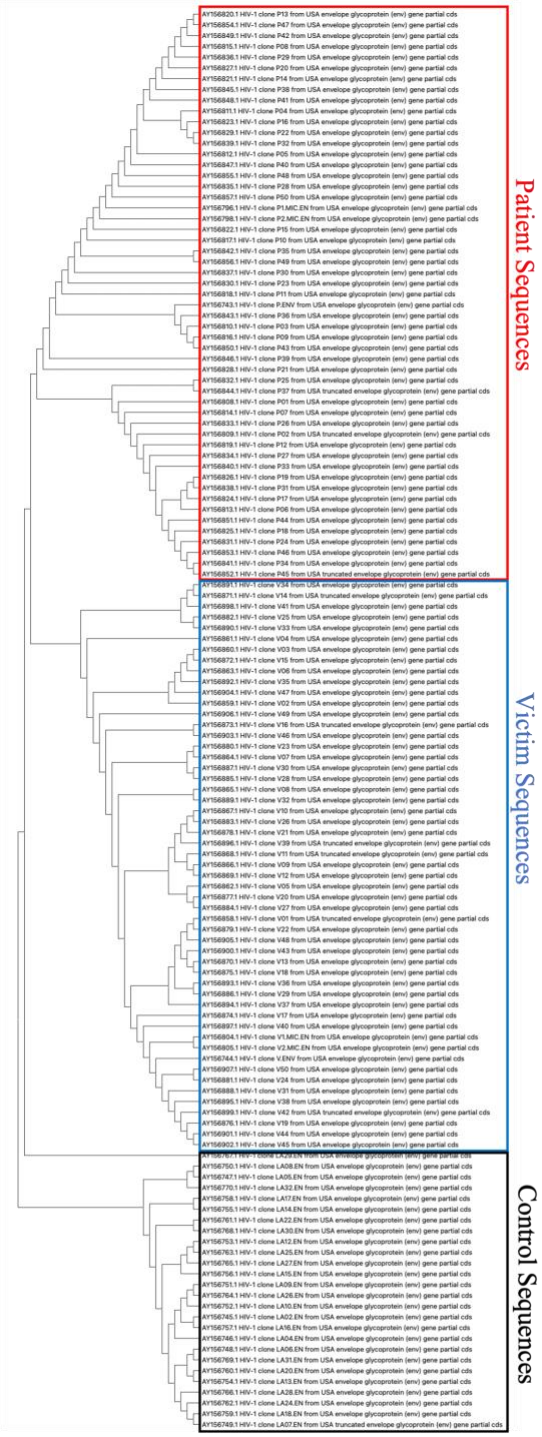

Fig. 5. Maximum Likelihood Tree in MEGA11 (Red Clade = P, Blue Clade = V, Black Clade = LA)

3.2 Visualizing Circular ML Tree for MAFFT-Aligned Data using iTOL

Below is the ML Tree for the MAFFT MSA using iTOL. Once again, three distinct clades were identified for P, V, and LA sequences. P and V sequences were most similar; however, each of these clades were separate with no sequences nested within a different clade. The arrow points to the node whose clade contains all P and V sequences, but which does not contain any LA sequences.

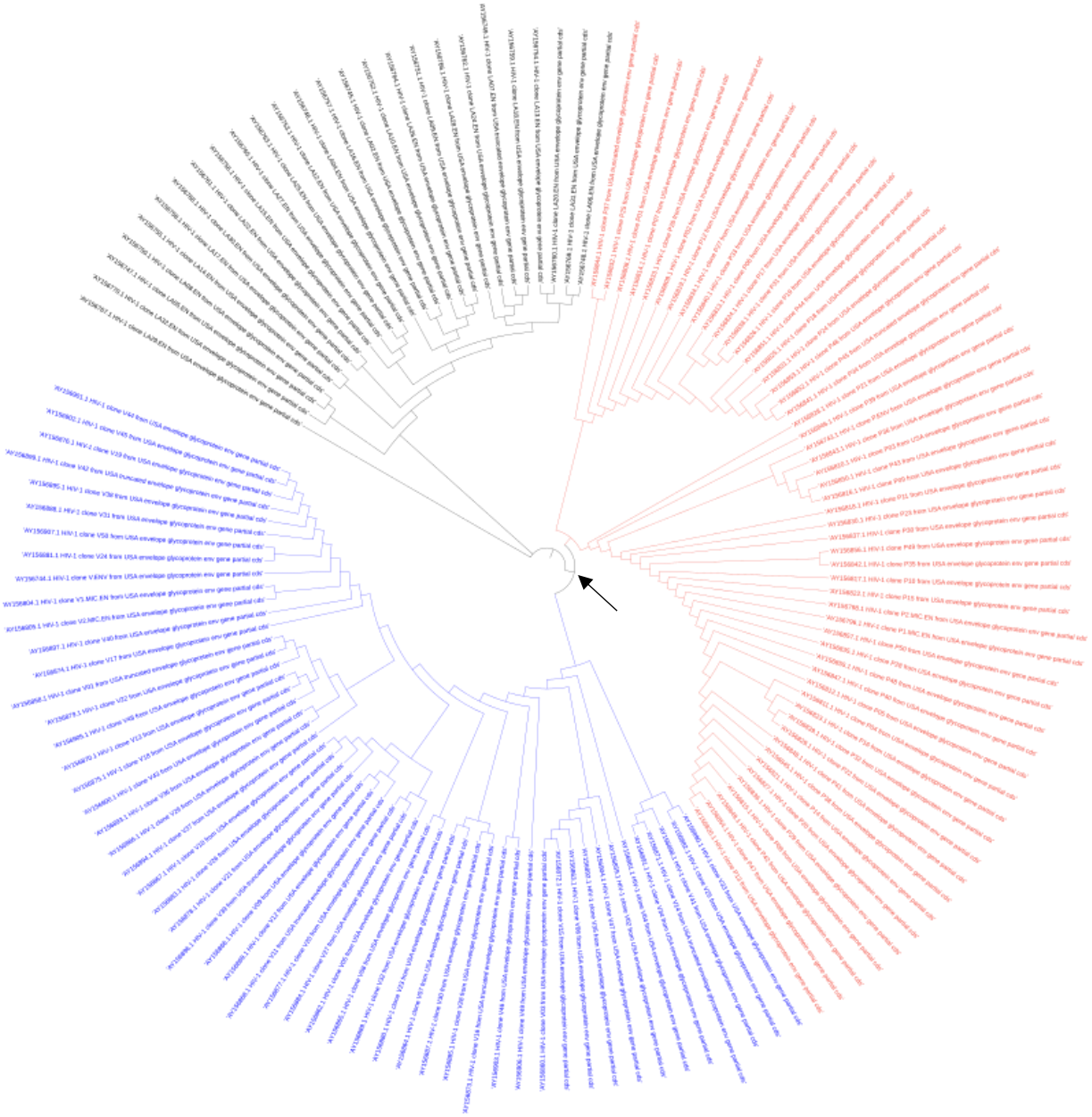

Fig. 6. Circular Maximum Likelihood Tree in iTOL (Red Clade = P, Blue Clade = V, Black Clade = LA, Arrow = Node containing P/V Clade)

4 Discussion

4.1 Interpreting ML Trees to Determine Similarity Between Sequences

The ML trees in Figures 5 and 6 both show three distinct clades for the 132 envelope sequences: patient, victim, and controls from the Lafayette metropolitan area. We noted that victim sequences were more similar to patient than to control sequences, due to the presence of a more-recent common ancestor. Victim and patient sequences belonged to the same clade, beginning at the node indicated by the arrow in Figure 6. This clade did not include LA sequences. This implies that, of all 132 envelope region samples in Lab3sequences.fasta, Trahan's HIV-strain is more likely to have derived from the patient's strain from Schmidt's office than from any of the representative control sequences from Lafayette. However, since none of the victim sequences were nested within a clade containing patient sequences, we cannot infer directionality of transfer or rule out the possibility of a common HIV-1 source for both Trahan and the patient. If V sequences were to have been nested within a P clade, we would have evidence to suggest that Trahan's strain had descended and evolved from the patient's strain of HIV-1. However, since the two groups' envelope region sequences only share a common ancestor, we may not go farther than to conclude that they are paraphyletic.

The outcomes of this phylogenetic approach do not provide conclusive evidence to corroborate Trahan's claims. Though paraphyletic and closely related, the victim and patient envelope sequences do not show evidence of directional transfer. We therefore cannot use these results to pronounce Schmdit guilty, as it is possible that both strains of HIV-1 arose from a separate source. Conversely, we cannot also ignore the relevance of these findings of high similarity; it remains possible that Schmidt injected Trahan with a vial containing the patient-strain, or even that he injected both Trahan and the patient with the same strain.

Other regions of the HIV-1 sequences better inform the specific relationship between Trahan's strain of HIV-1 and that of the patient. Metzker et al. also explored evolutionary patterns of the reverse transcriptase (RT) region, rather than just the envelope region, for the same cohort in this study. They found from their ML tree that several of Trahan's RT strains were nested *within* patient RT strains, providing strong evidence for the transfer of HIV-1 from the patient's sample in Schmidt's office to the victim. In these nested victim sequences, they also found regions of azidothymidine (AZT) resistance highly similar to those of ancestral patient sequences. These specific AZT-resistance regions were not found in any of the control sequences (Metzker *et al.*, 2002). This further corroborates the theory of directional transfer from the patient sample in Schmidt's office to Trahan. These findings can be used to eliminate certain possibilities that arose from our investigation alone–it is now clear that Trahan inherited HIV-1 from the patient's strain.

4.2 Drawing Conclusions on the Utility of These Insights for the Trial

It is critical to note that even jointly considering the findings of our study and those of Metzker et al. is insufficient to conclusively pronounce Schmidt guilty or innocent of injecting Trahan with a vial of HIV-infected blood. Though there is strong evidence of Trahan's strain descending and evolving from the patient's strain of HIV-1, prosecutors must eliminate further possible routes of infection. For example, they must establish that the victim and patient did not have any prior direct contact that could have led to viral transmission without Schmidt's injection, however improbable this outcome might appear. We remind readers that, while genetic testing and phylogenetic modeling are efficient, powerful tools for criminal trials, they must be used to

supplement cases: they cannot be taken as the case itself. Nevertheless, we conclude that there is sufficient evidence of HIV-transfer from the patient's blood to Trahan for the prosecution to continue building its case and to proceed with eliminating other potential routes of infection: especially considering the breadth of evidence confirming Schmidt's aggressive and threatening behavior towards the victim.

References

Arenas,M. (2015) Trends in substitution models of molecular evolution. *Front. Genet.*, **6**.

Darriba,D. *et al.* (2020) ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models. *Mol. Biol. Evol.*, **37**, 291–294.

Dye,L. (2002) Scientists Use Virus to Trace Assault Suspect. *ABC News*.

Letunic,I. and Bork,P. (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinforma. Oxf. Engl.*, **23**, 127–128.

Madeira,F. *et al.* (2022) Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Res.*, gkac240.

Metzker,M.L. *et al.* (2002) Molecular evidence of HIV-1 transmission in a criminal case. *Proc. Natl. Acad. Sci. U. S. A.*, **99**, 14292–14297.

Posada,D. (2008) jModelTest: Phylogenetic Model Averaging. *Mol. Biol. Evol.*, **25**, 1253–1256.

Science News Staff (1997) Novel DNA Evidence May Get Its Day in Court.

Stecher,G. *et al.* (2020) Molecular Evolutionary Genetics Analysis (MEGA) for macOS. *Mol. Biol. Evol.*, **37**, 1237–1239.

Tamura,K. *et al.* (2021) MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Mol. Biol. Evol.*, **38**, 3022–3027.

Woodhams,M.D. *et al.* (2015) A New Hierarchy of Phylogenetic Models Consistent with Heterogeneous Substitution Rates. *Syst. Biol.*, **64**, 638–650.