Posterior summary of trees

Instructor: Julia A. Palacios

In Bayesian phylogenetics, many possible genealogies can explain the data

Human Influenza A H3N2

Sampled Genealogies from the posterior distribution obtained from BEAST (Drummond et al. 2012).

Sequence data obtained from Gisaid Epi Flu database.

410 sequences from Human Influenza A H3N2 virus at the HA segment from each region: Singapore, New York and Chile. Samples were selected at random from (2014,2019).

Sequences were aligned using MAFFT.

Mutation model assumed: SRD06 codon-partition substitution model.

In Bayesian phylogenetics, many possible genealogies can explain the data

Human Influenza A H3N2

Questions:

How different are the evolutionary processes of influenza across different regions? across temporal seasons?

What is a typical evolutionary history of influenza?

Posterior summary of trees

- Tree space is a discrete-continuous high dimensional space.
- Multiple ways to define the center of a tree (Billera et al., 2001).
- **Densitree** provides a visualization of a sample of trees from the posterior distribution.
- **TreeAnnotator** uses the Maximum Clade Credibility (MCC) heuristic to summarize the posterior by a single tree.

Maximum Clade Credibility (MCC)

- It picks a topology: it takes the posterior tree with the maximum product of posterior probabilities of its internal nodes.
- It then assigns heights for each clade based on a point estimate from posterior trees containing that clade.
- It can lead to negative branch lengths

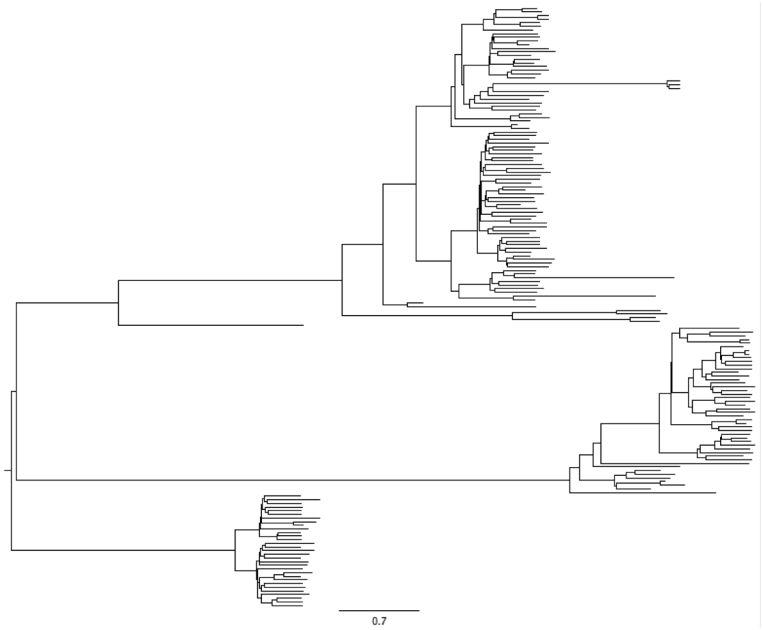
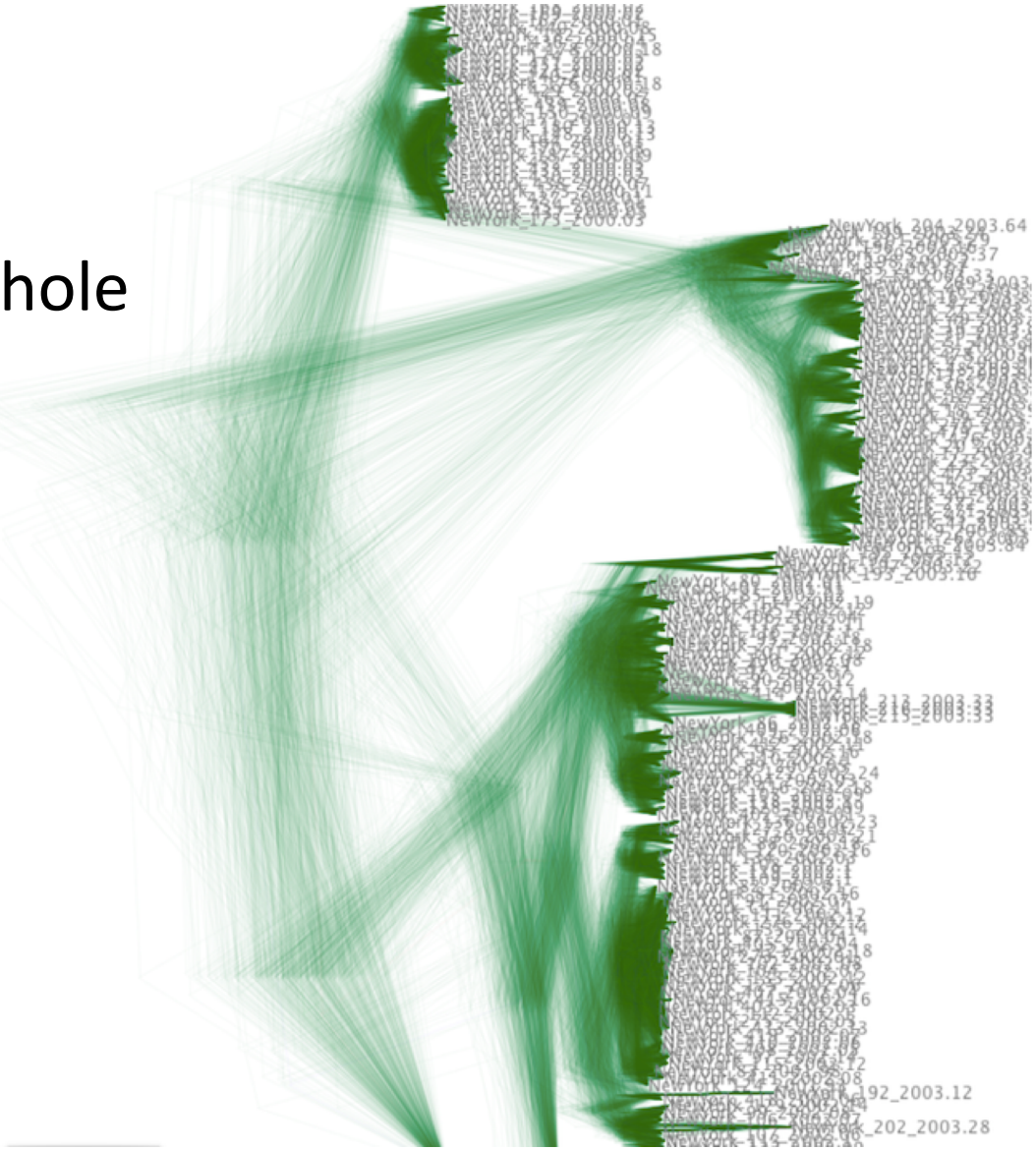


Figure: MCCT from Ne inference tutorial H3N2 Human influenza in New York

Densitree

- You can visualize the whole posterior distribution.



Tree distances

Phylogenetic trees

- Robinson-Foulds, Nearest Neighbor Interchange, Subree-Prune-and-Regraft, Tree Bisection and Reconnection.
- Geodesic distance between phylogenetic trees (Billera, Holmes and Vogtmann, 2001; Owen and Provan, 2009)
- Kendall-Colijn (2018)

Tree Shapes:

- Colijn-Plazzotta (2017)

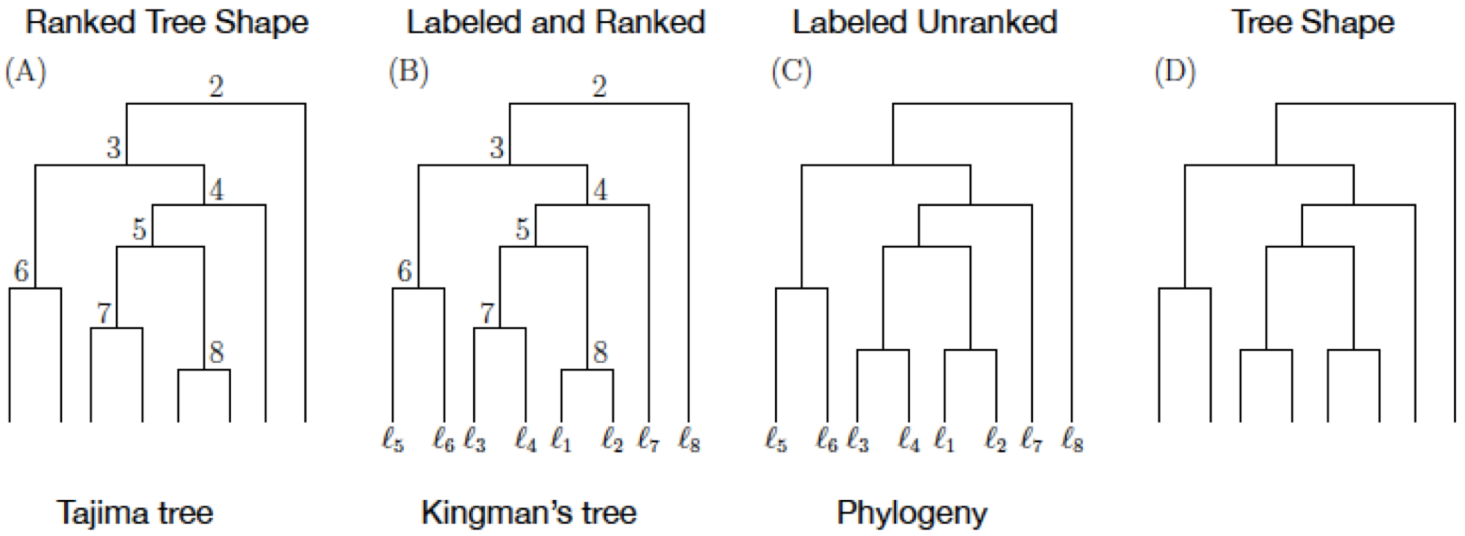
Why a metric on the space of unlabeled trees?

Motivation for defining a metric:

- Construct a decision theoretic statistical inference.
- Summarize unlabeled tree distributions.
- Compare different prior distributions on unlabeled trees.
- Compare different empirical distributions on unlabeled trees and model comparison.
- Develop approximate inference methods.

Unlabeled ranked tree shapes

Tree topologies:



Tajima tree

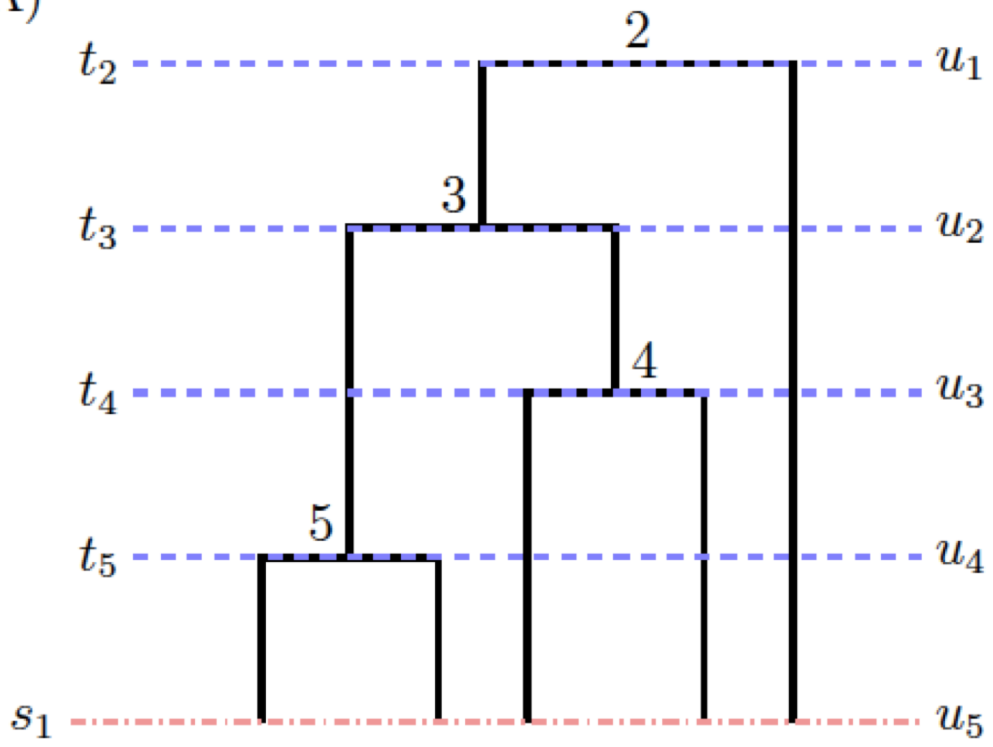
Kingman's tree

Phylogeny

Ranked tree shapes as matrices

There is a unique encoding of a ranked tree shapes as lower triangular integer-valued matrix.

(A)



(B)

$$\mathbf{F} = \begin{pmatrix} 2 & 0 & 0 & 0 \\ 1 & 3 & 0 & 0 \\ 1 & 2 & 4 & 0 \\ 1 & 1 & 3 & 5 \end{pmatrix}$$

$F_{i,j}$ indicates the number of branches extant at time (u_{j+1}, u_j) that do not bifurcate during (u_{i+1}, u_j)

Distances between ranked tree shapes

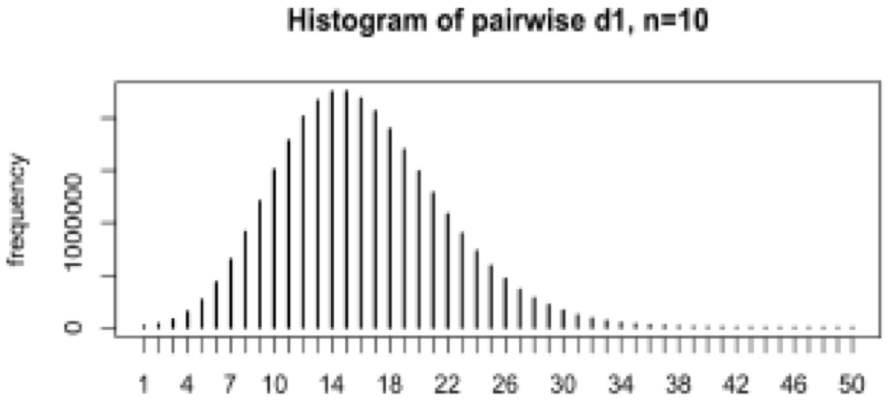
We define two distance functions d_1 and d_2 on the space of ranked tree shapes with n leaves. For $T_1^R, T_2^R \in \mathcal{T}_n^R$ and their corresponding $\mathbf{F}$-matrix representations $\mathbf{F}^{(1)}, \mathbf{F}^{(2)} \in \mathcal{F}_n$,

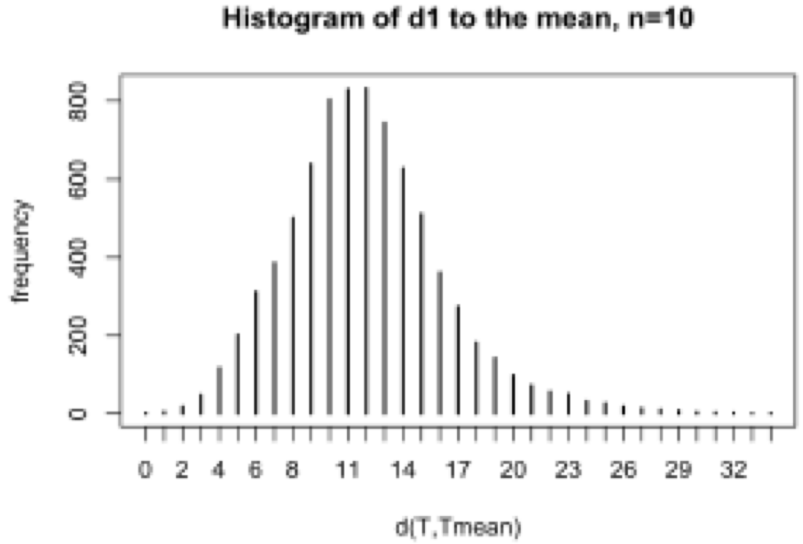
$$d_1(T_1^R, T_2^R) = \sum_{i,j} \left| F_{i,j}^{(1)} - F_{i,j}^{(2)} \right|,$$

$$d_2(T_1^R, T_2^R) = \sqrt{\sum_{i,j} \left(F_{i,j}^{(1)} - F_{i,j}^{(2)} \right)^2}.$$

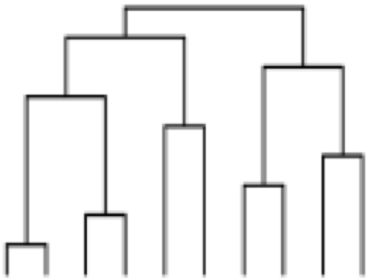
The two distances are metrics since both inherit properties of L_1 norm (Manhattan distance) and L_2 norm (Frobenius norm).

Example when n=10

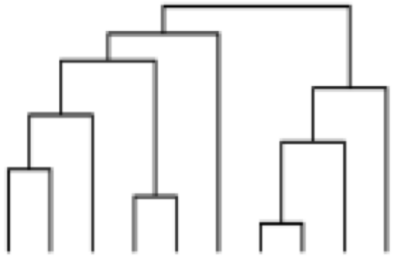




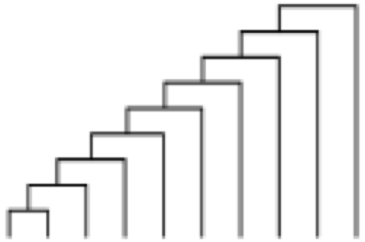
Most balanced



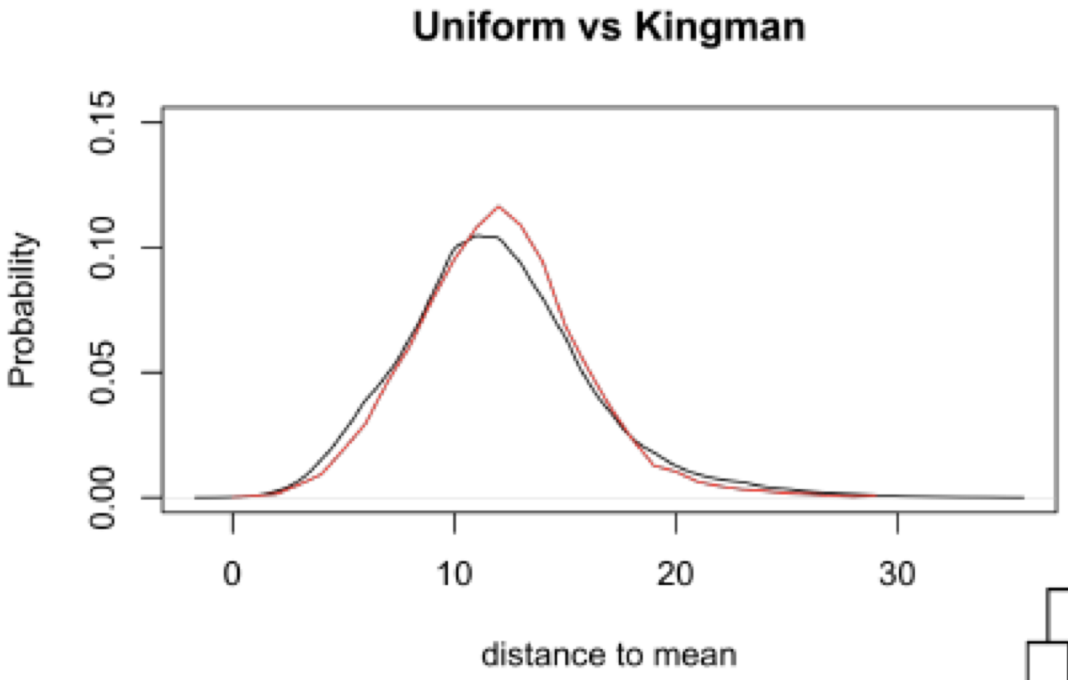
Mean

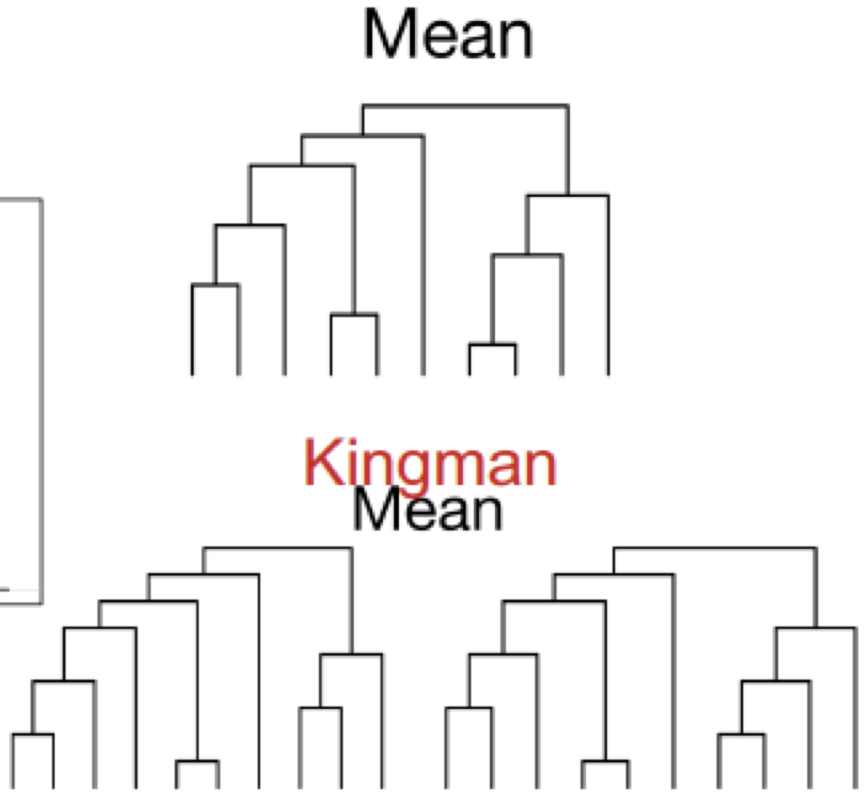


Most unbalanced



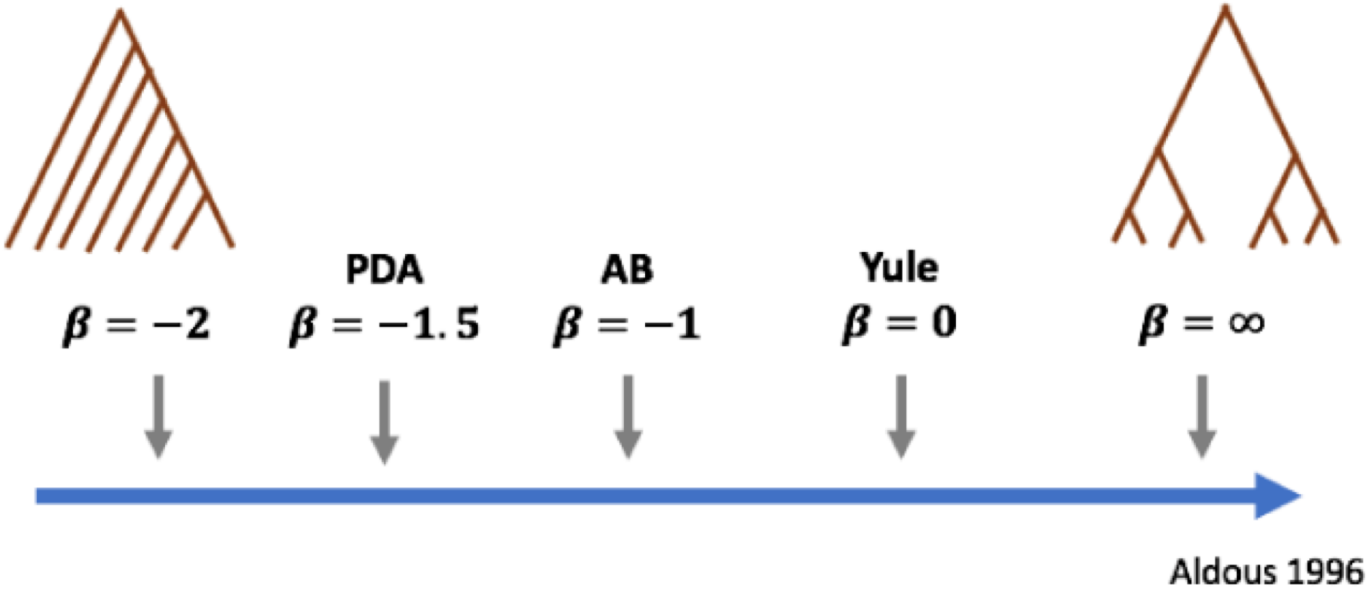
Example when n=10





Testing distance on ranked tree shapes...

Beta-splitting model on labeled tree shapes



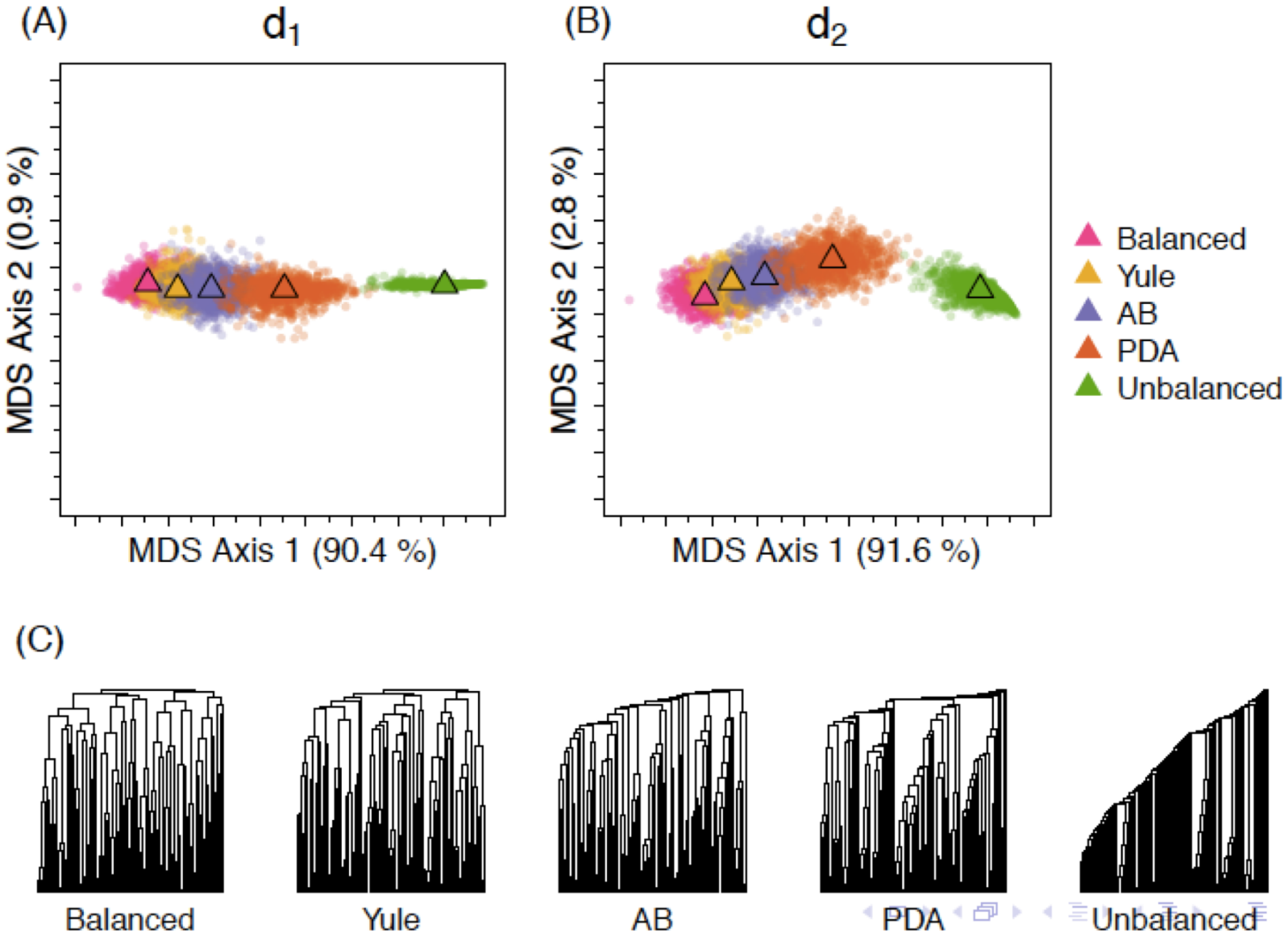
Can be adapted to unlabeled ranked tree shapes.

Testing distance on ranked tree shapes...

We simulated 1000 ranked tree shapes of $n = 100$ tips from the Beta-splitting:

- Balanced distribution $\beta = 100$
- Yule model $\beta = 0$
- AB model $\beta = -1$
- PDA $\beta = -1.5$
- Unbalanced $\beta = -1.9$

Simulations from Beta-splitting on ranked tree shapes



Adapting other tree distances on ranked tree shapes

- Colijn-Plazzotta **(CP)** metric on tree shapes

$$d_{\text{CP-RTS}}(T_1^R, T_2^R) = d_{\text{CP}}(\phi(T_1^R), \phi(T_2^R)),$$

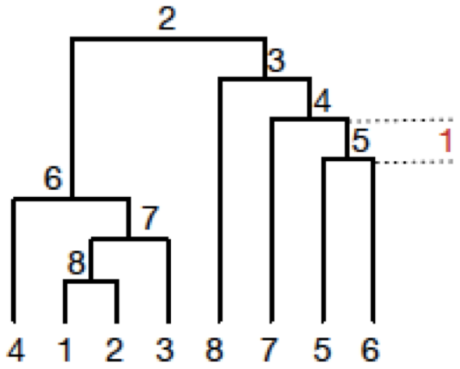
where ϕ returns the corresponding tree shape of a ranked tree shape by removing the labels of its internal nodes (rankings).

Adapting other tree distances on ranked tree shapes..

- Billera-Holmes-Vogtmann (**BHV**) on timed labeled phylogenies (feasible option)

$$d_{\text{BHV-RTS}}(T_1^R, T_2^R) = d_{\text{BHV}}(\psi(T_1^R), \psi(T_2^R)),$$

where ψ maps a ranked tree shape to its corresponding ranked labeled genealogy by assigning a uniquely defined label to each leaf and assigning a unit length to each change point time interval (u_i, u_{i-1}).



Adapting other tree distances on ranked tree shapes..

- Billera-Holmes-Vogtmann (**BHV**) on timed labeled phylogenies (unfeasible option for large n)

$$d_{\text{BHV-RTS}^*}(T_1^R, T_2^R) = \min_{\pi_i, \pi_j \in S_n} \left\{ d_{\text{BHV}}\left((\pi_i \circ \psi)(T_1^R), (\pi_j \circ \psi)(T_2^R) \right) \right\}.$$

Here, ψ assigns an initial labeling to a ranked tree shape and assigns unit branch lengths. π_i permutes the set of leaf labels.

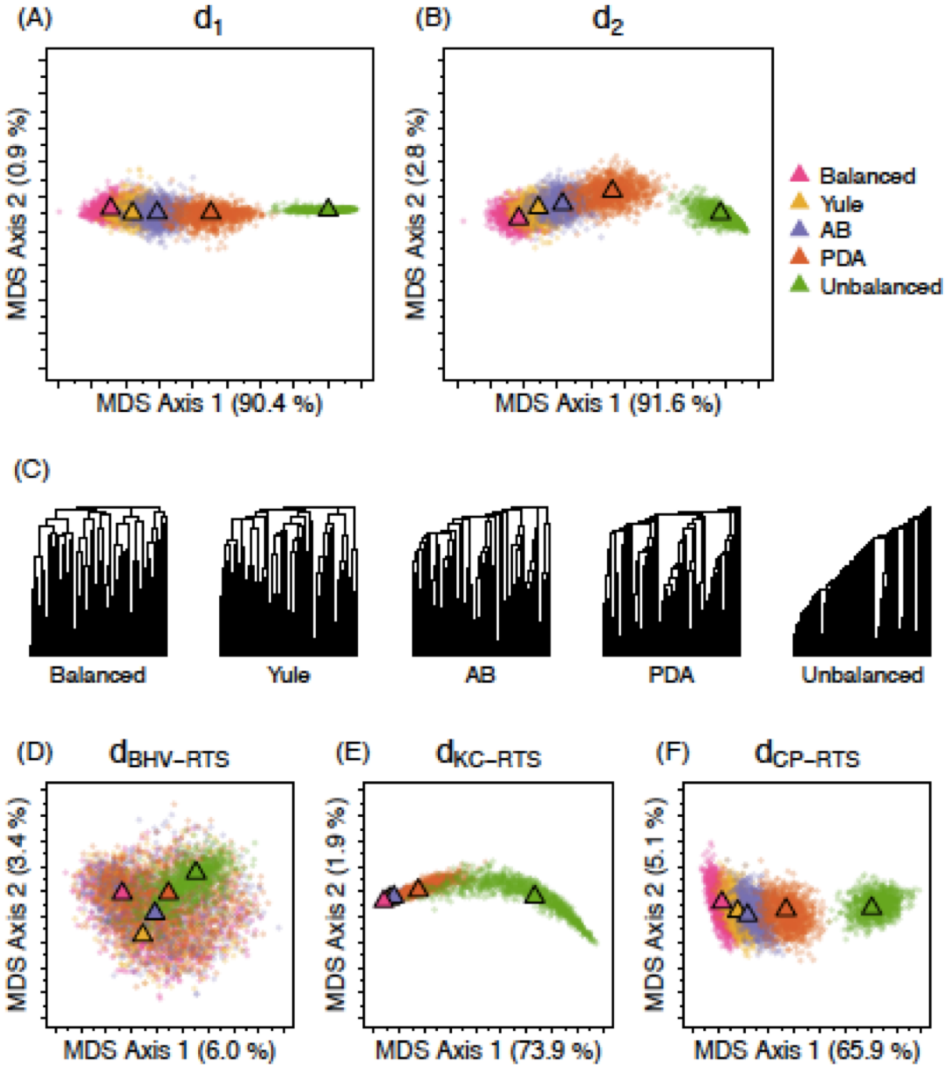
Adapting other tree distances on ranked tree shapes..

- The Kendall Colijn (KC) metric on labeled phylogenies and timed labeled phylogenies (regulated by a $\lambda = 0$ weight)

$$d_{\text{KC-RTS}}(T_1^R, T_2^R) = d_{\text{KC},0}(\eta(T_1^R), \eta(T_2^R)),$$

where η maps a ranked tree shape to a labeled unranked tree shape by removing internal node rankings and uniquely labeling leaves following the procedure described for $d_{\text{BHV-RTS}}$.

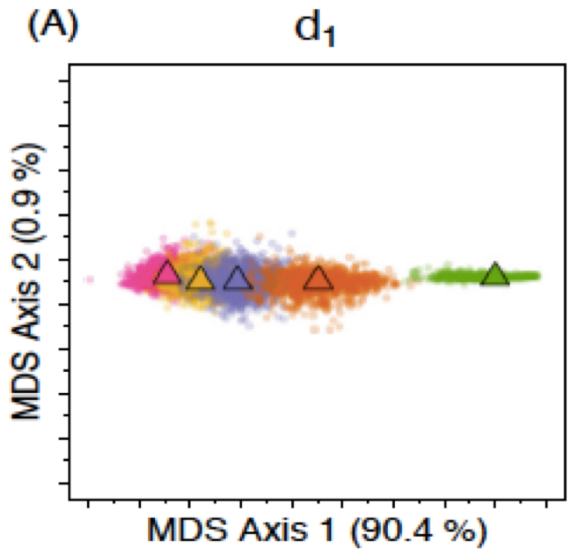
Comparing separation of different Beta-splitting distributed ranked tree shapes with other metrics



We calculate confusion tables as:

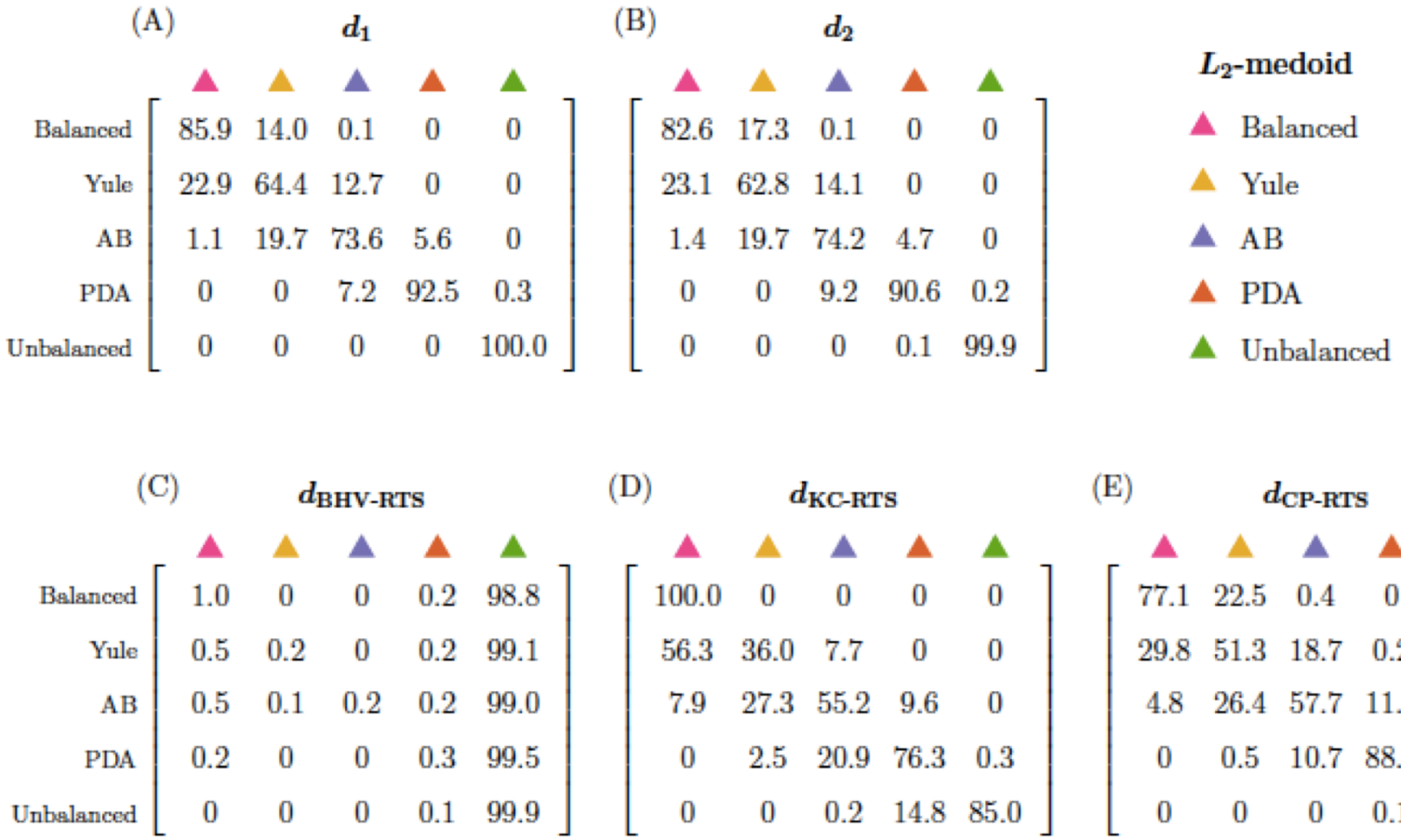
The percentage of trees that are closer to the observed medoids of other distributions.

	▲	▲	▲	▲	▲
Balanced	85.9	14.0	0.1	0	0

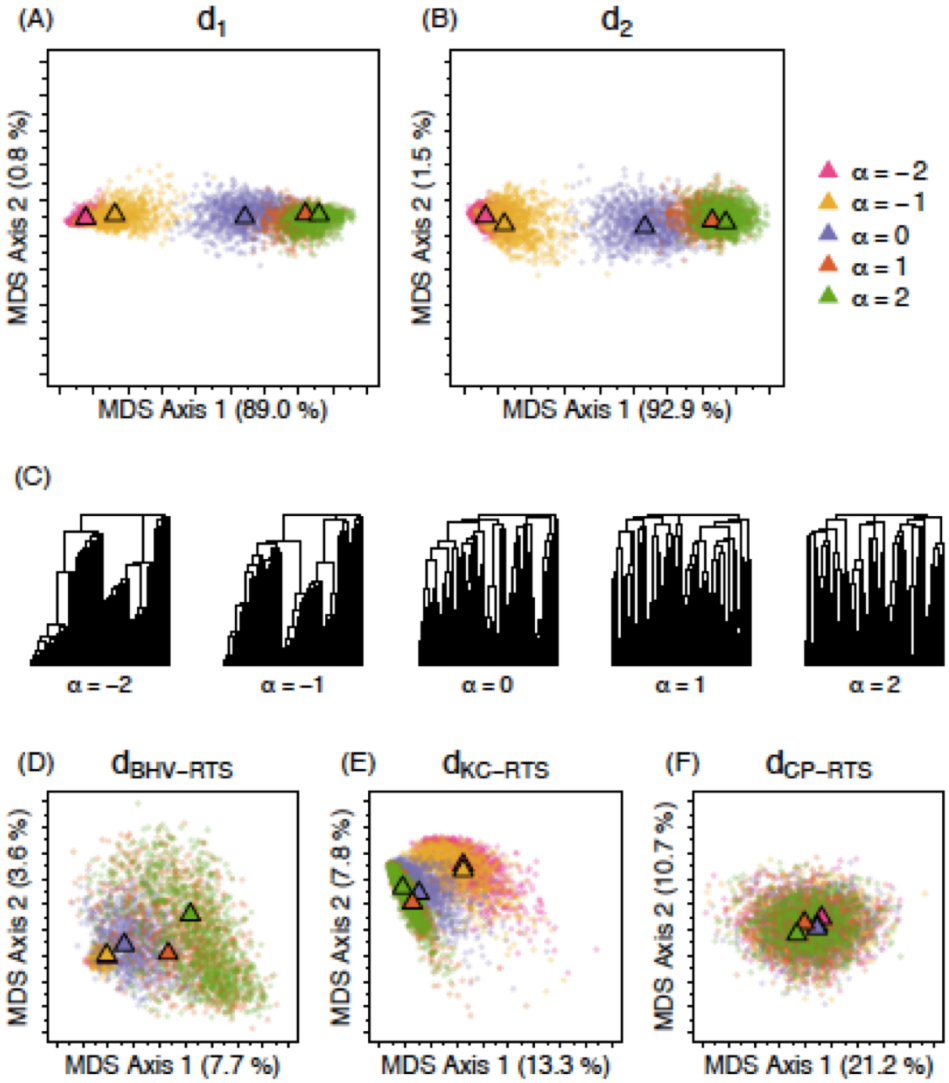


Comparing separation of different Beta-splitting distributed ranked tree shapes with other metrics

Confusion matrices: Diagonal shows the percentage of ranked tree shapes that are closer to their medoids.



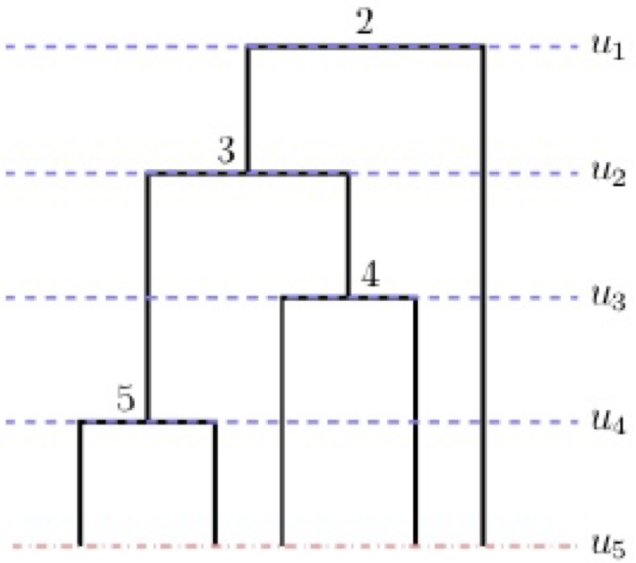
Simulation from Alpha-Beta splitting on ranked tree shapes



A distance between ranked genealogies

$$d_1^w(\mathbf{g}_1^R, \mathbf{g}_2^R) = \sum_{i,j} \left| F_{i,j}^{(1)} W_{i,j}^{(1)} - F_{i,j}^{(2)} W_{i,j}^{(2)} \right|,$$

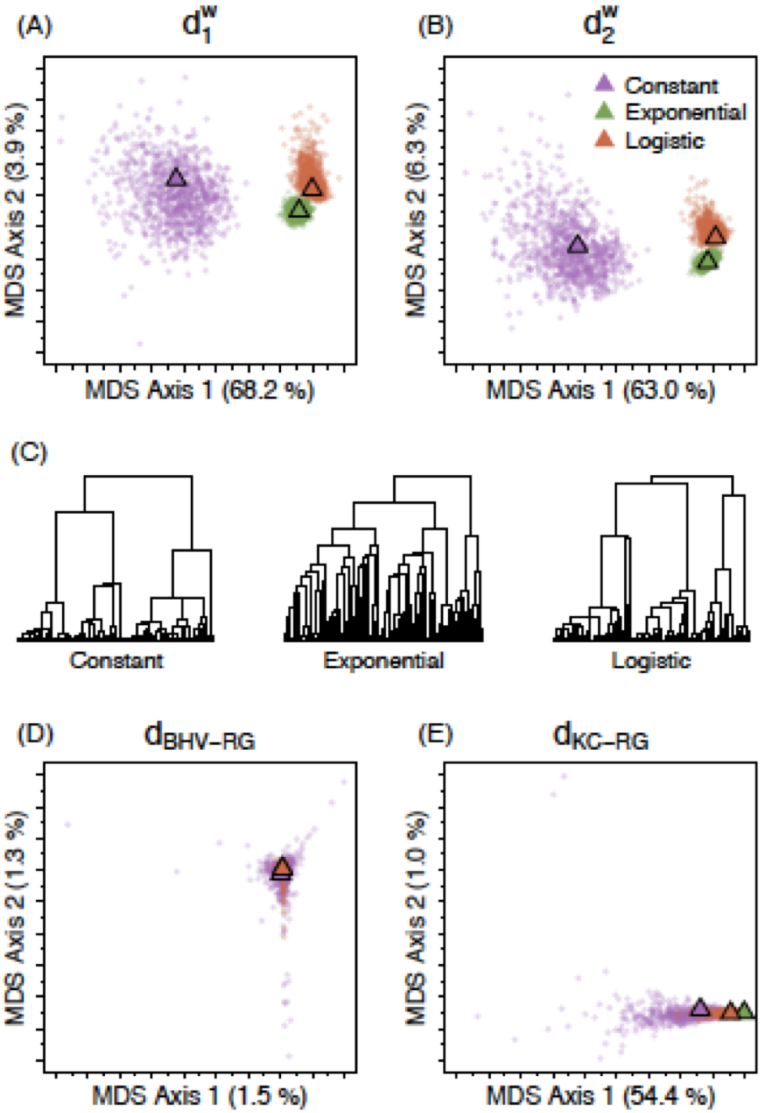
$$d_2^w(\mathbf{g}_1^R, \mathbf{g}_2^R) = \sqrt{\sum_{i,j} \left(F_{i,j}^{(1)} W_{i,j}^{(1)} - F_{i,j}^{(2)} W_{i,j}^{(2)} \right)^2},$$



$$\mathbf{F} = \begin{pmatrix} 2 & 0 & 0 & 0 \\ 1 & 3 & 0 & 0 \\ 1 & 2 & 4 & 0 \\ 1 & 1 & 3 & 5 \end{pmatrix}$$

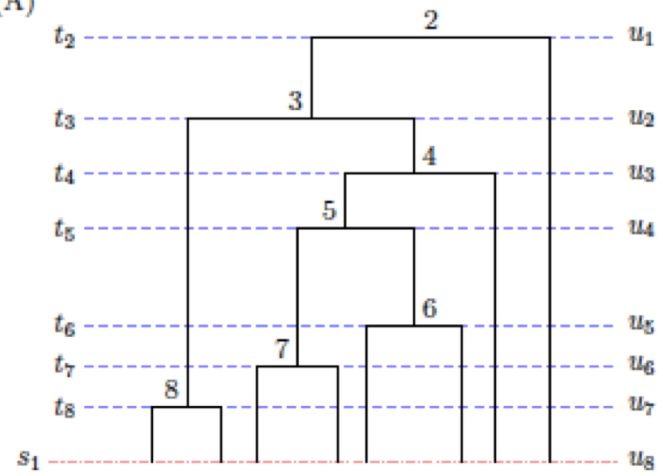
$$\mathbf{W} = \begin{pmatrix} u_1 - u_2 & 0 & 0 & 0 \\ u_1 - u_3 & u_2 - u_3 & 0 & 0 \\ u_1 - u_4 & u_2 - u_4 & u_3 - u_4 & 0 \\ u_1 & u_2 & u_3 & u_4 \end{pmatrix}$$

Simulation of genealogies with different branch distributions



Adapting the distances to heterochronous ranked tree shapes

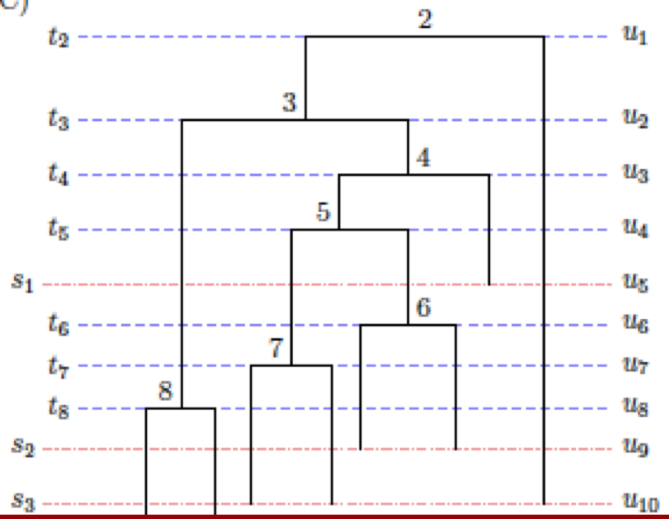
(A)



(B)

$$\mathbf{F} = \begin{pmatrix} 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 3 & 0 & 0 & 0 & 0 & 0 \\ 1 & 2 & 4 & 0 & 0 & 0 & 0 \\ 1 & 2 & 3 & 5 & 0 & 0 & 0 \\ 1 & 2 & 3 & 4 & 6 & 0 & 0 \\ 1 & 2 & 3 & 3 & 5 & 7 & 0 \\ 1 & 1 & 2 & 2 & 4 & 6 & 8 \end{pmatrix}$$

(C)



(D)

$$\mathbf{F} = \left(\begin{array}{cccc|cccc|cc} 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 2 & 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 2 & 3 & 5 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 1 & 2 & 2 & 4 & 4 & 0 & 0 & 0 & 0 & 0 \\ 1 & 2 & 2 & 3 & 3 & 5 & 0 & 0 & 0 & 0 \\ 1 & 2 & 2 & 2 & 2 & 4 & 6 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 3 & 5 & 7 & 0 & 0 \\ \hline 1 & 1 & 1 & 1 & 1 & 1 & 3 & 5 & 5 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 2 & 2 \end{array}\right)$$

Human influenza A/H3N2

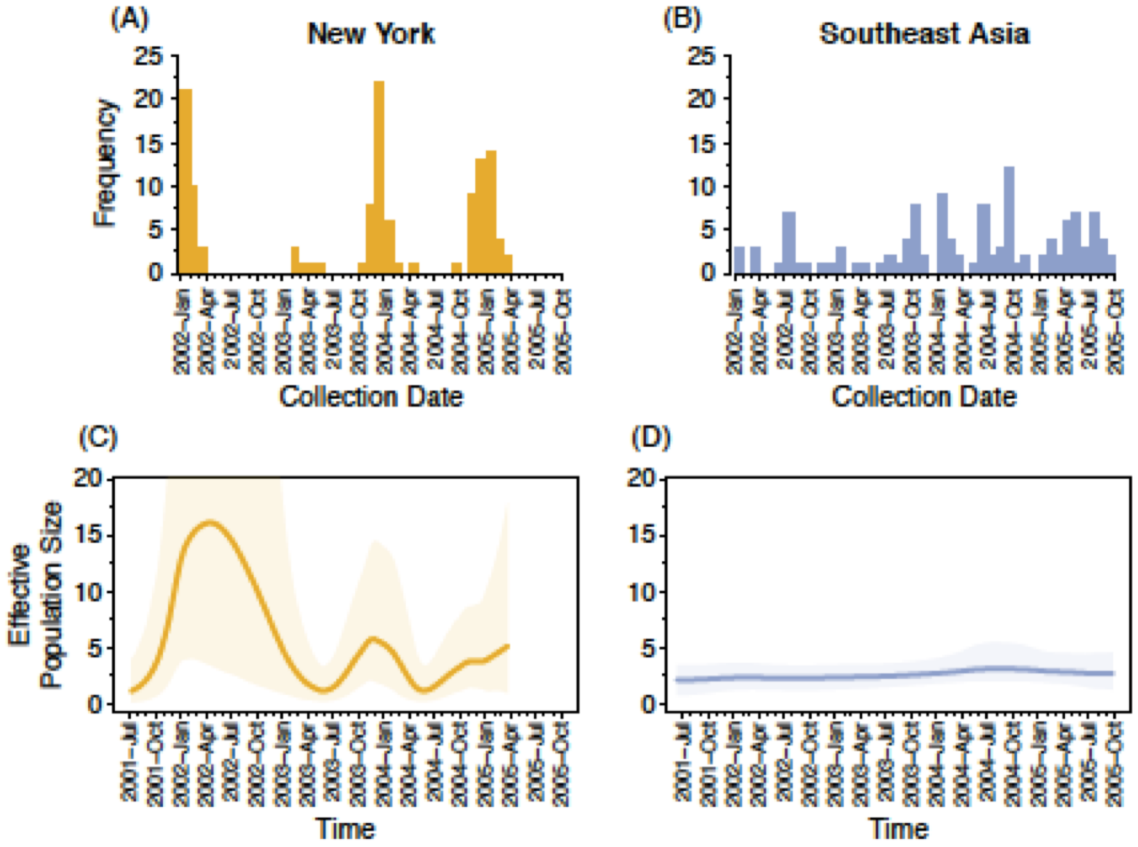
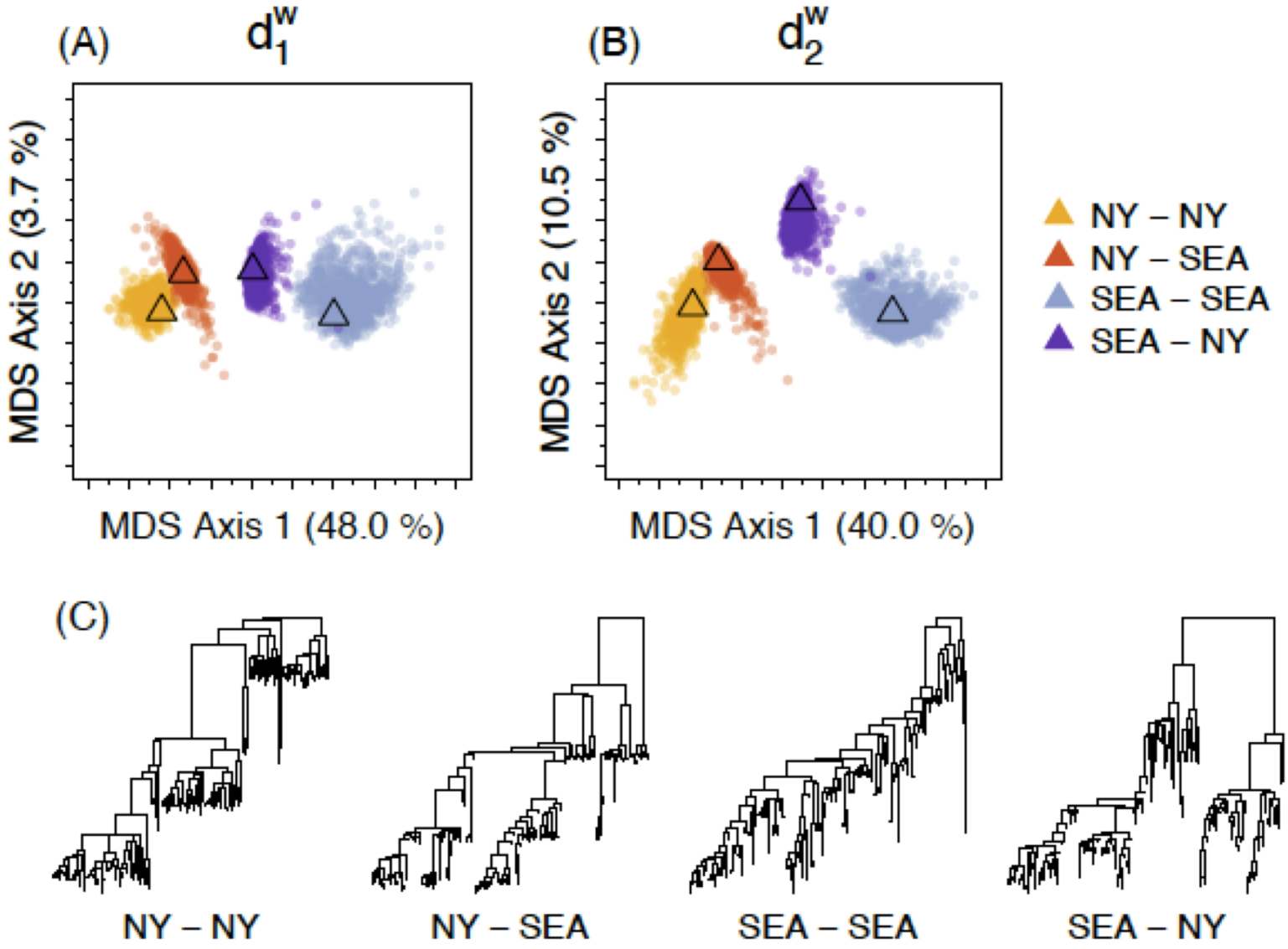
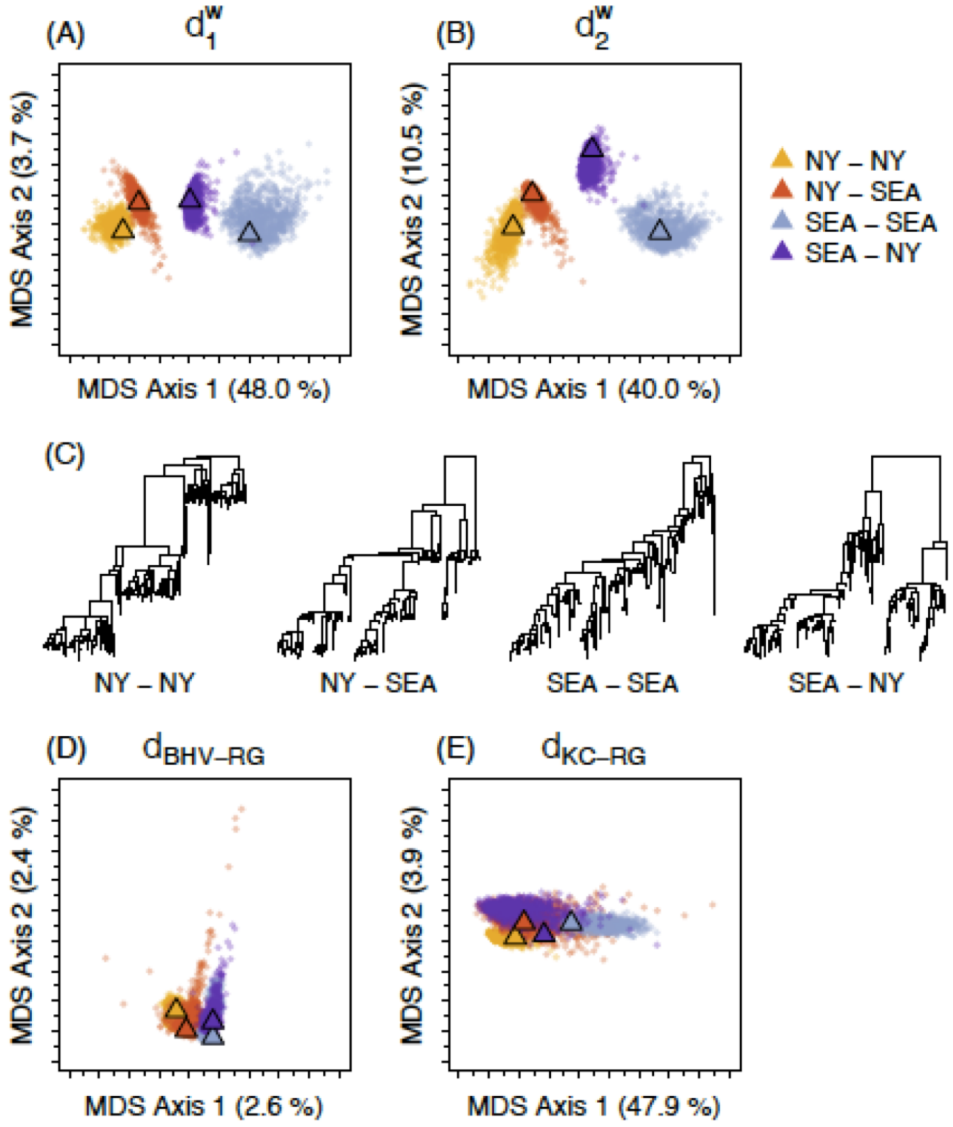


Fig. 7. Human influenza A/H3N2 virus collection dates and inferred effective population size trajectories. (A) Collection date histogram, New York; (B) Collection date histogram, Southeast Asia; (C) Inferred effective population size trajectory with BEAST, New York; and (D) Southeast Asia.

Human influenza A/H3N2



Human influenza A/H3N2



Convergence diagnostic

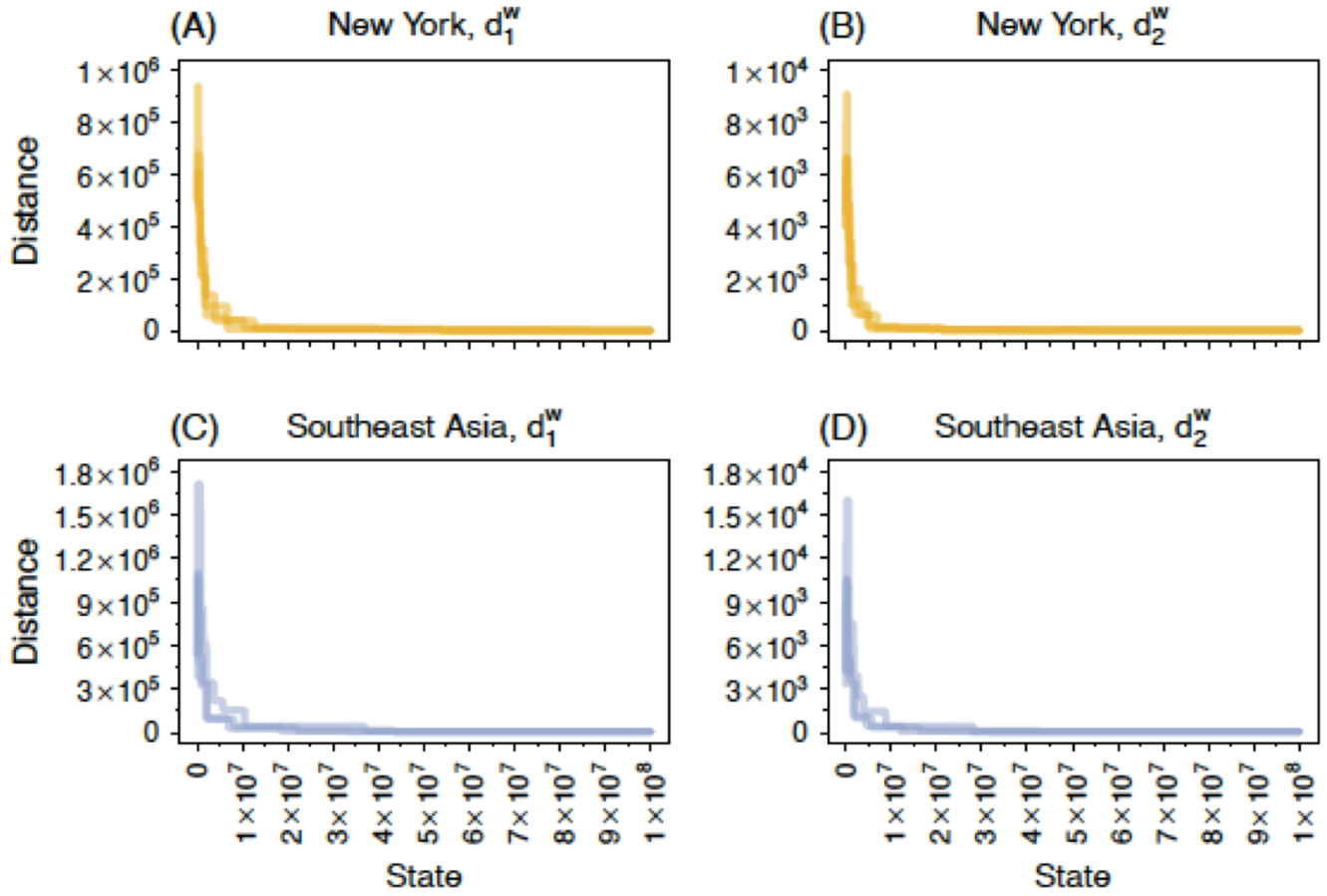


Fig. 9. Assessment of convergence of MCMC BEAST chains. Each curve corresponds to the distance between the running posterior L_2-medoid ranked genealogy and the global posterior L_2-medoid ranked genealogy after every 10^5 iterations for each chain. (A) New York, d_1^w; (B) New York, d_2^w; (C) Southeast Asia, d_1^w; (D) Southeast Asia, d_2^w.

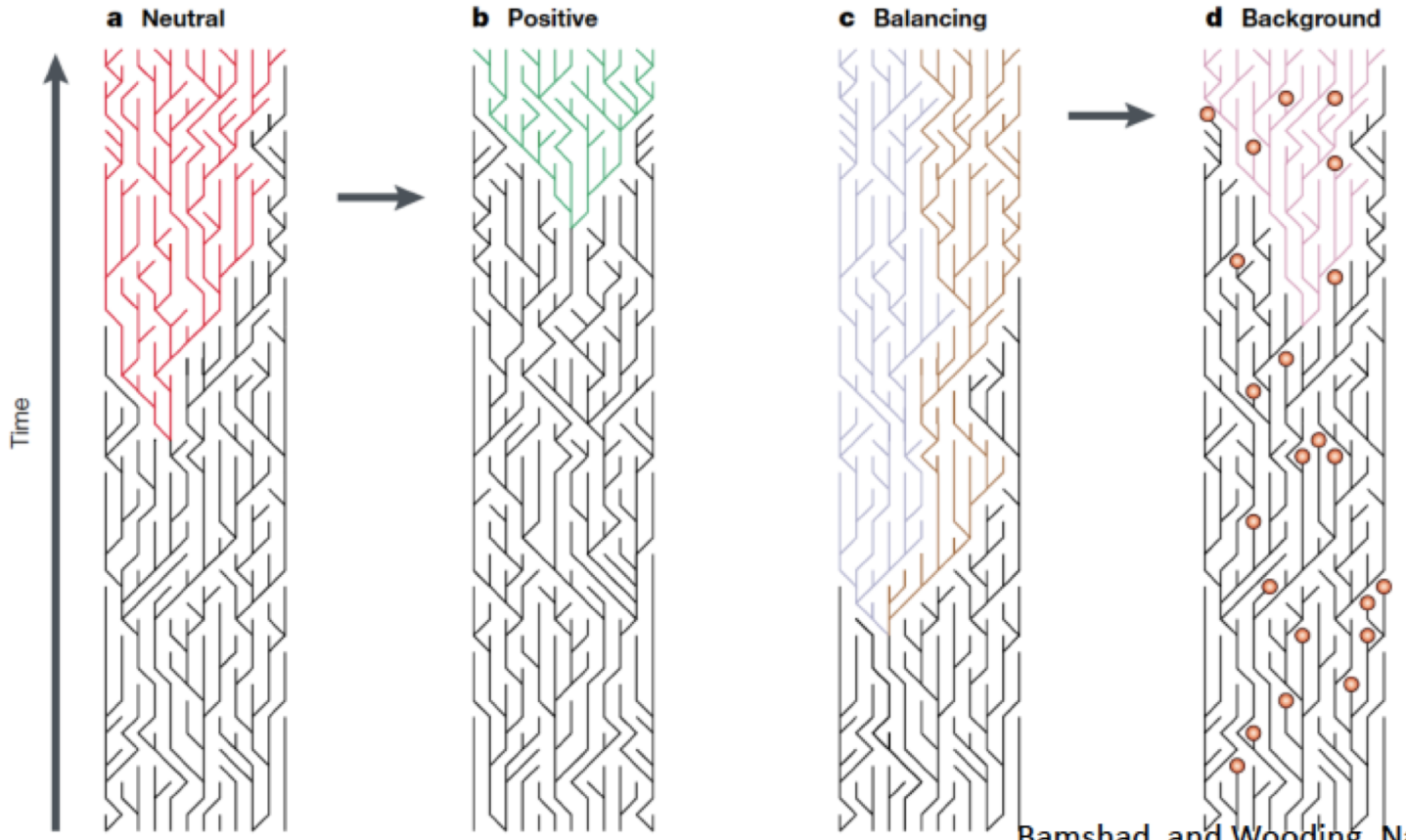
Tutorial and References

- https://github.com/JuliaPalacios/phylodyn/blob/master/vignettes/Distance_RankedGenealogies.Rmd

Kim J, Rosenberg NA, Palacios JA. **A Metric Space of Ranked Tree Shapes and Ranked Genealogies.** bioRxiv:2019.12.23.887125

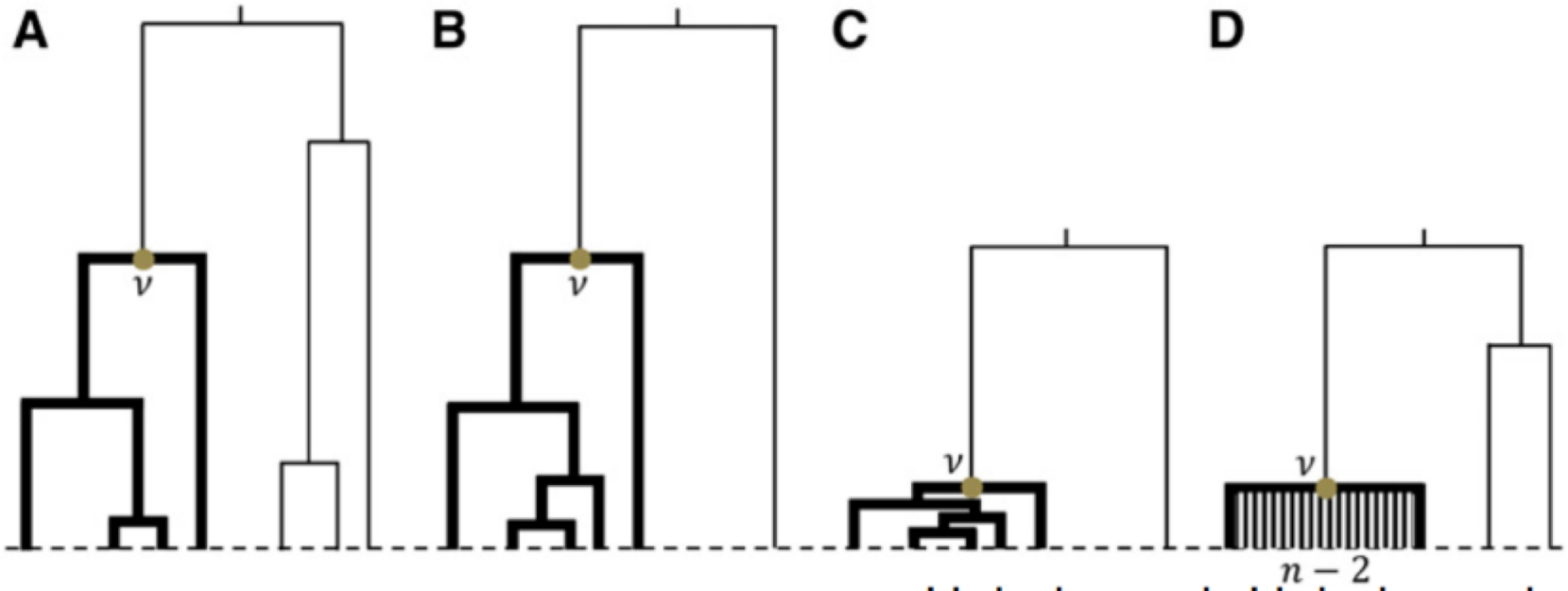
- Chakerian J, Holmes S, Computational tools for evaluating phylogenetic and hierarchical clustering trees. *JCGS*, 2012.
- Maliet O, Gascuel F and Lambert, A. Ranked Tree Shapes, Nonrandom Extinctions, and the Loss of Phylogenetic Diversity. *Syst. Biol.* 2018.

Genealogical signatures of natural selection in the Human genome



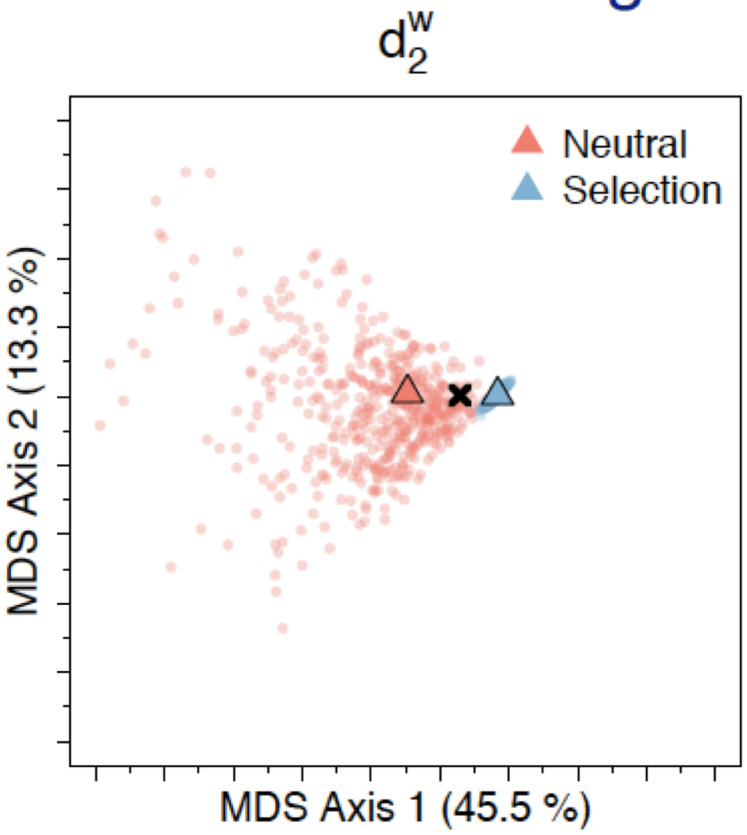
Bamshad and Wooding, Nat. Rev. Genet, 2003

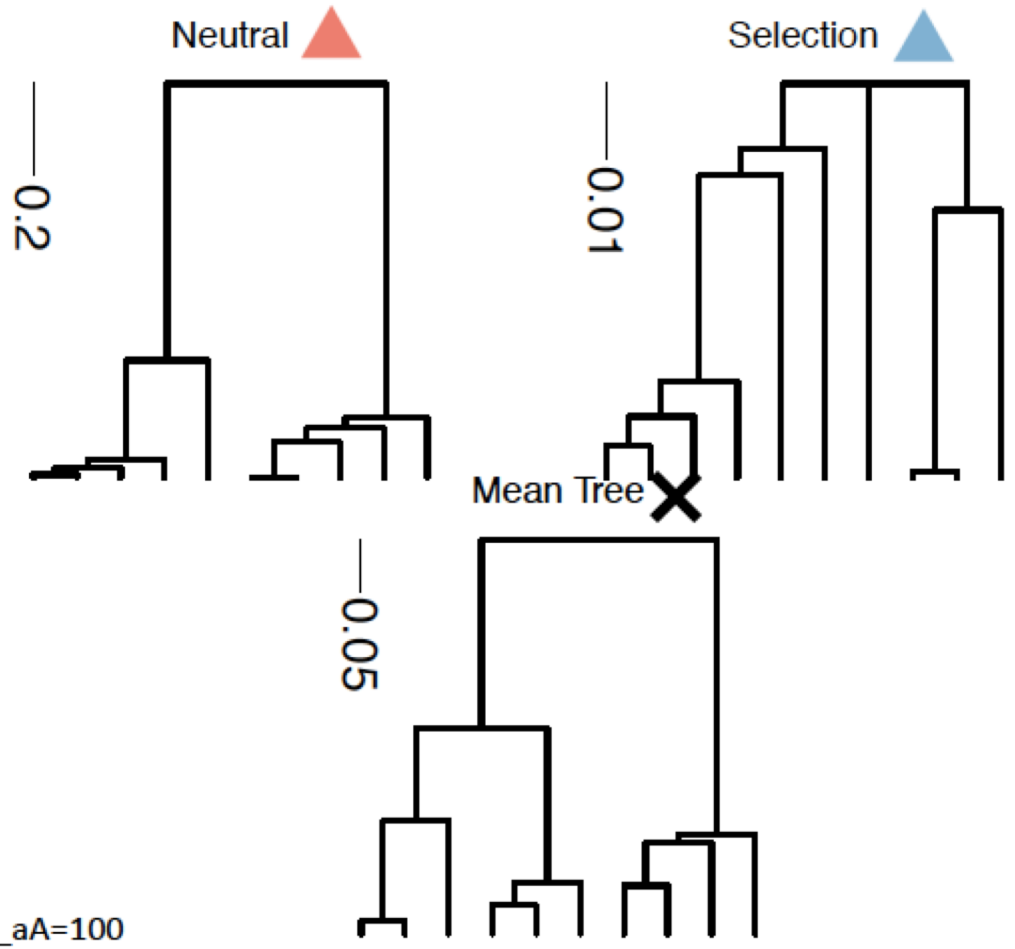
Genealogical signatures of natural selection in the Human genome



Yang et al, Genetics 2019

Simulation at a single locus: Neutral vs Selection





Neutral: n = 10 , Ne=100,000, θ=5

Positive selection: n=10, Ne=100,000, θ=5, s_AA=200, s_aA=100

Comparing empirical distributions