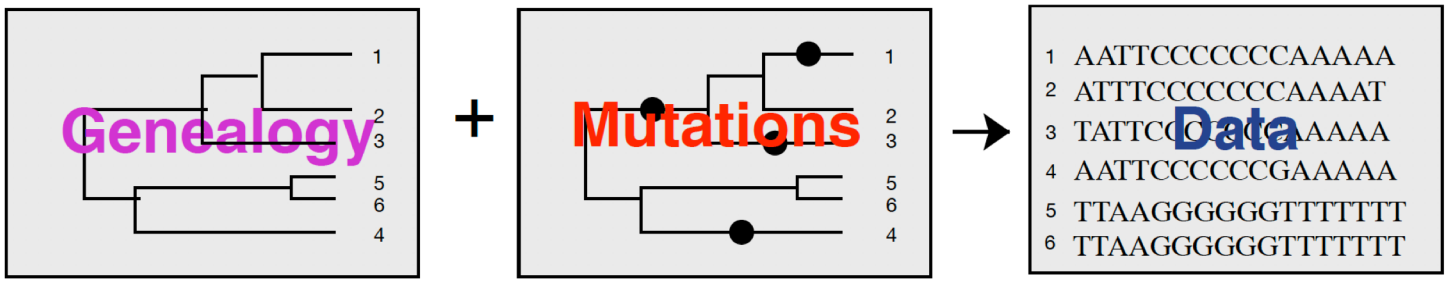
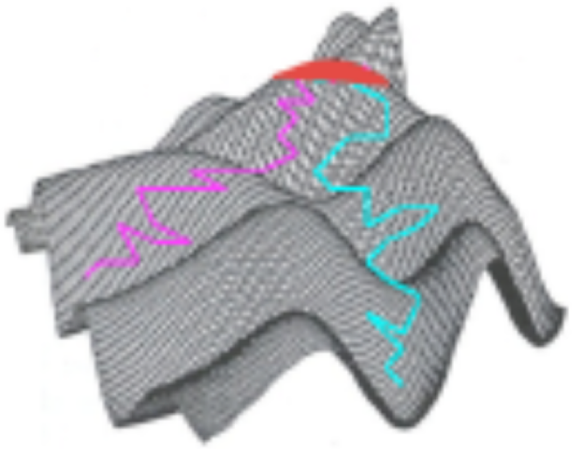
Posterior summary of trees

Intructor: Julia A. Palacios

Statistical Phylogenetics seeks to infer genealogies from molecular data





Goal: $\mathrm{P}(T \mid Y) \propto \mathrm{P}(Y \mid T, \mu)\mathrm{P}(T)\mathrm{P}(\mu)$

In Bayesian phylogenetics, many possible genealogies can explain the data

Which tree is representative of the sample?

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

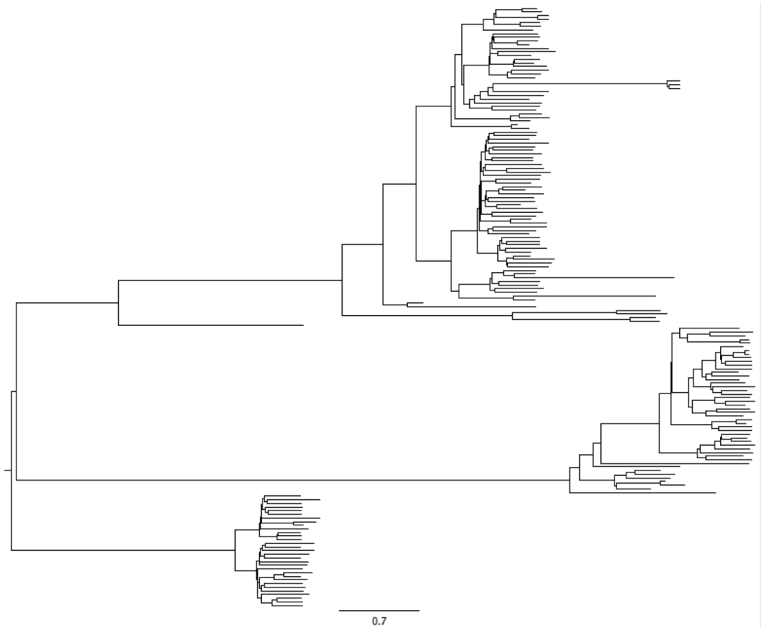
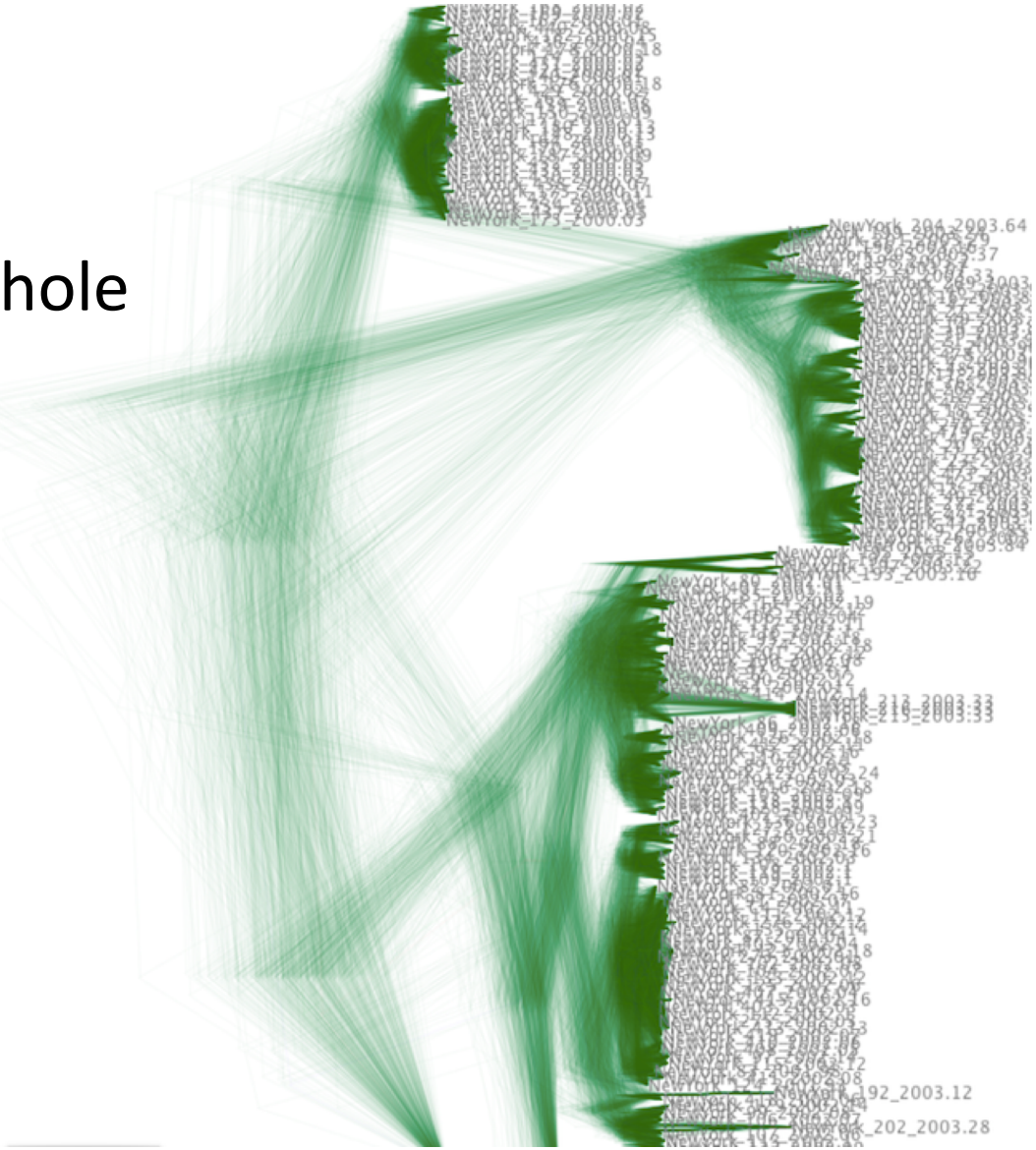


Figure: MCCT from Ne inference tutorial H3N2 Human influenza in New York

Densitree

- You can visualize the whole posterior distribution.



Other tree distances

Phylogenetic trees (labeled and unranked)

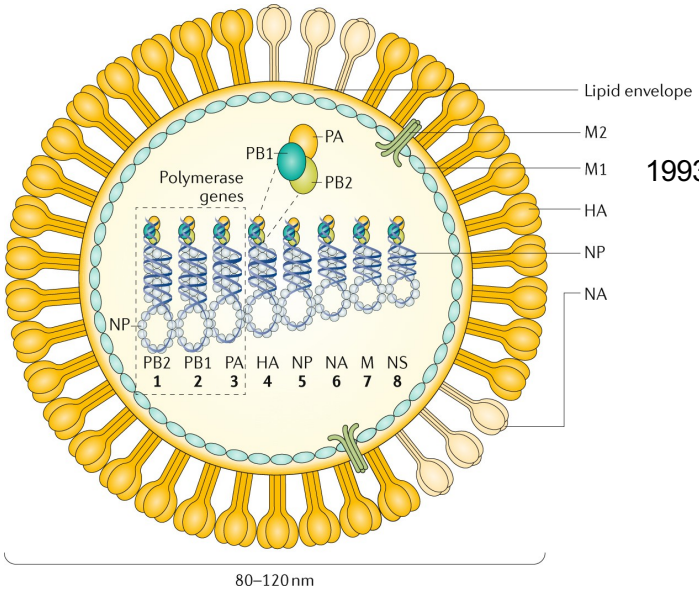
- Robinson-Foulds, Nearest Neighbor Interchange, Subree-Prune-and-Regraft, Tree Bisection and Reconnection.
- Geodesic distance between phylogenetic trees (Billera, Holmes and Vogtmann, 2001; Owen and Provan, 2009)
- Kendall-Colijn (2018)

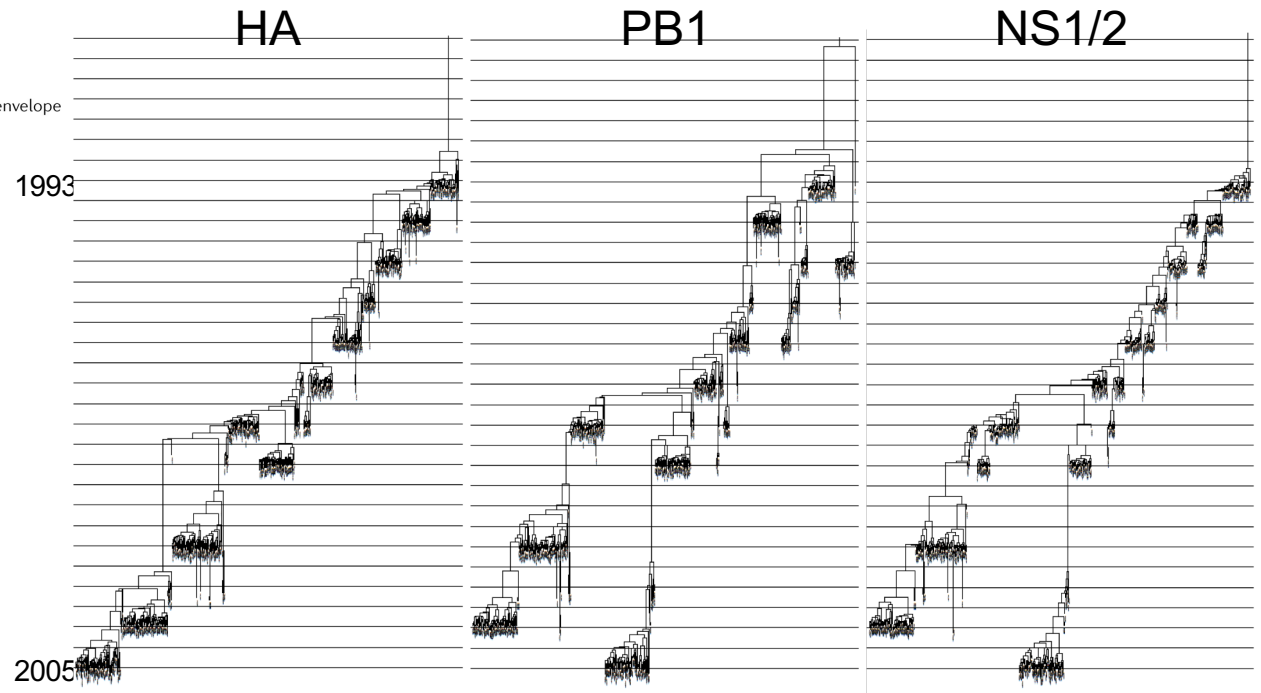
Tree Shapes (unlabeled and unranked)

- Colijn-Plazzotta (2017)

How different are the evolutionary processes across genes?

Influenza A/H3N2





Krammer, et al., 2018
Rambaut, et al., 2008

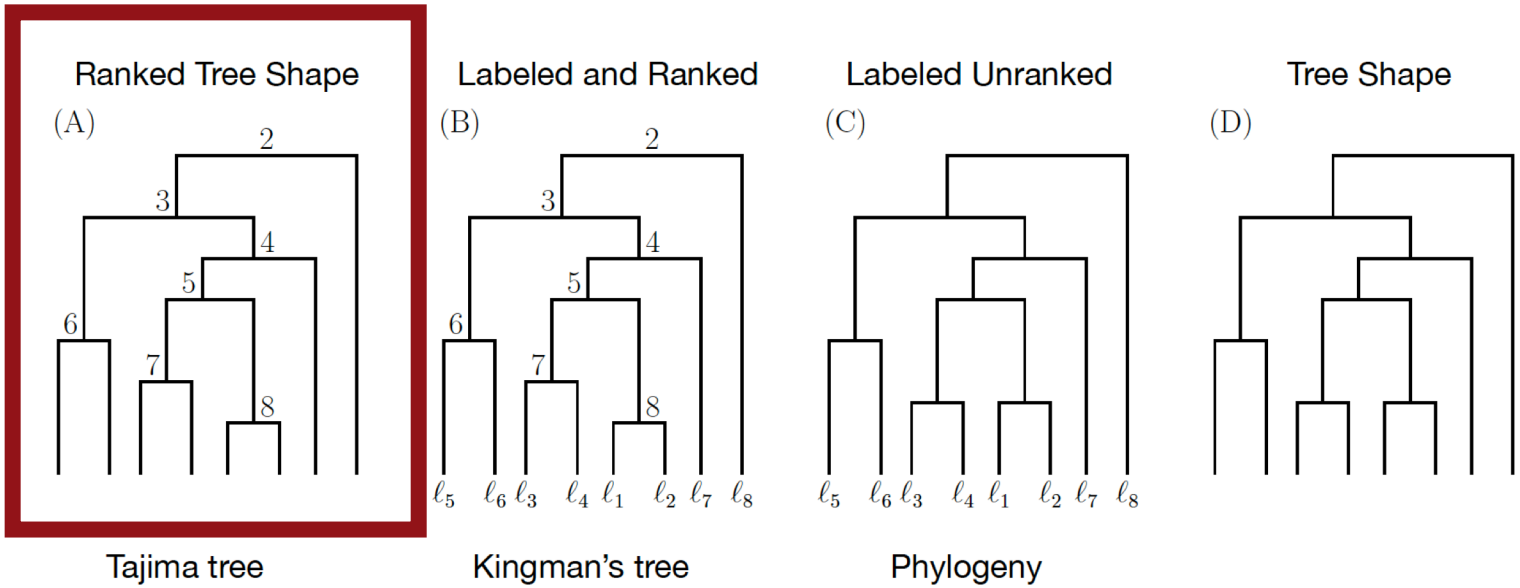
Why a metric on the space of unlabeled trees?

$d($  $)$?

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.

Tree resolutions

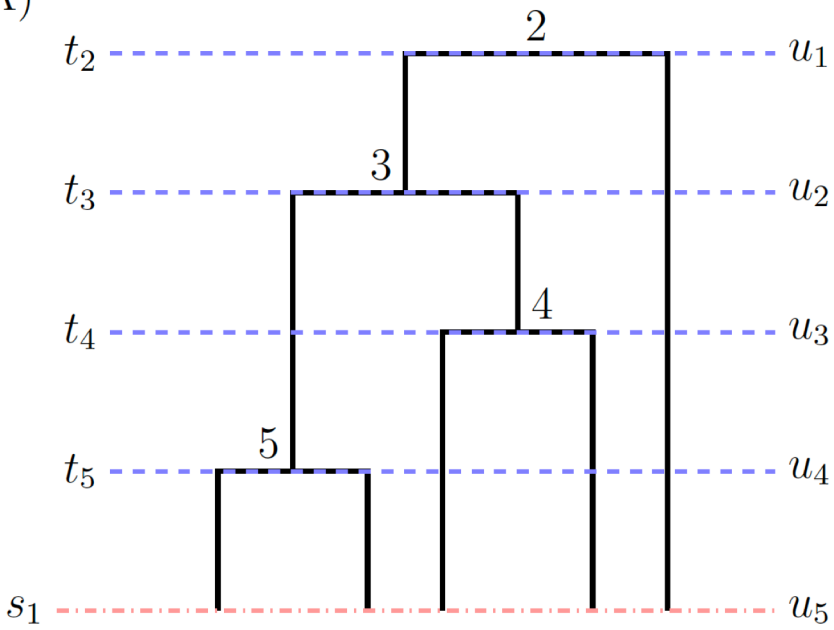


Ranked tree shapes are rooted binary trees with unlabeled leaves with an increasing ordering of internal nodes (root to leaves).

Ranked tree shapes as matrices

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

(A)



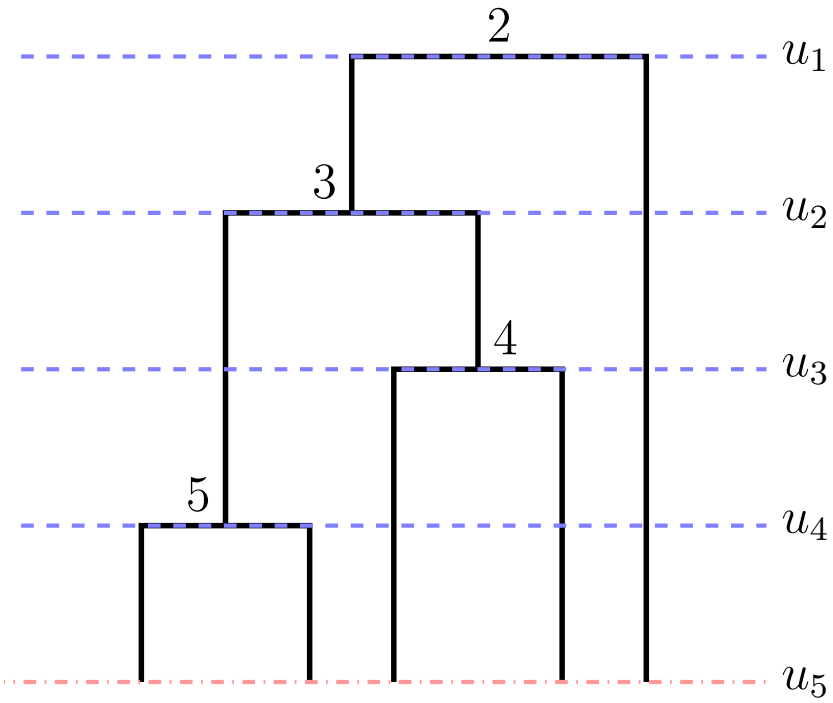
(B)

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

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

Kim et al., *PNAS* (2020)

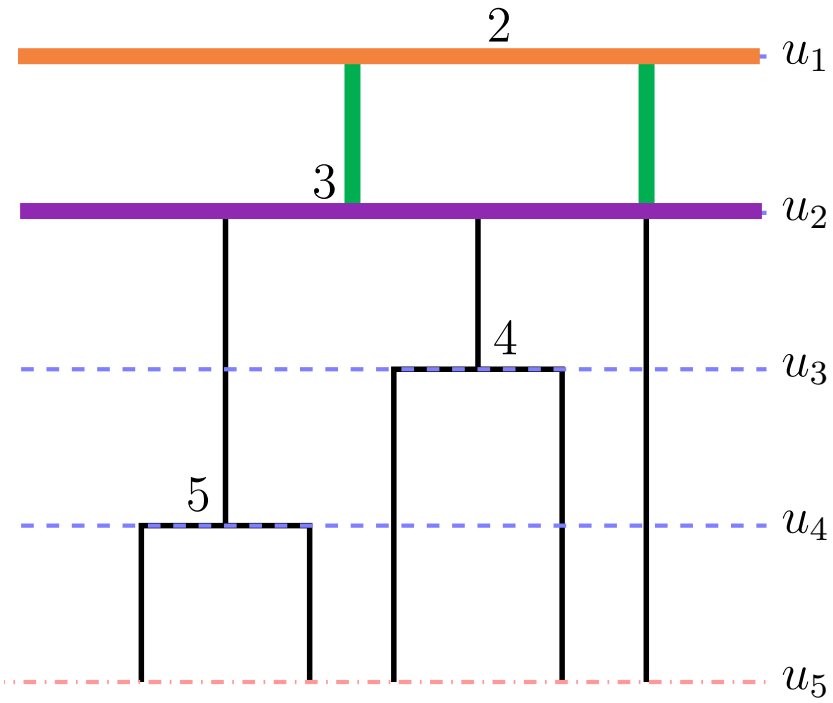
Unique encoding of a ranked tree shape



Start Time

End Time	u_1	u_2	u_3	u_4
u_2		0	0	0
u_3			0	0
u_4				0
u_5				

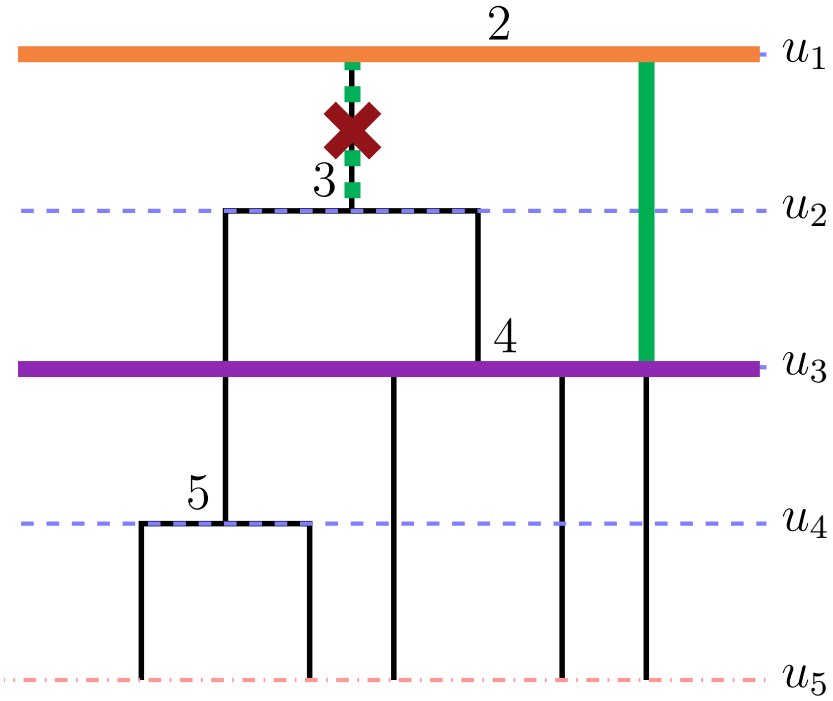
Unique encoding of a ranked tree shape



Start

End	u_1	u_2	u_3	u_4
u_2	2	0	0	0
u_3			0	0
u_4				0
u_5				

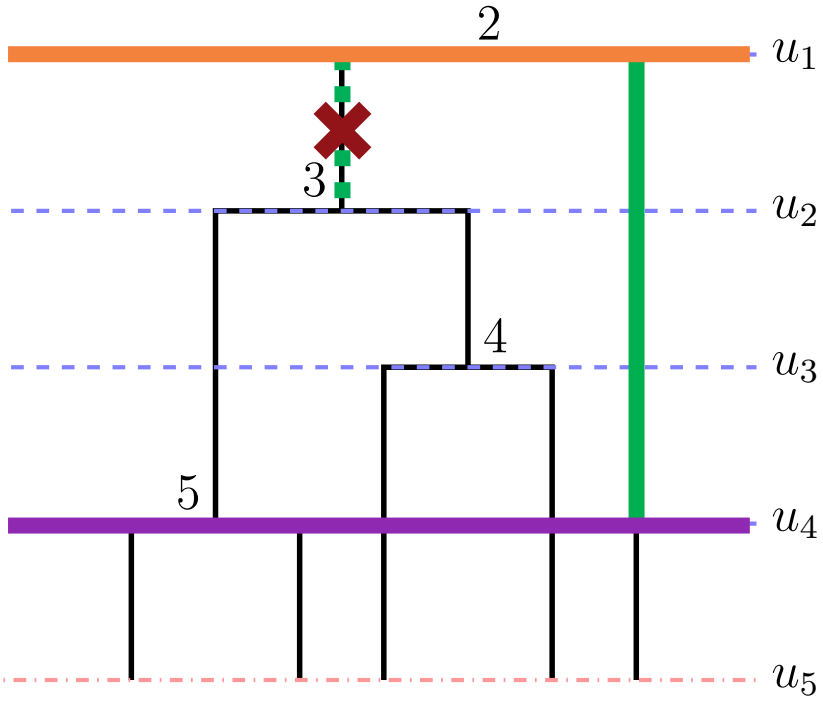
Unique encoding of a ranked tree shape



Start

End	u_1	u_2	u_3	u_4
u_2	2	0	0	0
u_3	1		0	0
u_4				0
u_5				

Unique encoding of a ranked tree shape

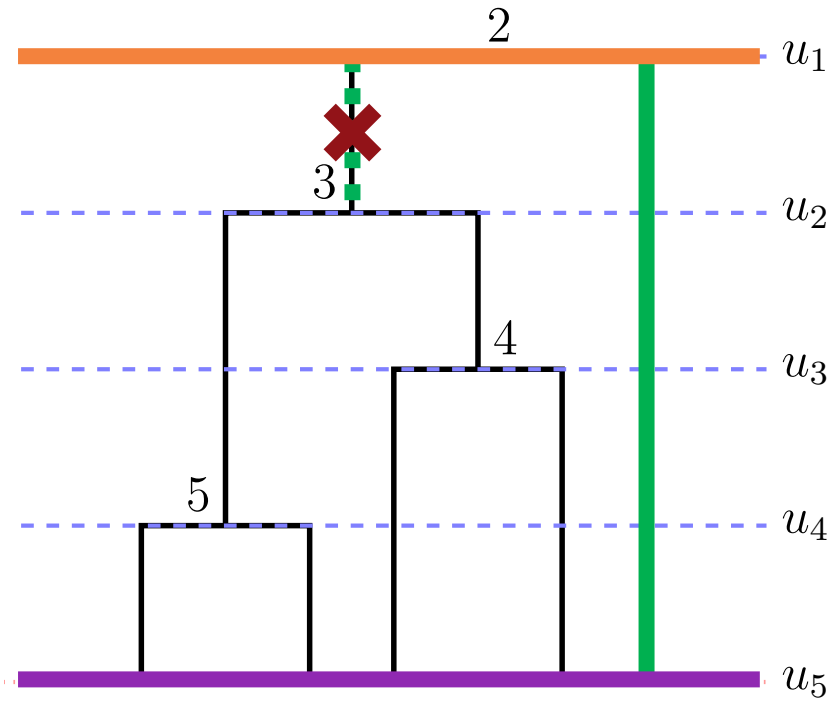


Start

End	u_1	u_2	u_3	u_4
u_2	2	0	0	0
u_3	1		0	0
u_4	1			0
u_5				

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

Unique encoding of a ranked tree shape

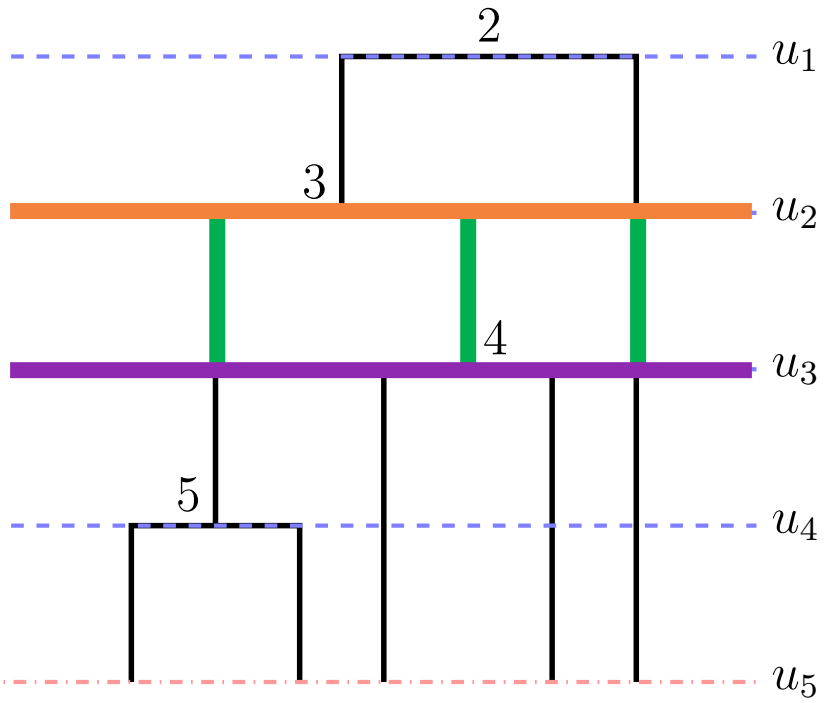


Start

End	u_1	u_2	u_3	u_4
u_2	2	0	0	0
u_3	1		0	0
u_4	1			0
u_5	1			

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

Unique encoding of a ranked tree shape

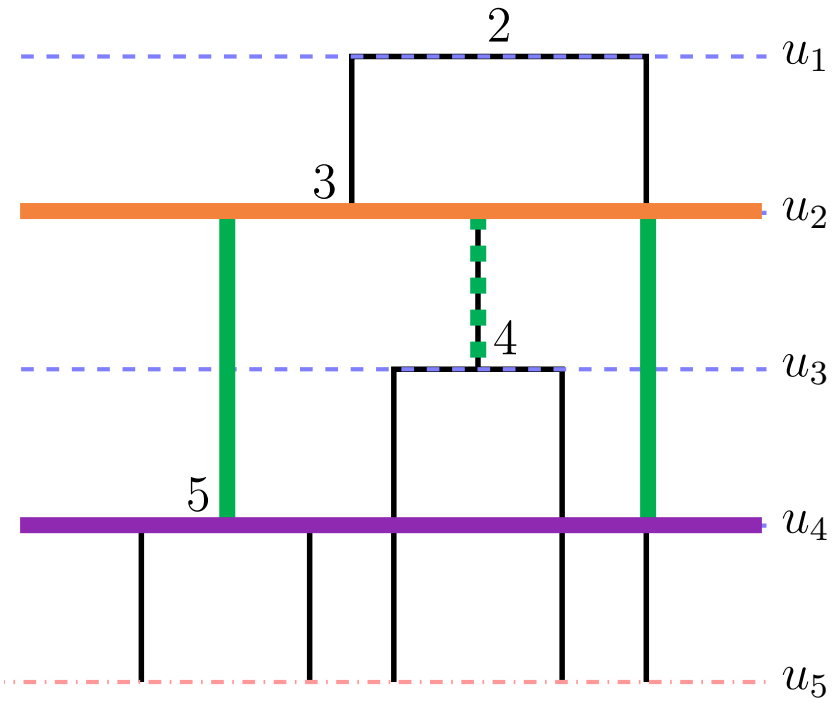


Start

End	u_1	u_2	u_3	u_4
u_2	2	0	0	0
u_3	1	3	0	0
u_4	1			0
u_5	1			

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

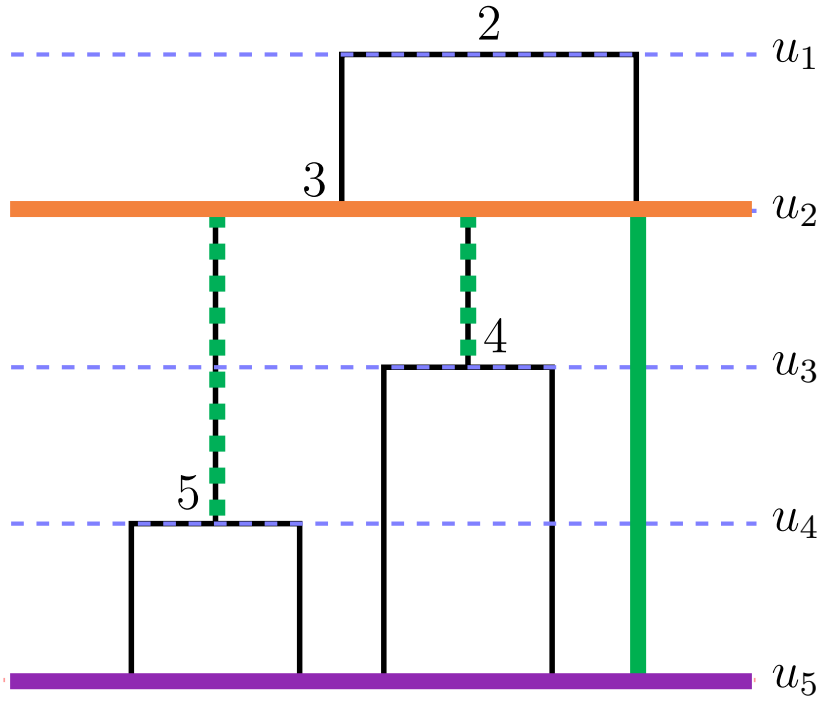
Unique encoding of a ranked tree shape



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

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

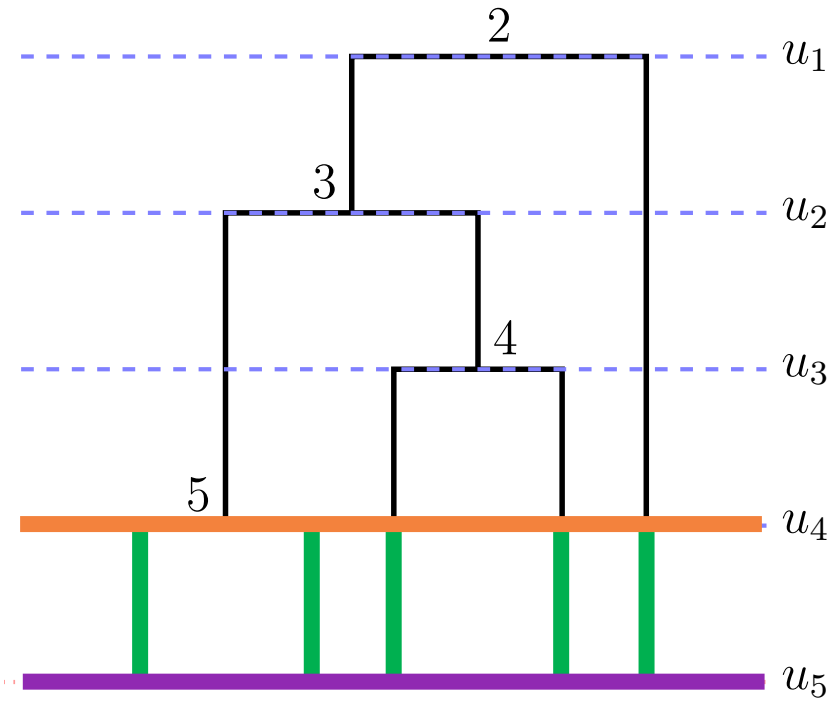
Unique encoding of a ranked tree shape



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

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

Unique encoding of a ranked tree shape



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

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

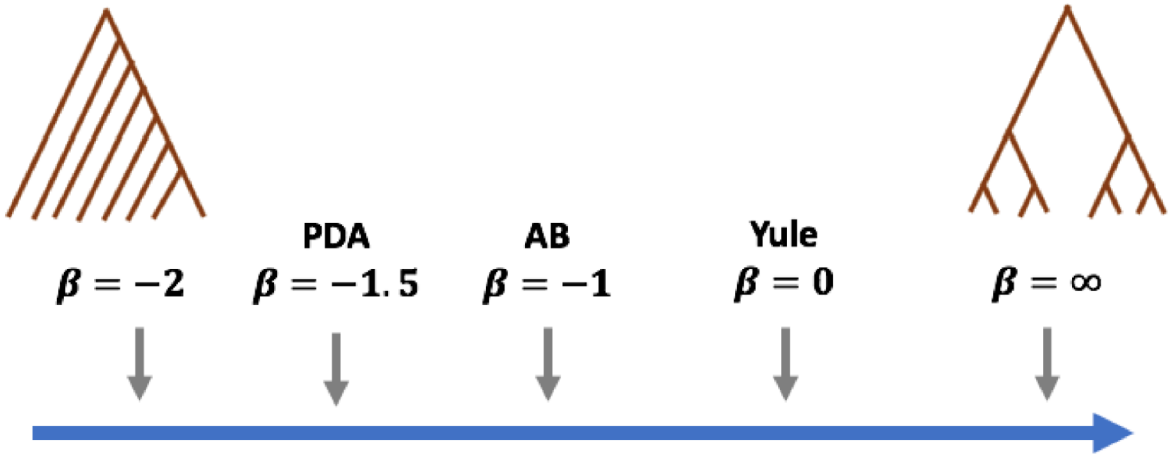
Metrics on ranked tree shapes

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

$$d_2(T^{(1)}, T^{(2)}) = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{i} \left(F_{i,j}^{(1)} - F_{i,j}^{(2)} \right)^2}$$

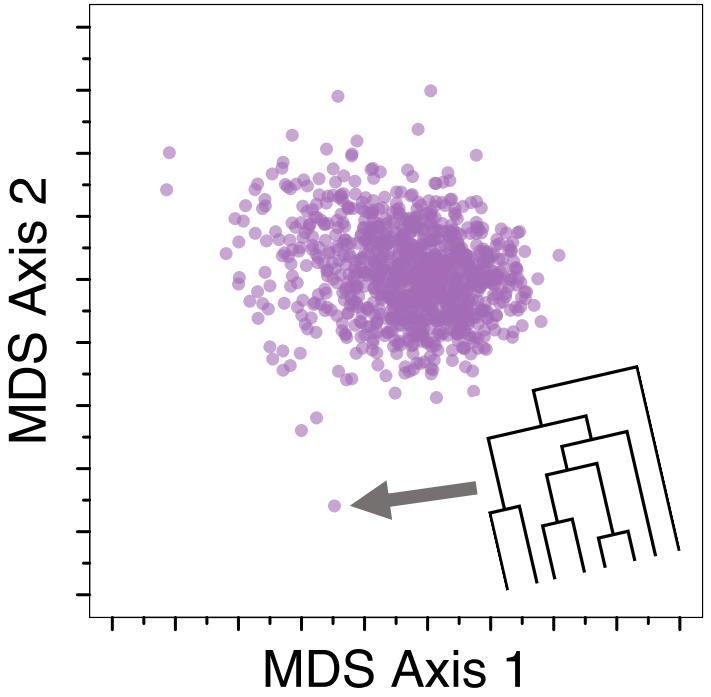
Testing distance on ranked tree shapes...

Beta-splitting model

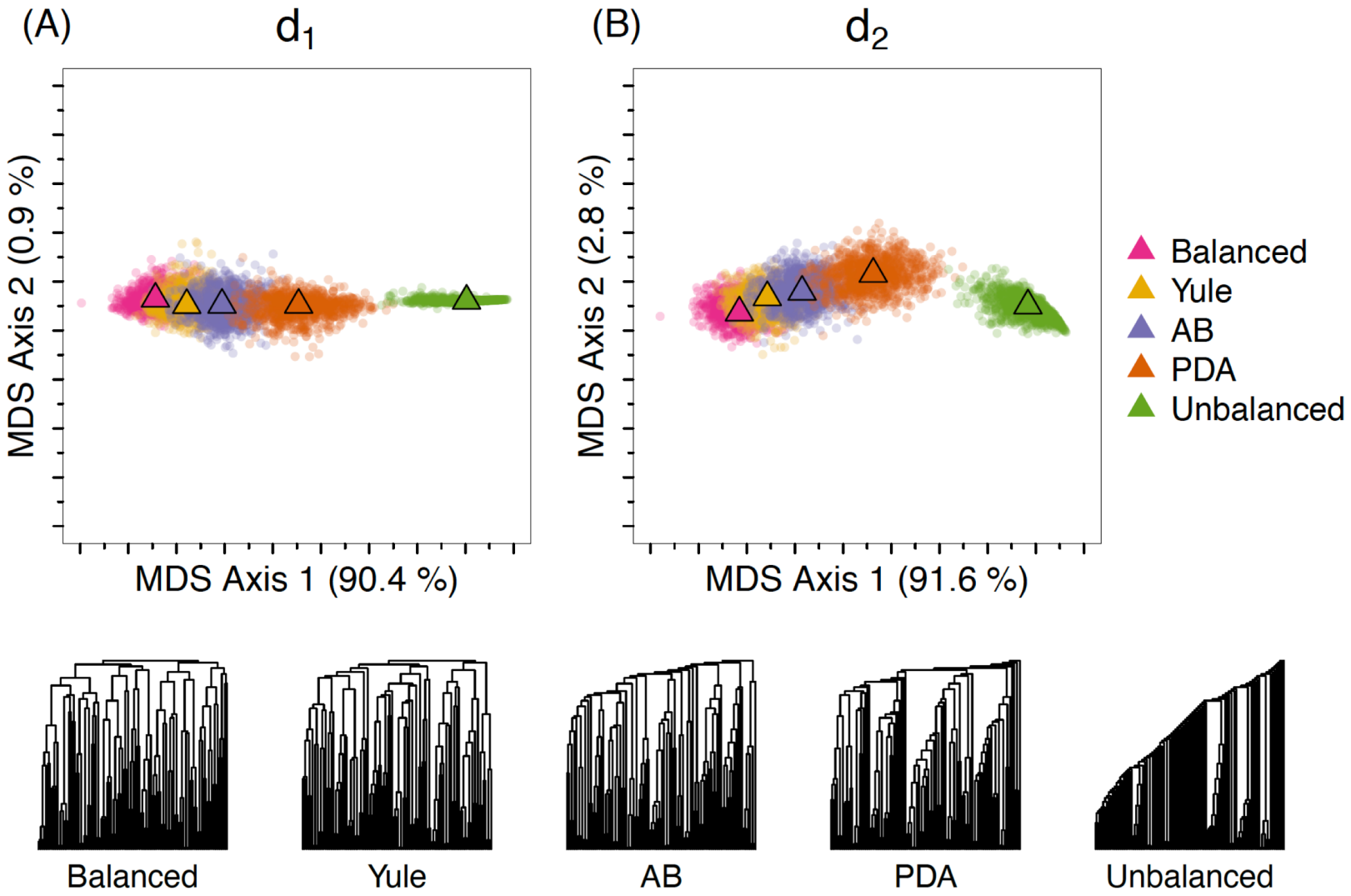


MDS Visualization for comparing tree distributions

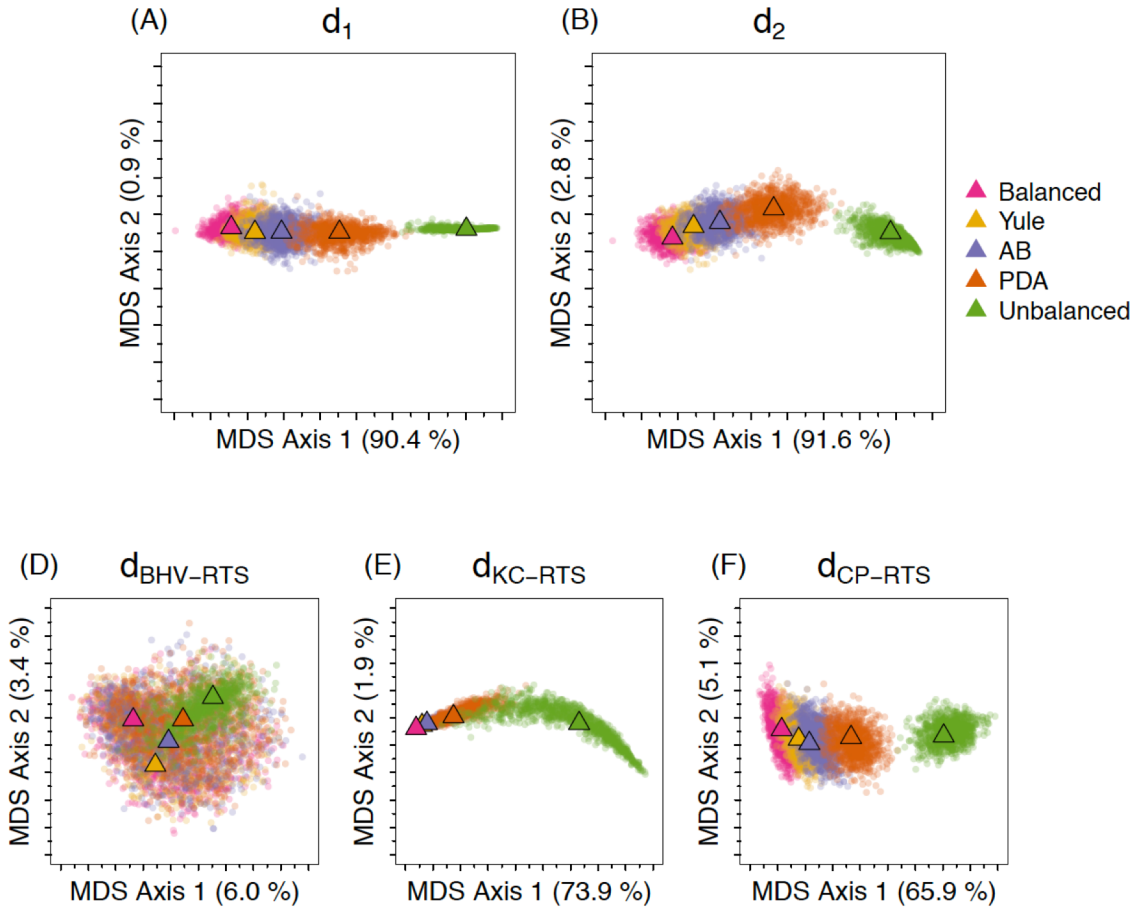
$$D = \begin{bmatrix} d(T_1, T_1) & d(T_1, T_2) & \cdots & d(T_1, T_N) \\ d(T_2, T_1) & d(T_2, T_2) & \cdots & d(T_2, T_N) \\ \vdots & \vdots & \ddots & \vdots \\ d(T_N, T_1) & d(T_N, T_2) & \cdots & d(T_N, T_N) \end{bmatrix}$$



Simulations from Beta-splitting on ranked tree shapes



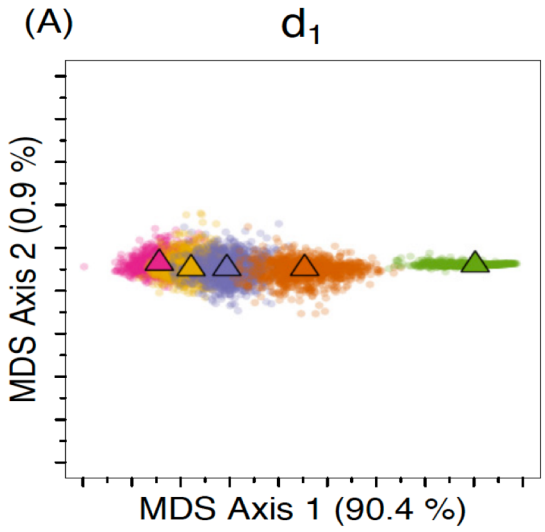
Comparing separation of different Beta-splitting distributed ranked tree shapes with other metrics



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.

(A) d_1

	▲	▲	▲	▲	▲
Balanced	85.9	14.0	0.1	0	0
Yule	22.9	64.4	12.7	0	0
AB	1.1	19.7	73.6	5.6	0
PDA	0	0	7.2	92.5	0.3
Unbalanced	0	0	0	0	100.0

(B) d_2

▲	▲	▲	▲	▲
82.6	17.3	0.1	0	0
23.1	62.8	14.1	0	0
1.4	19.7	74.2	4.7	0
0	0	9.2	90.6	0.2
0	0	0	0.1	99.9

L_2-medoid

- ▲ Balanced
- ▲ Yule
- ▲ AB
- ▲ PDA
- ▲ Unbalanced

(C) $d_{\text{BHV-RTS}}$

	▲	▲	▲	▲	▲
Balanced	1.0	0	0	0.2	98.8
Yule	0.5	0.2	0	0.2	99.1
AB	0.5	0.1	0.2	0.2	99.0
PDA	0.2	0	0	0.3	99.5
Unbalanced	0	0	0	0.1	99.9

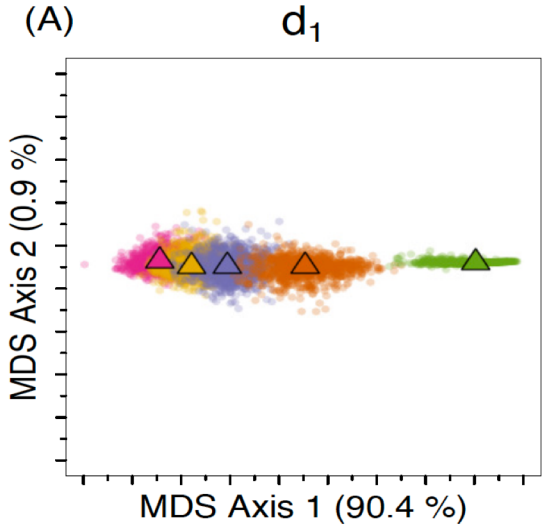
(D) $d_{\text{KC-RTS}}$

▲	▲	▲	▲	▲
100.0	0	0	0	0
56.3	36.0	7.7	0	0
7.9	27.3	55.2	9.6	0
0	2.5	20.9	76.3	0.3
0	0	0.2	14.8	85.0

(E) $d_{\text{CP-RTS}}$

▲	▲	▲	▲	▲
77.1	22.5	0.4	0	0
29.8	51.3	18.7	0.2	0
4.8	26.4	57.7	11.1	0
0	0.5	10.7	88.8	0
0	0	0	0.1	99.9

Nonparametric Test For equality in ranked tree shape distribution



Test statistic:

$$C^{x,y} = \frac{1}{2(N-1)} \sum_{j=1}^{N} \left[1_{d(X_j,\bar{Y}) \leq d(X_j,\bar{X})} + 1_{d(Y_j,\bar{X}) \leq d(Y_j,\bar{Y})} \right]$$

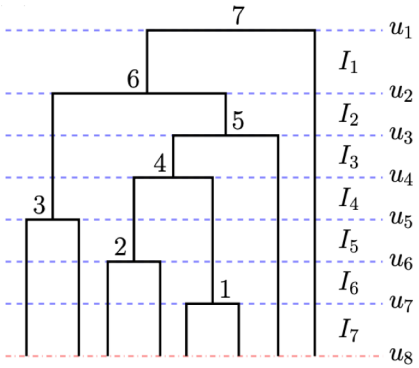
Permutation p value:

$$P = \frac{1 + \sum_{k=1}^{N_{\text{perm}}} 1_{C_k^{x,y} \leq C^{x,y}}}{1 + N_{\text{perm}}}$$

A distance between unlabeled genealogies

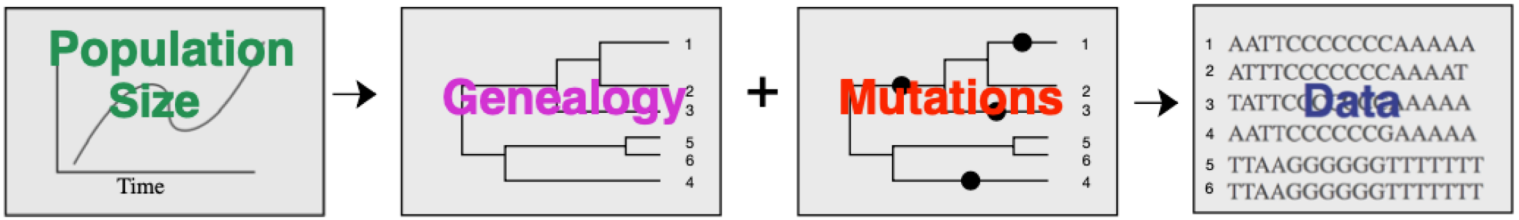
$$d_1(G_1, G_2) := \sum_{i,j} |(F_1)_{ij}(W_1)_{ij} - (F_2)_{ij}(W_2)_{ij}|,$$

$$d_2(G_1, G_2) := \sqrt{\sum_{i,j} ((F_1)_{ij}(W_1)_{ij} - (F_2)_{ij}(W_2)_{ij})^2},$$



$$\mathbf{F} = \begin{pmatrix} 2 & & & & & & \\ 1 & 3 & & & & & \\ 1 & 2 & 4 & & & & \\ 1 & 2 & 3 & 5 & & & \\ 1 & 1 & 2 & 4 & 6 & & \\ 1 & 1 & 2 & 3 & 5 & 7 & \\ 1 & 1 & 2 & 2 & 4 & 5 & 8 \end{pmatrix} W = \begin{pmatrix} u_1 - u_2 & & & & & & \\ u_1 - u_3 & u_2 - u_3 & & & & & \\ u_1 - u_4 & u_2 - u_4 & u_3 - u_4 & & & & \\ u_1 - u_5 & u_2 - u_5 & u_3 - u_5 & u_4 - u_5 & & & \\ u_1 - u_6 & u_2 - u_6 & u_3 - u_6 & u_4 - u_6 & u_5 - u_6 & & \\ u_1 - u_7 & u_2 - u_7 & u_3 - u_7 & u_4 - u_7 & u_5 - u_7 & u_6 - u_7 & \\ u_1 & u_2 & u_3 & u_4 & u_5 & u_5 & u_6 \end{pmatrix}$$

Coalescent models on genealogies



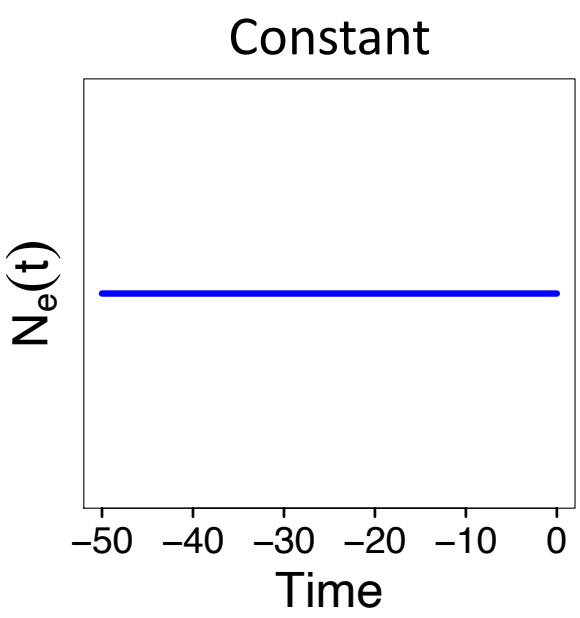
Coalescent prior: (the pure death proc. and the discrete jump chain are independent)

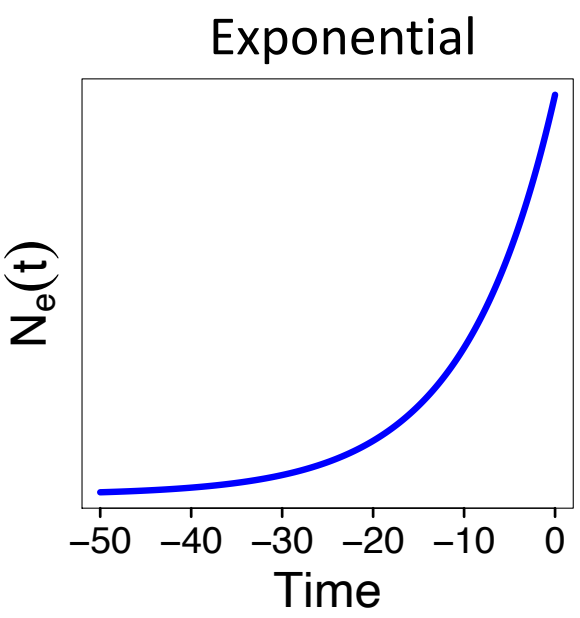
$$P[\mathbf{g} = (\mathbf{F}, \mathbf{u}) \mid N_e(t)] = P(\mathbf{F}) \prod_{k=2}^{n} P[u_k \mid u_{k+1}, N_e(t)]$$

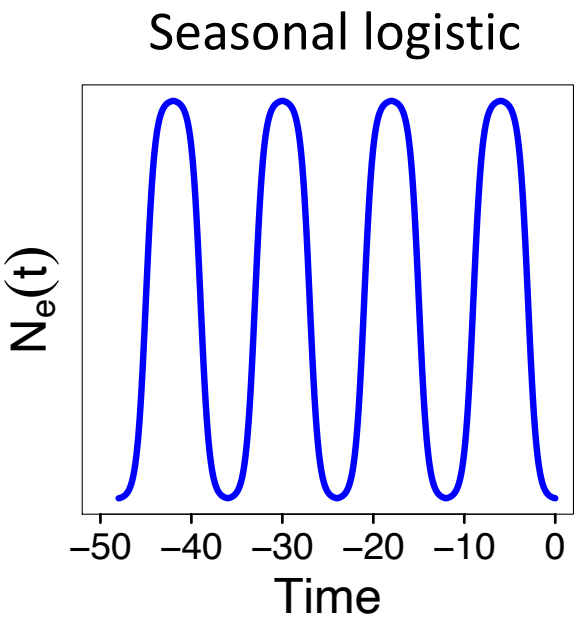
$$P[u_k \mid u_{k+1}, N_e(t)] = \frac{C_k}{N_e(u_k)} \exp\left[- \int_{u_{k+1}}^{u_k} \frac{C_k dt}{N_e(t)} \right]$$

where $C_k = \binom{k}{2}$ is the coalescent factor that depends on the number of lineages $k = 2, \ldots, n$.

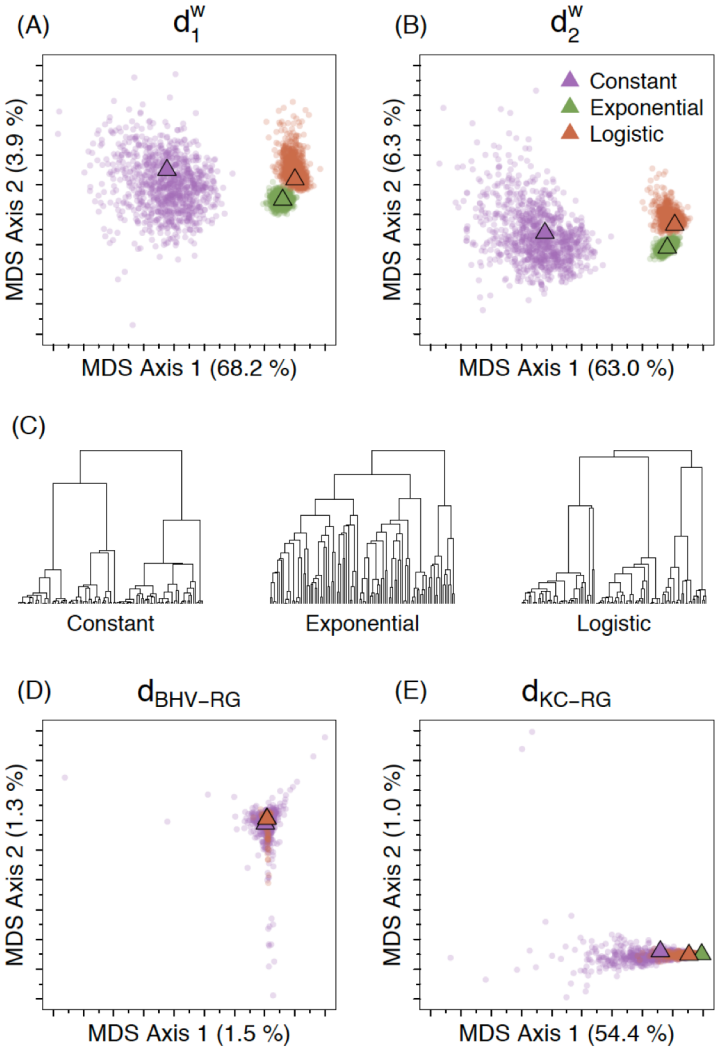
Simulations with different Ne(t) trajectories





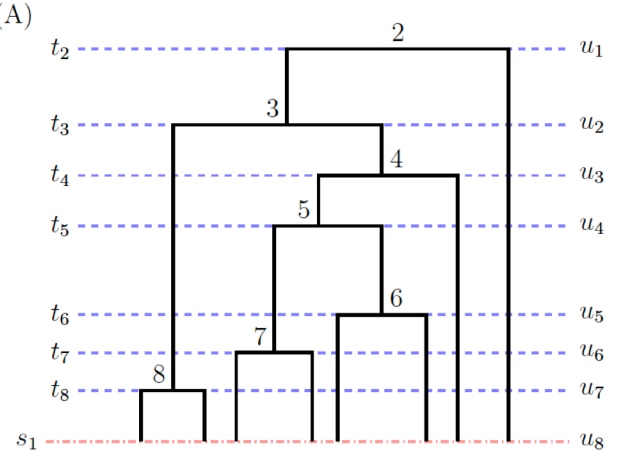


Simulation of genealogies with different branch distributions



Adapting the distances to heterochronous ranked tree shapes

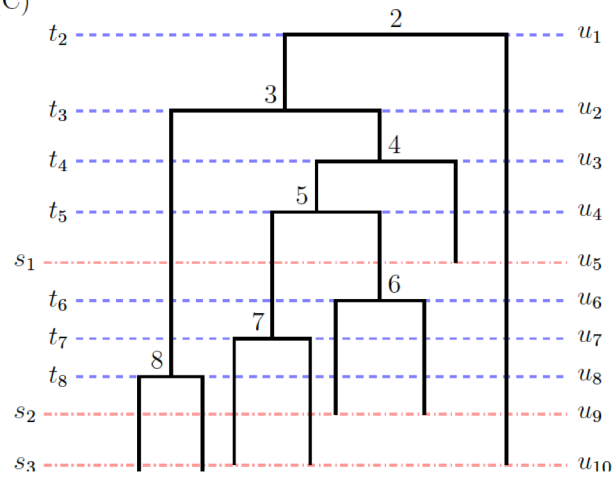
(A)



(B)

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

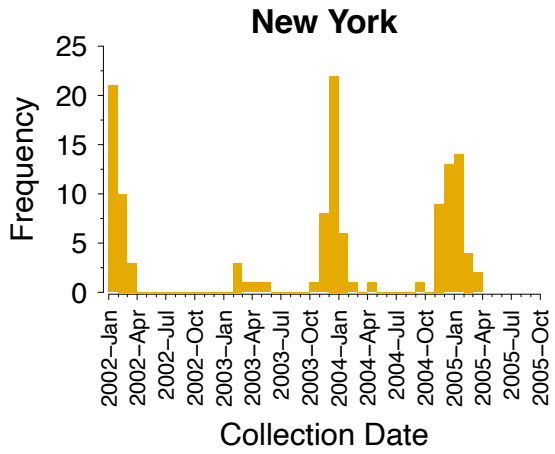
(C)

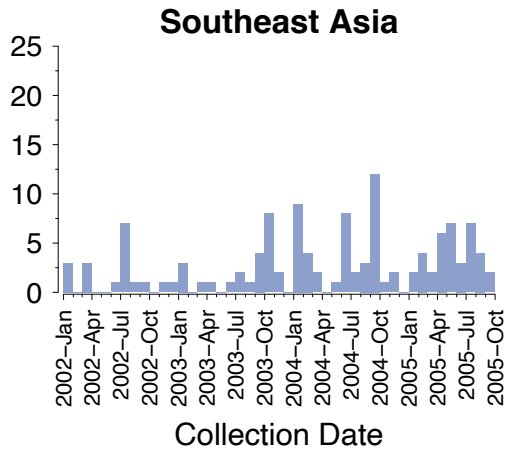


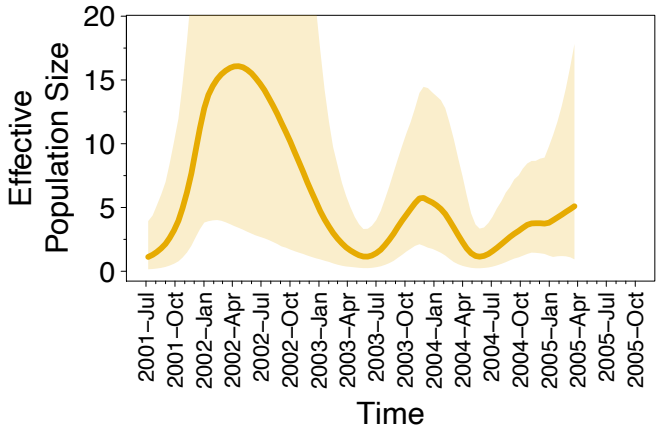
(D)

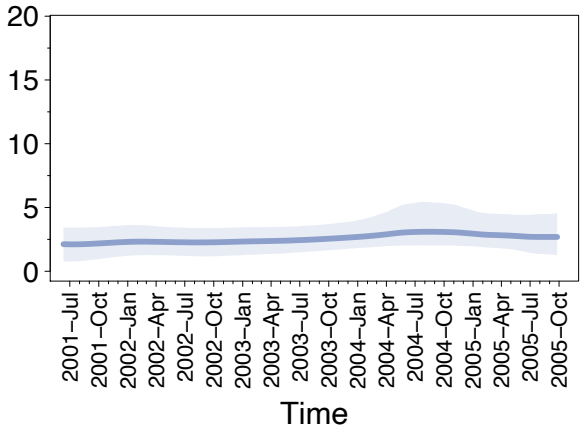
$$\mathbf{F} = \left(\begin{array}{cccc|cccc|cc} 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 2 & 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 2 & 3 & 5 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 1 & 2 & 2 & 4 & 4 & 0 & 0 & 0 & 0 & 0 \\ 1 & 2 & 2 & 3 & 3 & 5 & 0 & 0 & 0 & 0 \\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

Human influenza A/H3N2

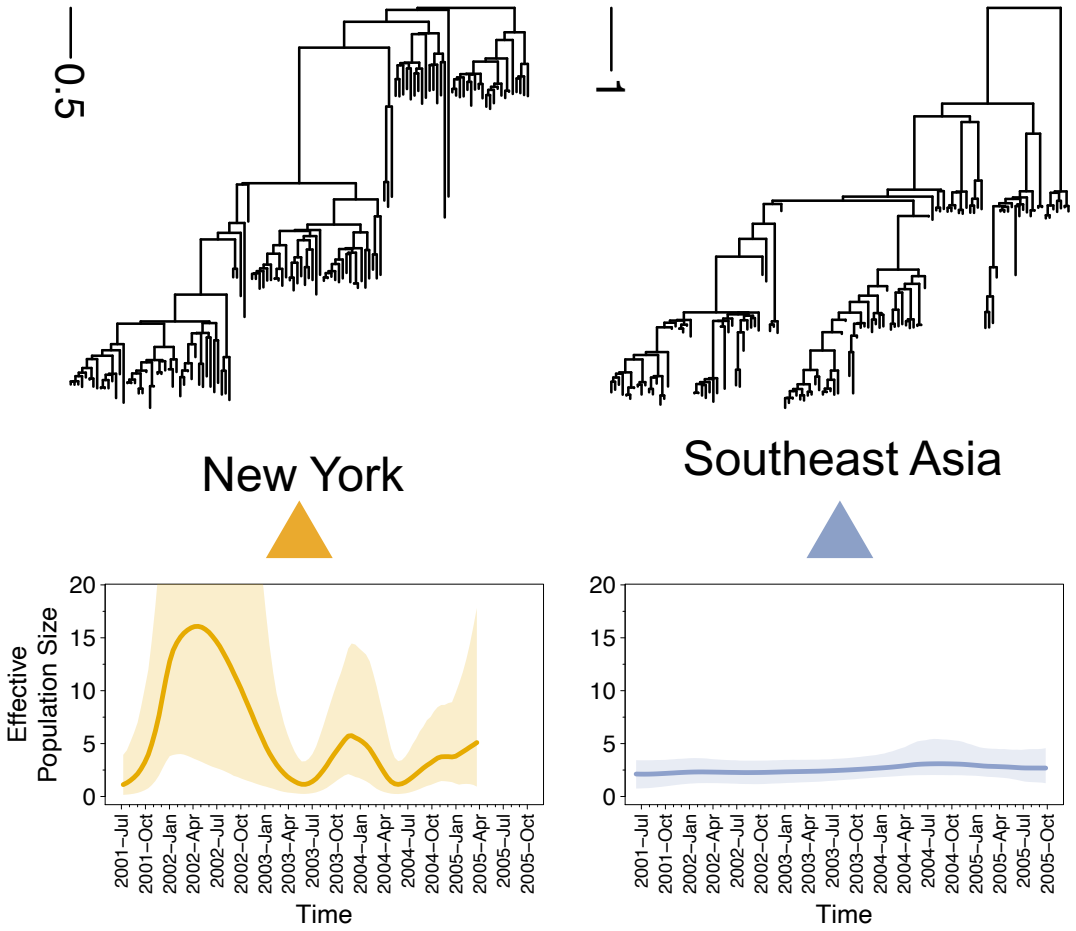


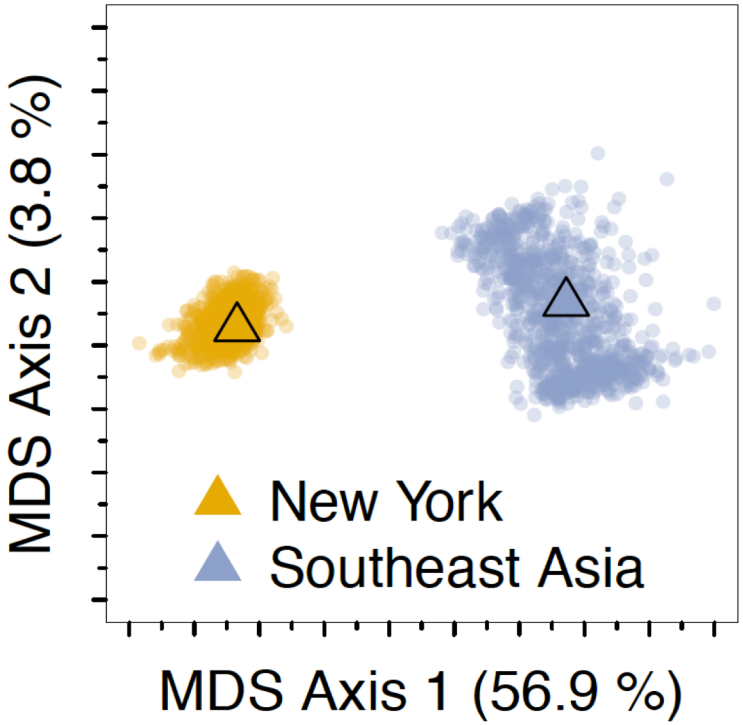






Distance metric differentiate evolutionary histories





Convergence diagnostic

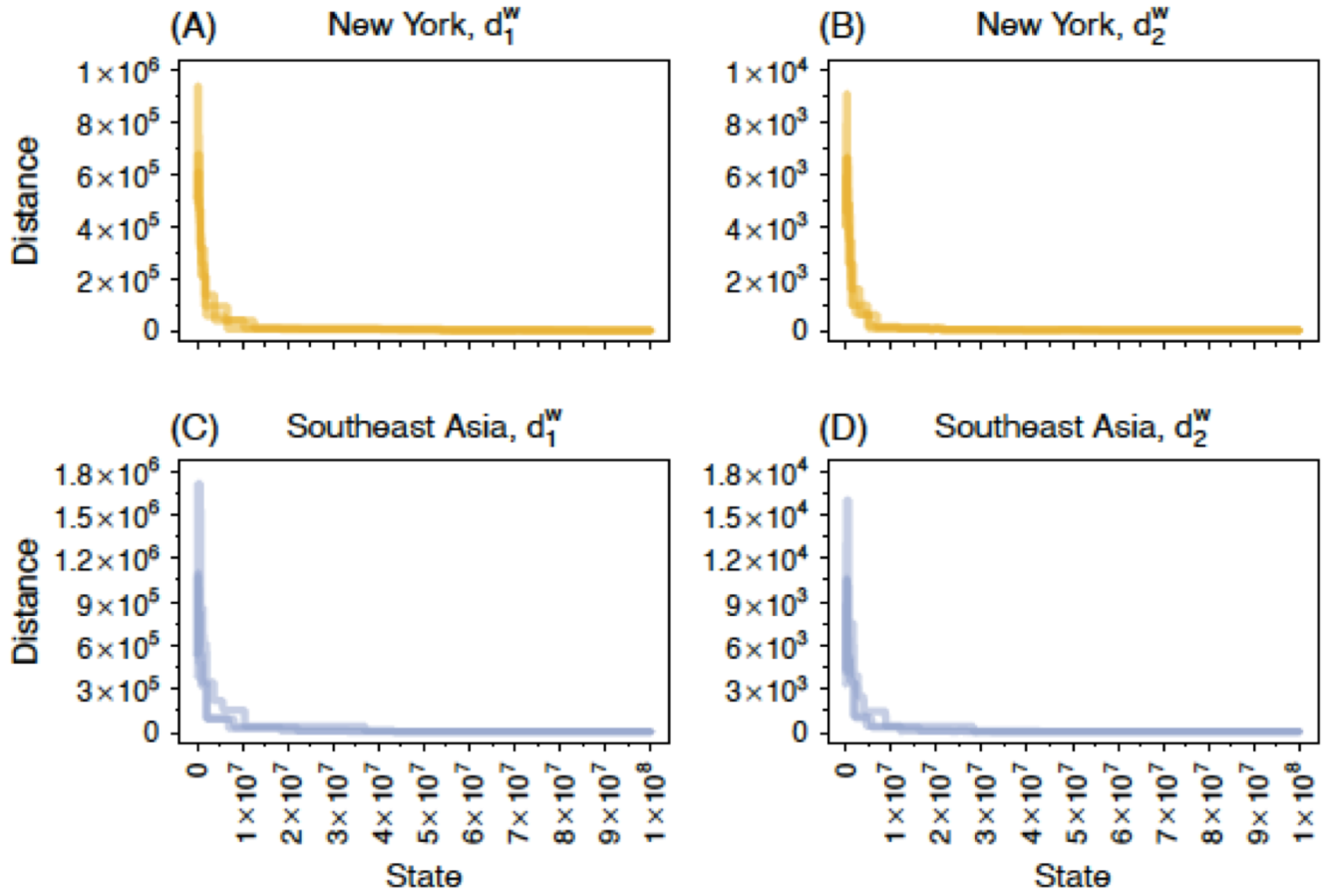
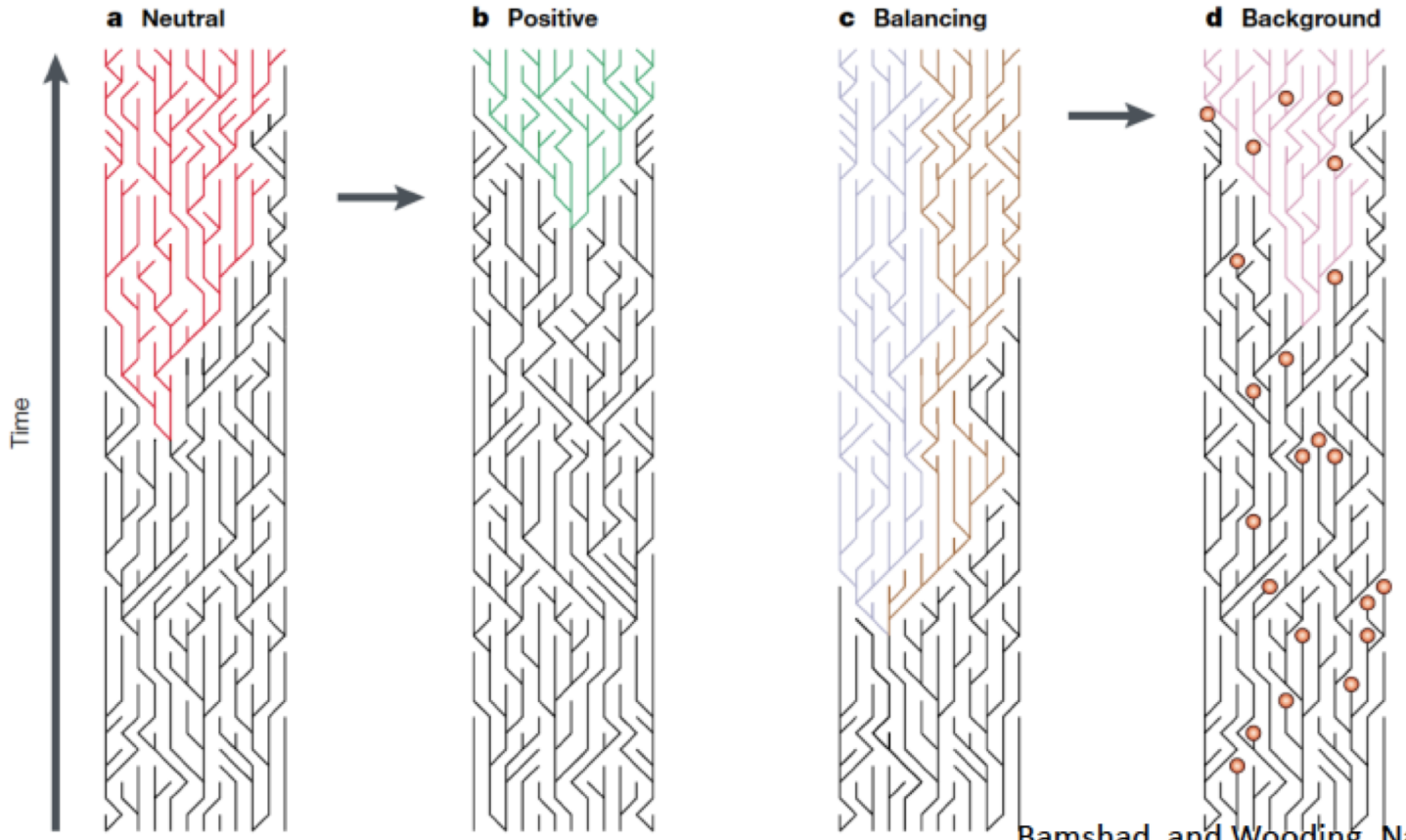


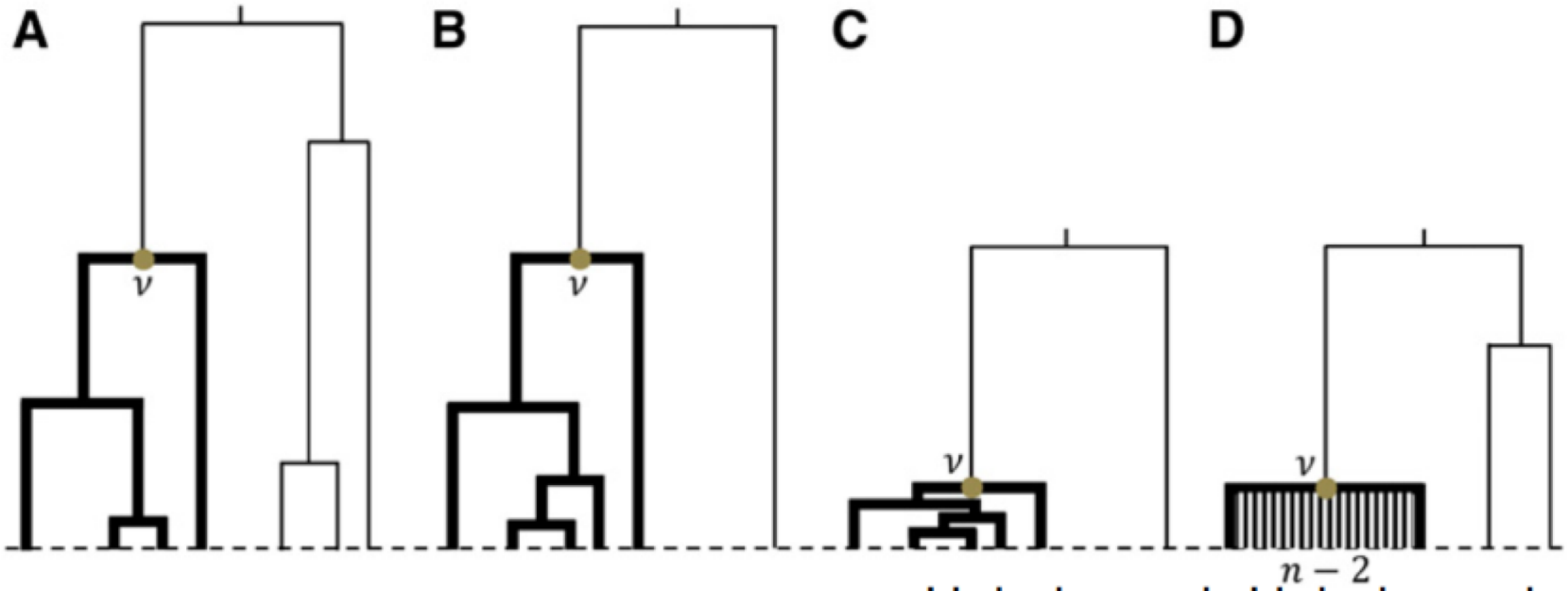
Fig. 9. Assessment of convergence of MCMC BEAST chains. Each curve corresponds to the distance between the running posterior L_2-medoid ranked genealogy and the global posterior L_2-medoid ranked genealogy after every 10^5 iterations for each chain. (A) New York, d_1^w; (B) New York, d_2^w; (C) Southeast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

Genealogical signatures of natural selection in the Human genome



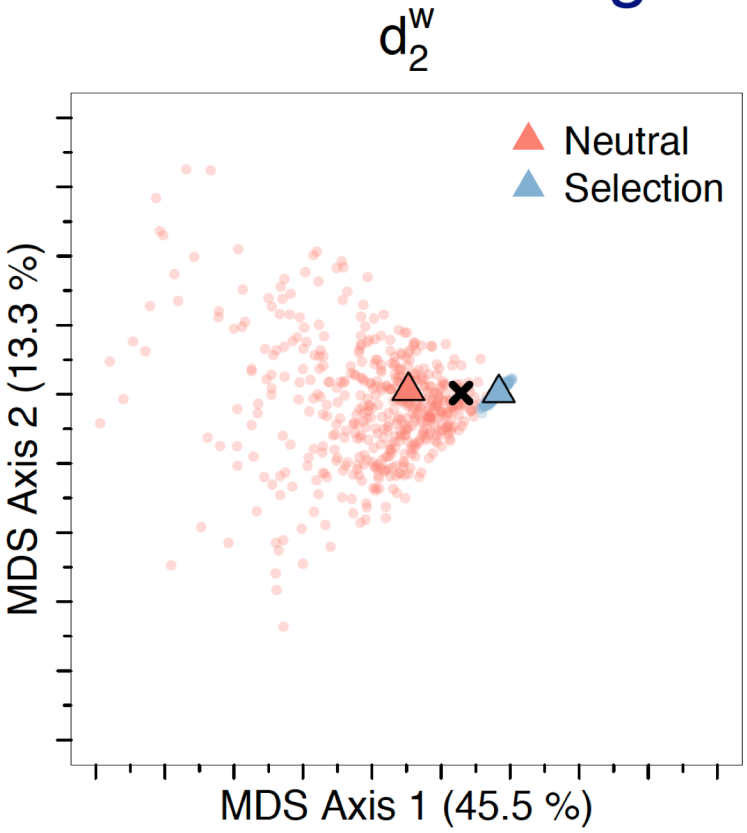
Bamshad and Wooding, Nat. Rev. Genet, 2003

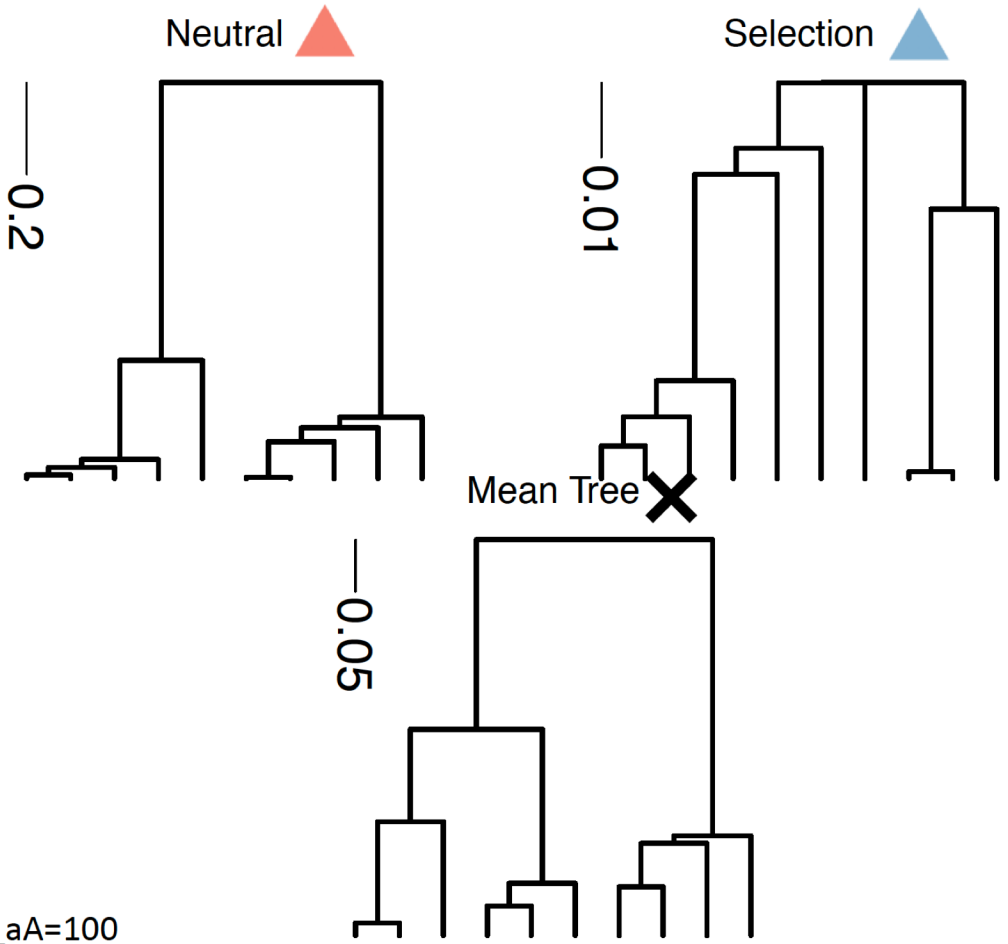
Genealogical signatures of natural selection in the Human genome



Yang et al, Genetics 2019

Simulation at a single locus: Neutral vs Selection



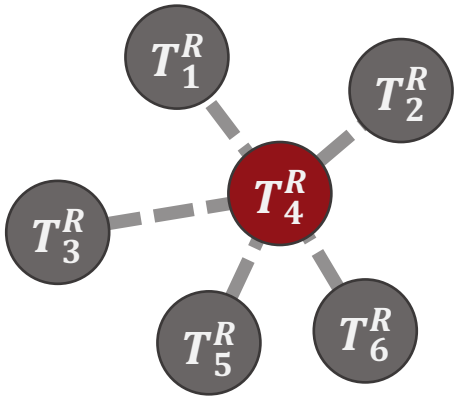


Neutral: n = 10 , Ne=100,000, θ=5

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

Some summary statistics

$$\overline{T} := \underset{T \in \{T_1^R, T_2^R, \ldots, T_s^R\}}{\arg\min} \sum_{j=1}^{s} d^2(T, T_j^R)$$ Medoid



$$\sigma^2 := \frac{1}{s} \sum_{j=1}^{s} d^2(\overline{T}, T_j^R)$$ Sample Variation

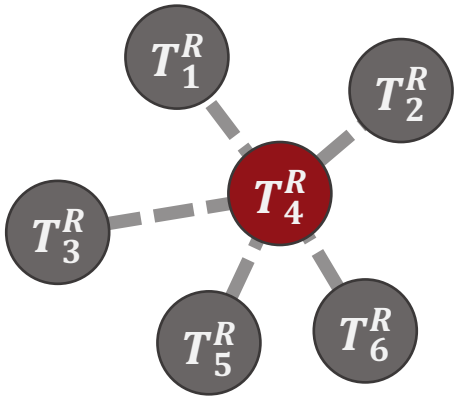
Test Statistic

$$C^{x,y} = \frac{1}{2(N-1)} \sum_{j=1}^{N} \left[1_{d(X_j, \bar{Y}) \leq d(X_j, \bar{X})} + 1_{d(Y_j, \bar{X}) \leq d(Y_j, \bar{Y})} \right]$$

$$P = \frac{1 + \sum_{k=1}^{N_{\text{perm}}} 1_{C_k^{x,y} \leq C^{x,y}}}{1 + N_{\text{perm}}}$$

Some summary statistics

$$\overline{T} := \underset{T \in \{T_1^R, T_2^R, \ldots, T_s^R\}}{\arg\min} \sum_{j=1}^{s} d^2(T, T_j^R)$$ Medoid



$$\sigma^2 := \frac{1}{s} \sum_{j=1}^{s} d^2(\overline{T}, T_j^R)$$ Sample Variation

Challenges

- Medoids may not be unique and computationally expensive.
- Permutation test is computationally expensive.
- Unknown statistical properties of statistics.

Finding a representative tree

Current approaches for labeled trees:

- Majority-rule consensus tree (MRC)
- Maximum clade credibility (MCC)
- Median tree based on metrics on labeled trees

Current approaches for unlabeled ranked trees:

Finding a representative tree

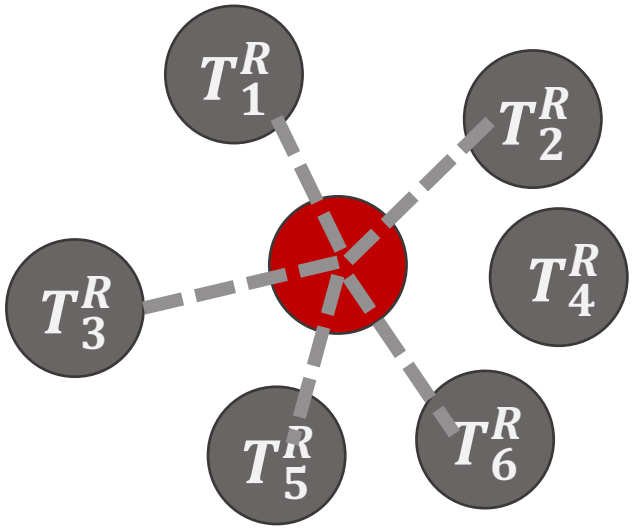
The tree shape that minimizes the expected squared distance

$$\bar{T} \in \underset{x \in \mathcal{T}_n}{\operatorname{argmin}} \sum_{y \in \mathcal{T}_n} d(x,y)^2 \mu(y).$$

The genealogy that minimizes the expected squared distance

$$\bar{G} \in \underset{G \in \mathcal{G}_n}{\operatorname{argmin}} \int_{H \in \mathcal{G}_n} d(G,H)^2 d\nu(H).$$

Fréchet Mean

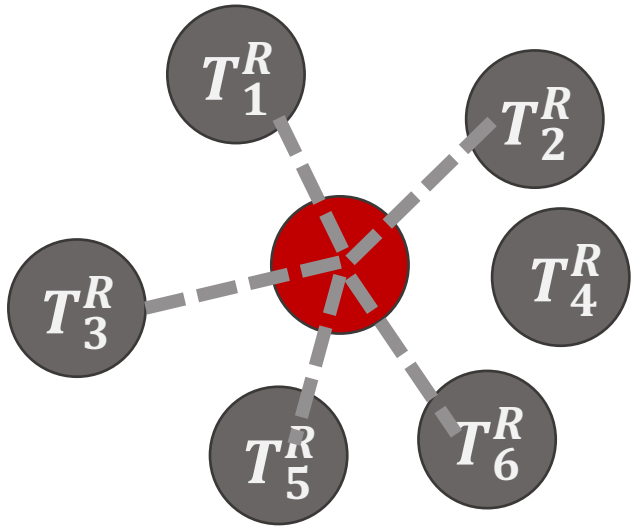


Fréchet mean genealogy

Fréchet Mean

In particular for d_2

$$\bar{G} \in \operatorname*{argmin}_{G \in \mathcal{G}_n} \int_{H \in \mathcal{G}_n} d(G,H)^2 d\nu(H).$$



Proposition 2. Let ν be a probability measure on $\mathcal{G}_n$, the space of isochronous genealogies, such that the tree topology and branching event times are independent under ν. The Fréchet mean $\bar{G}_2 = (F^*, u^*)$ under the d_2 metric can be obtained by separately finding F^* and u^*.

Finding the Fréchet mean as MIP

In particular, using d_2

$$\bar{F}_2 \in \underset{F \in \mathcal{F}_n}{\operatorname{argmin}} \sum_{H \in \mathcal{F}_n} \sum_{k,l} (F_{kl} - H_{kl})^2 \mu(H)$$

$$= \underset{F \in \mathcal{F}_n}{\operatorname{argmin}} \sum_{H \in \mathcal{F}_n} \sum_{k,l} (F_{kl}^2 - 2F_{kl}H_{kl}) \mu(H)$$

$$= \underset{F \in \mathcal{F}_n}{\operatorname{argmin}} \sum_{k,l} \{F_{kl}^2 - 2F_{kl}M_{kl}\}$$

Where:

$$M_{kl} = \sum_{H \in \mathcal{F}_n} H_{kl} \cdot \mu(H)$$

Optimization with Gurobi

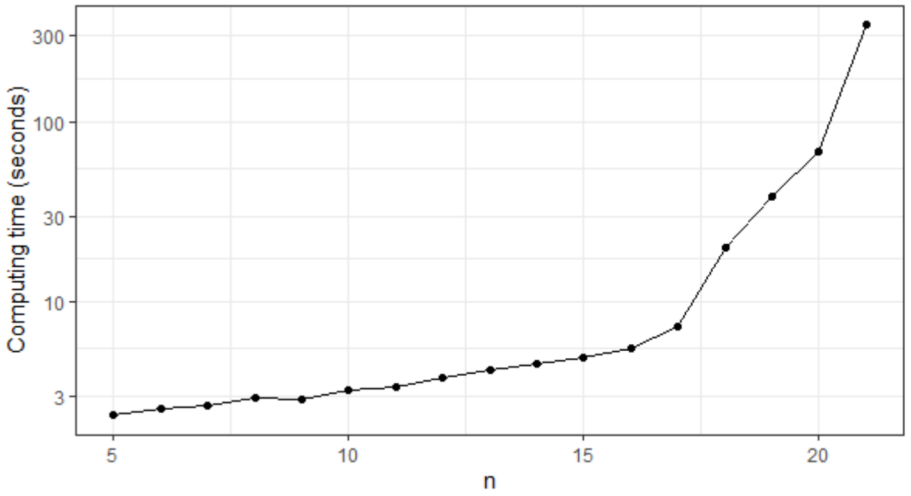


Figure 3: Running time for exact computation for Fréchet mean using Gurobi, plotted against dimension of $\mathbf{F}$-matrices. $B = 1000$ trees were generated for each n. Computations done on personal laptop, with Intel i7 processor.

Cardinality of search space

$$|\mathcal{T}|_n \sim 2\,(2/\pi)^{n+1} \cdot n!$$

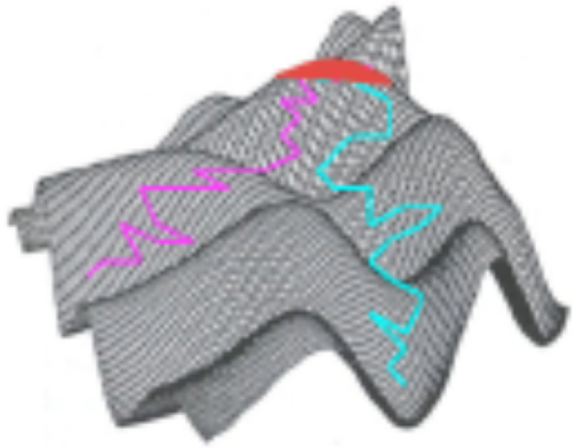
Finding the Fréchet mean with combinatorial optimization

Simulated Annealing:

Minimize $E(x) = \sum_{i=1}^{m} d(x, y_i)^2$

Maximize $\exp\{-E(x)/R\}$

at any given temperature $R > 0$



In particular, at the temperature schedule:

$R_k = \alpha^k R_0$ for some high initial temperature R_0.

Construct a **Markov chain** with symmetric proposal distribution and MH acceptance:

$$a_k = \exp\left(-\frac{E(x_k)}{R_k} + \frac{E(x_{k-1})}{R_{k-1}}\right) \wedge 1.$$

Other summaries

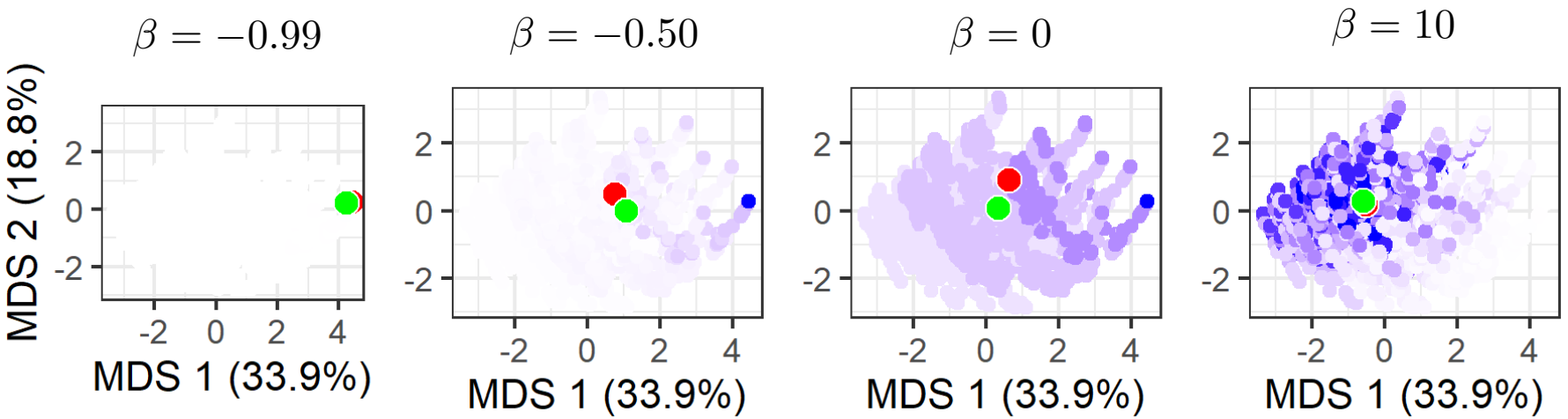
Population Fréchet Variance

$$V = \sum_{y \in \mathcal{T}_n} d(y, \bar{T})^2 \cdot \mu(y), \quad \text{where} \quad \bar{T} = \operatorname*{argmin}_{x \in \mathcal{T}_n} \sum_{y \in \mathcal{T}_n} d(x, y)^2 \cdot \mu(y)$$

Sample Fréchet Variance

$$V_m = \frac{1}{m} \sum_{i=1}^{m} d(y_i, \bar{T})^2, \quad \text{where} \quad \bar{T} = \operatorname*{argmin}_{x \in \mathcal{T}_n} \sum_{i=1}^{m} d(x, y_i)^2$$

Results on Blum-François Beta-splitting model



- Expected value
- Fréchet mean

SARS-CoV-2

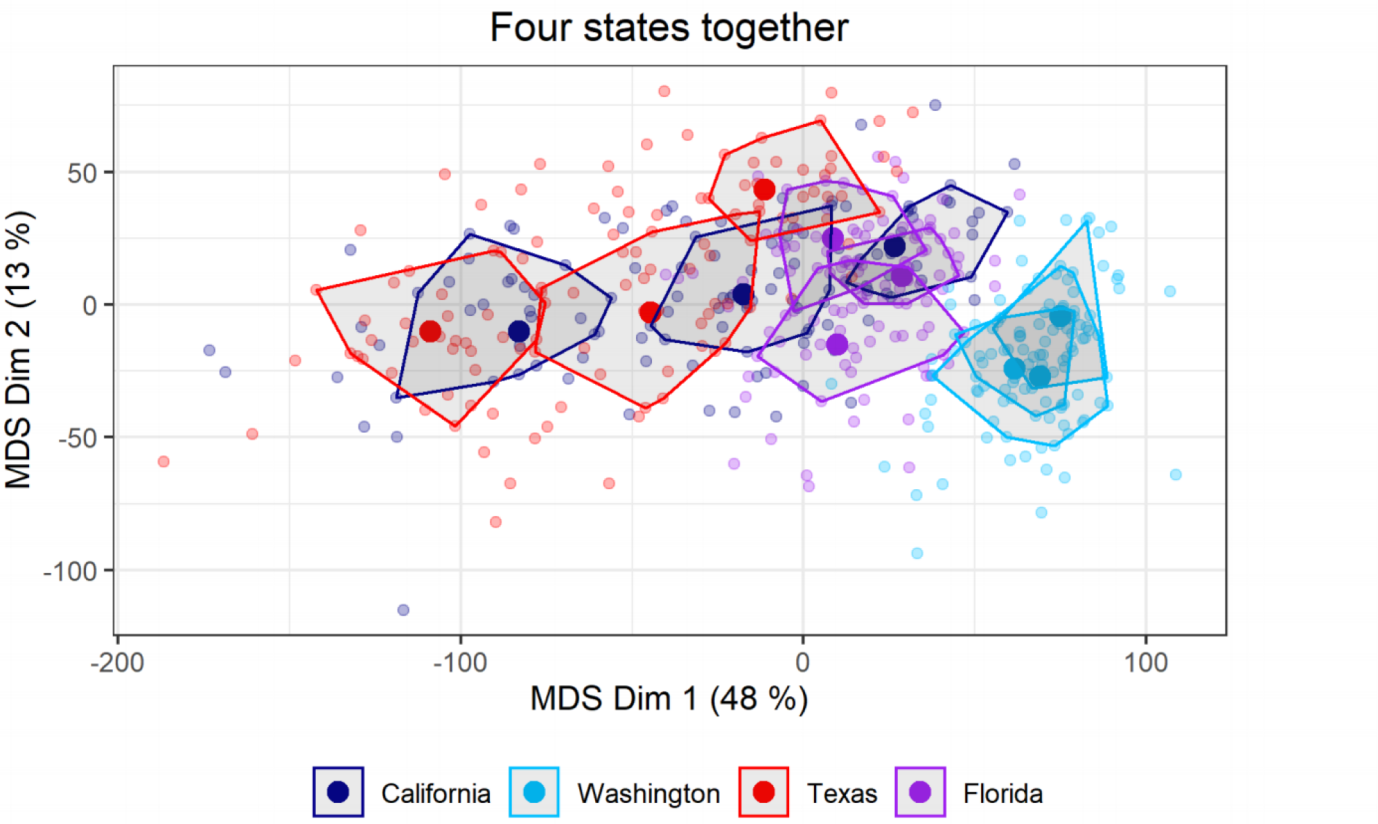


Figure 13: **MDS plot of multiple samples from California, Washington, Florida, and Texas.** Three samples of 20 trees of $n = 100$ samples randomly chosen among GISAID sequences in Feb-May 2020 per location. The Fréchet means are calculated using average coalescent times and marked as red dots The shaded region corresponds to 50% credible convex hulls around the Fréchet means.

Tutorial and Reference

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

Distance metrics for ranked evolutionary trees.

Kim J, Rosenberg NA, Palacios JA

PNAS (2020) in press.

- https://github.com/RSamyak/fmatrix/blob/main/README.md

Statistical summaries of unlabelled evolutionary trees and ranked hierarchical clustering trees.

Samyak R, Palacios JA

arXiv:2106.02724.