State-space exploration of Tajima Trees

Julia A. Palacios
Joint with A. Véber, J. Wakeley and S. Ramachandran

Department of Statistics
Department of Biomedical Data Science
Stanford University

CIRM
July 2018

Motivation: Estimation of Effective Population Size

Effective Population Size Trajectory $Ne(t)$

$N_e(t)$ is a measure of relative genetic diversity over time.

Motivation: Estimation of Effective Population Size

Effective Population Size Trajectory $Ne(t)$

$N_e(t)$ is a measure of relative genetic diversity over time.

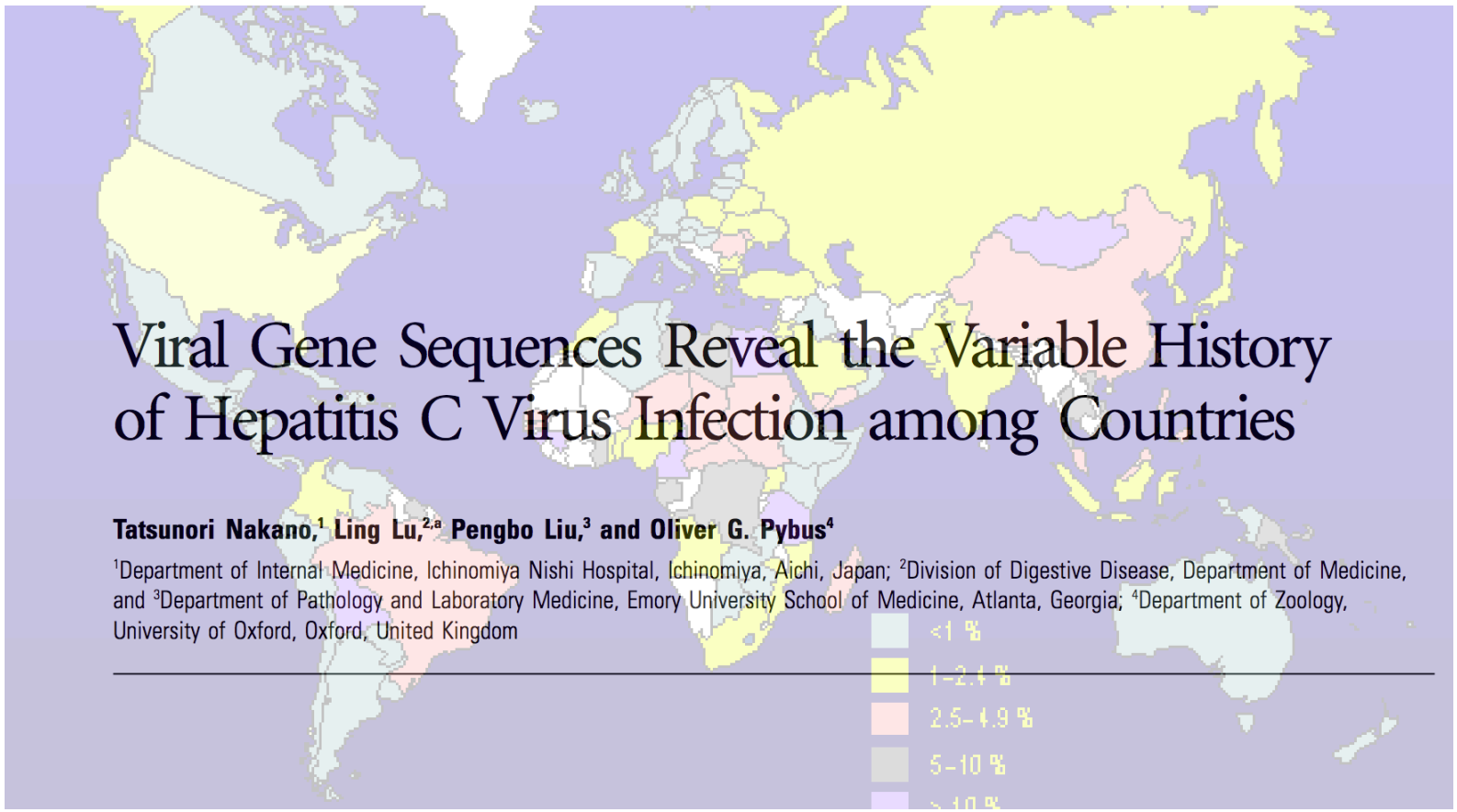
Why is it important?

Motivation: Estimation of Effective Population Size

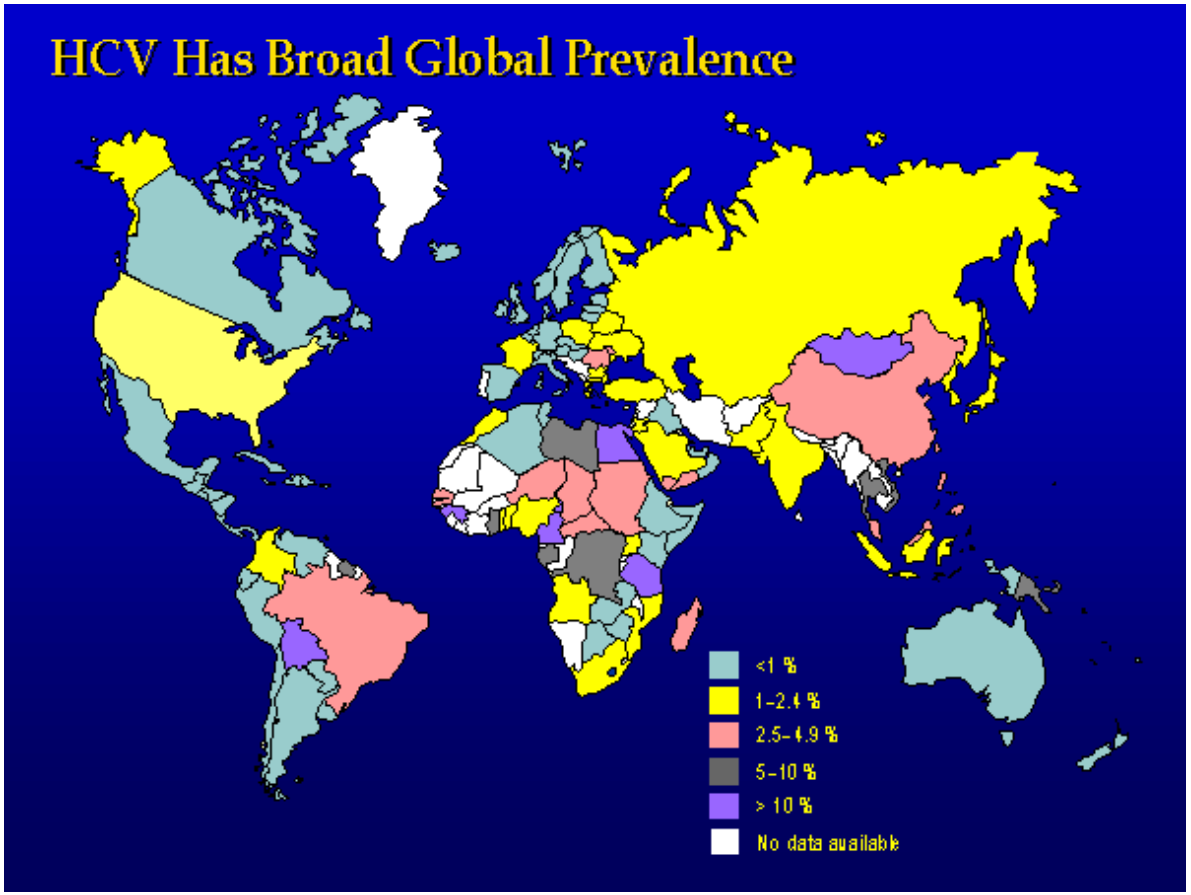
Effective Population Size Trajectory $Ne(t)$

$N_e(t)$ is a measure of relative genetic diversity over time.

Why is it important?



Example 1: Hepatitis C Virus



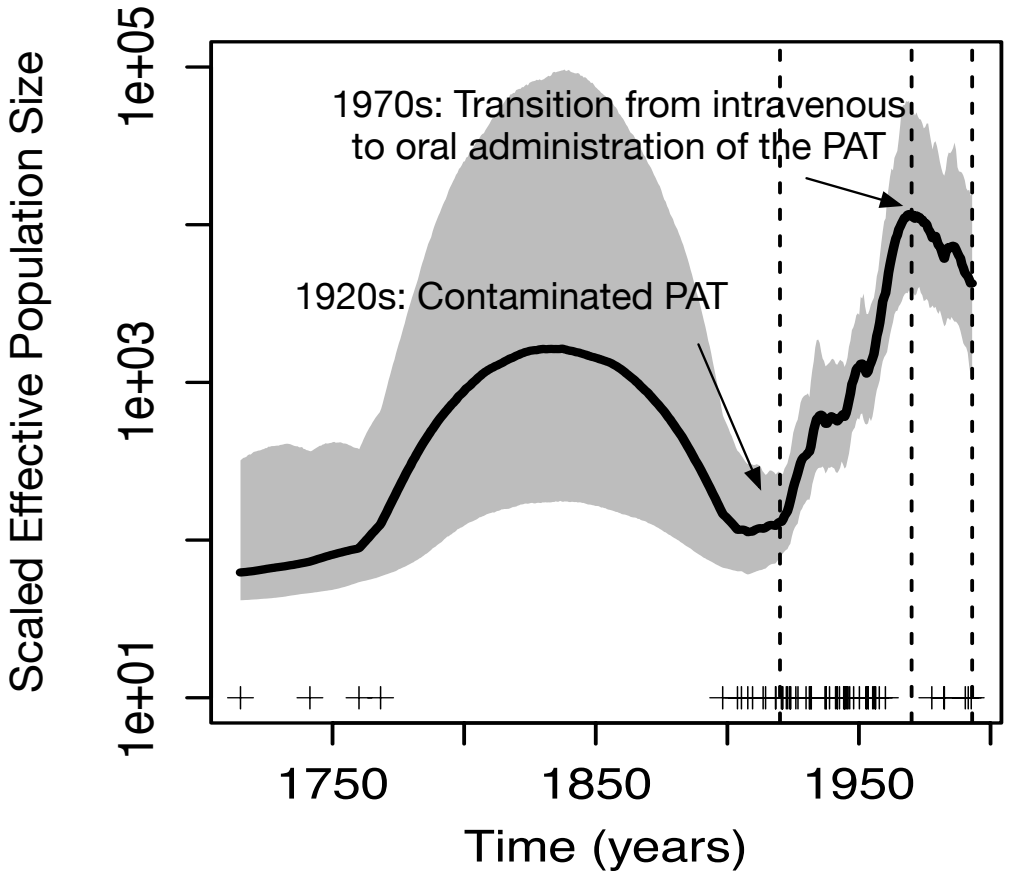
Prevalence of HCV - WHO 1999

- Identified in 1989
- Spread by blood to blood contact
- $\approx$3% of infected population worldwide
- 8,000 - 10,000 deaths per year in the USA
- Egypt has the highest prevalence

Example 1: Hepatitis C Virus

- 62 samples in 1993 from the E1 gene (411bp)
- Parenteral antischistosomal therapy (PAT) was practiced from 1920s to 1980s
- In the 1970s started a transition from the intravenous to the oral administration of the PAT

Example 1: Hepatitis C Virus

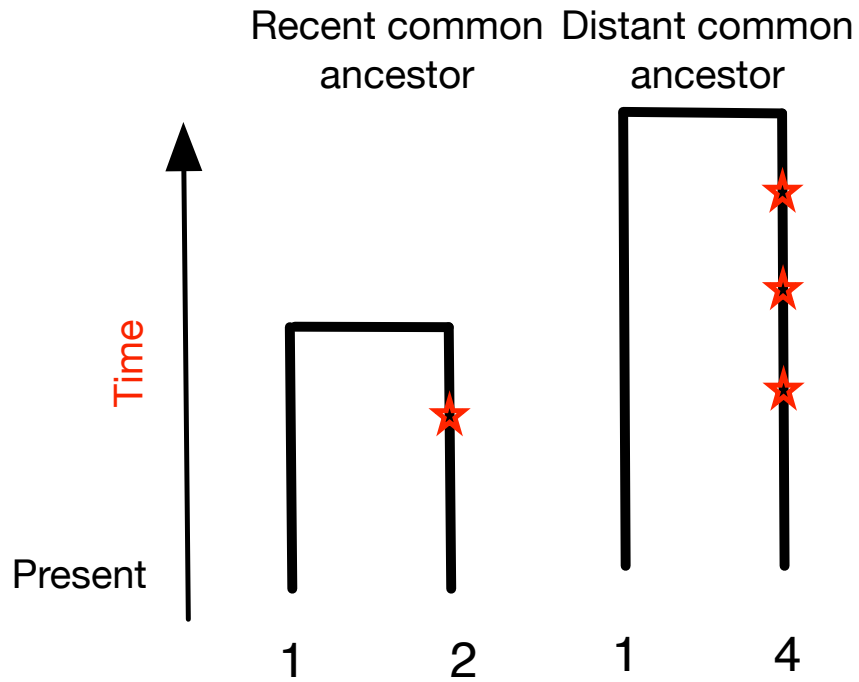


[Palacios and Minin, Biometrics 2013]

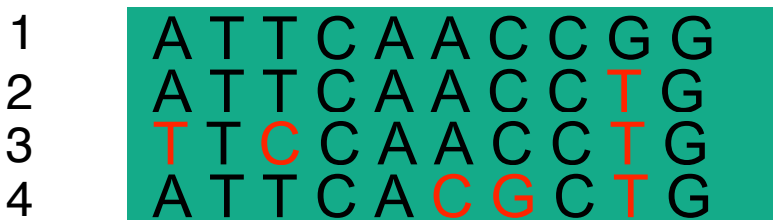
- 62 samples in 1993 from the E1 gene (411bp)
- Parenteral antischistosomal therapy (PAT) was practiced from 1920s to 1980s
- In the 1970s started a transition from the intravenous to the oral administration of the PAT

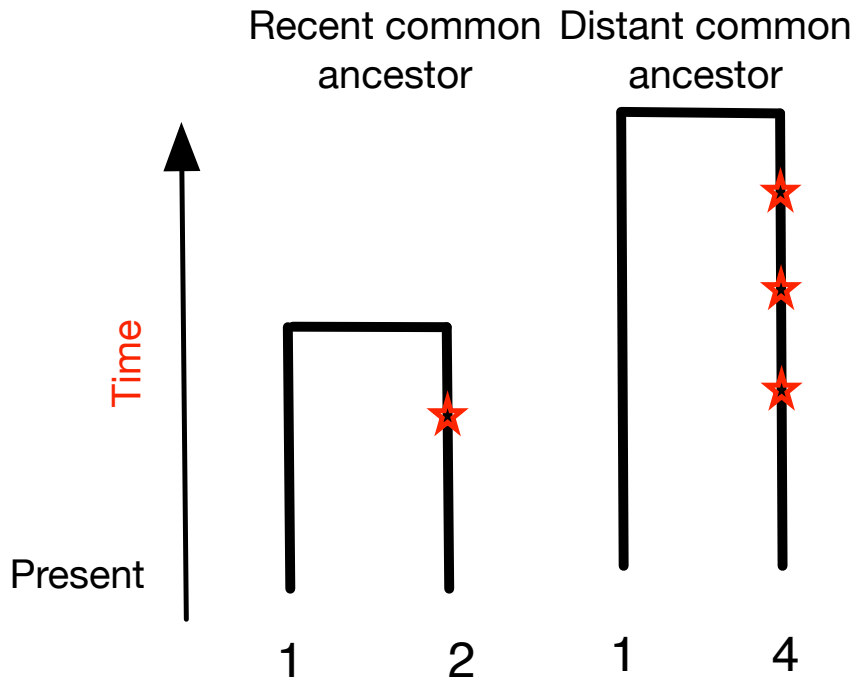
Present-day DNA data inform us about the past

1	A T T C A A C C G G
2	A T T C A A C C T G
3	T T C C A A C C T G
4	A T T C A C G C T G



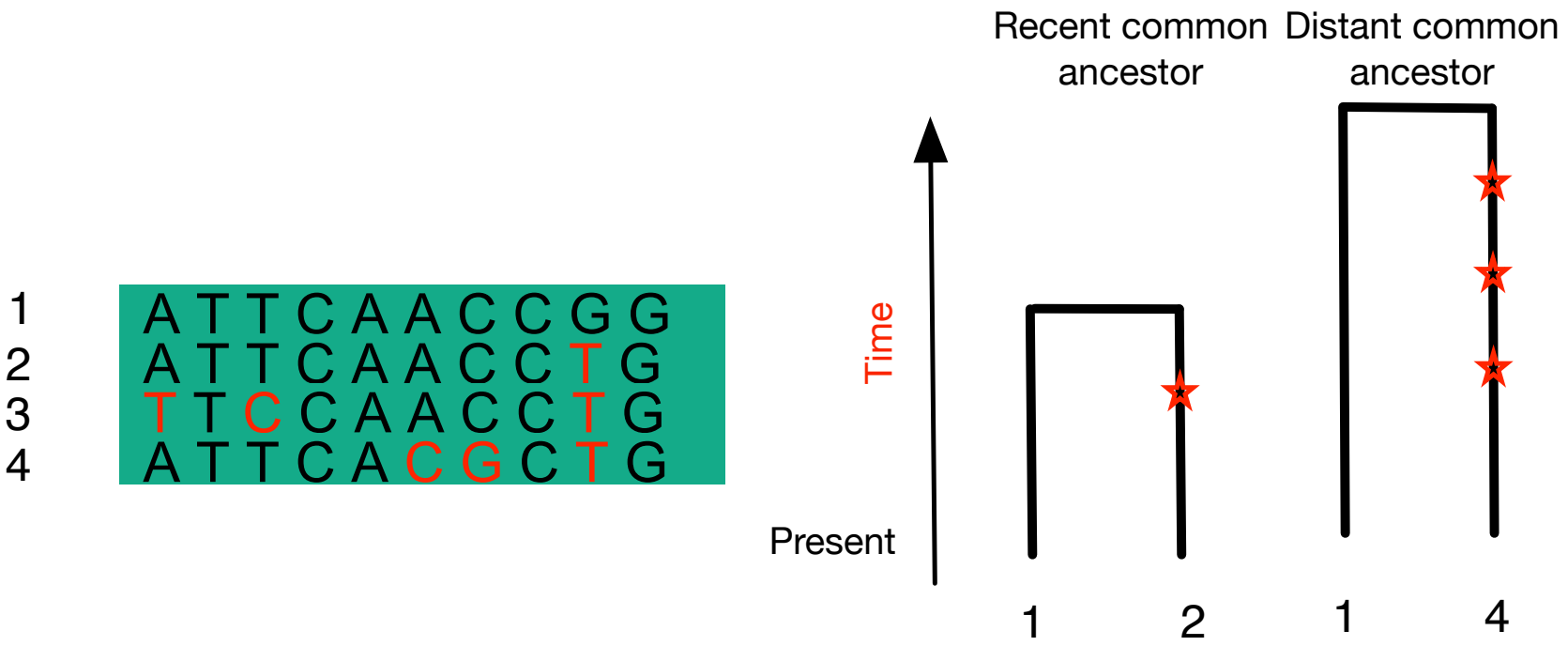
Present-day DNA data inform us about the past





- In humans, mutation rate is estimated to be $\approx 10^{-8}$ per base per generation.

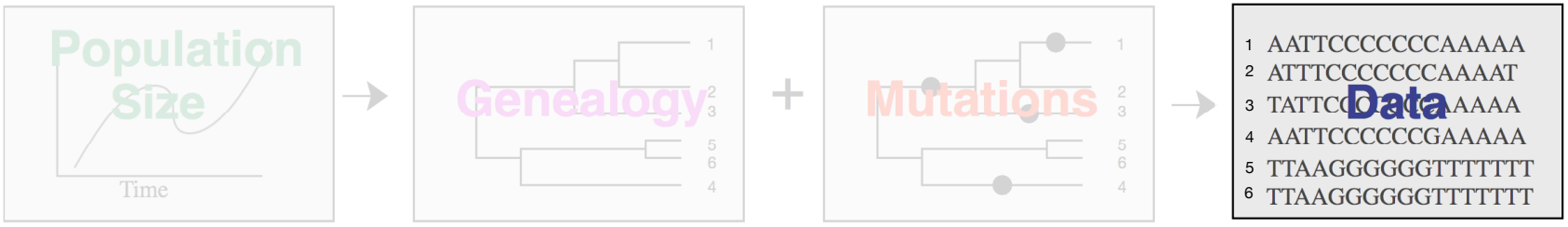
Present-day DNA data inform us about the past



- In humans, mutation rate is estimated to be $\approx 10^{-8}$ per base per generation.
- Recent ancestry indicates small population sizes.

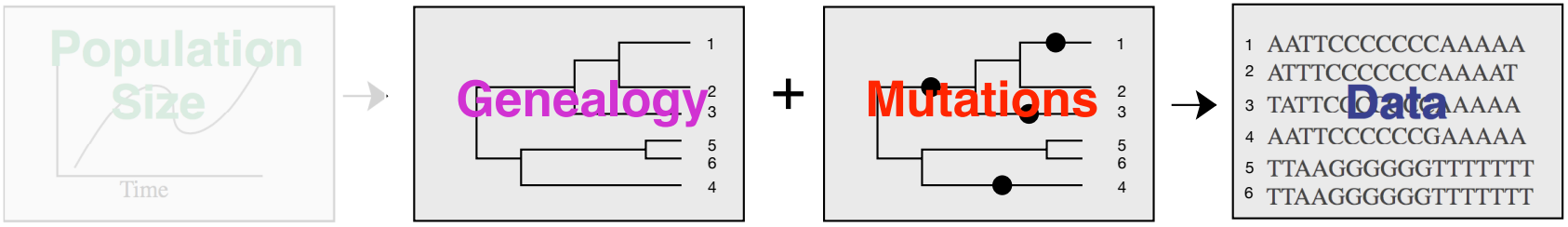
Goal: Estimation of Effective Population Size

Coalescent-based model



Goal: Estimation of Effective Population Size

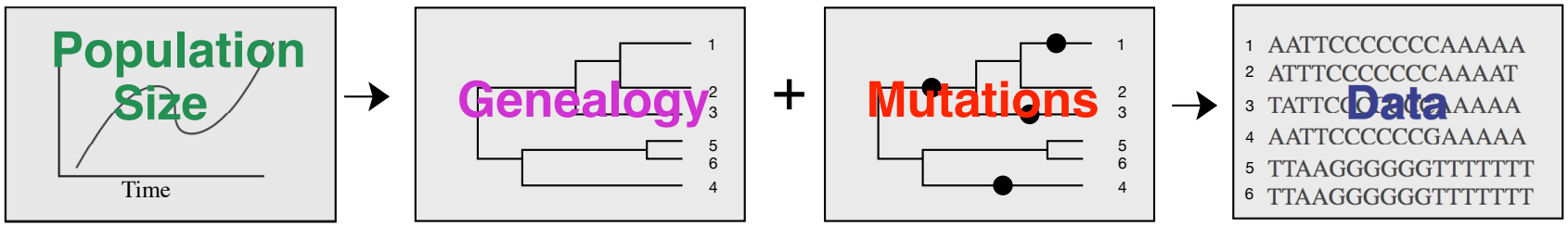
Coalescent-based model



- Ancestral process: coalescent process of genealogies
- Mutation process: Poisson process along the branches of the genealogy

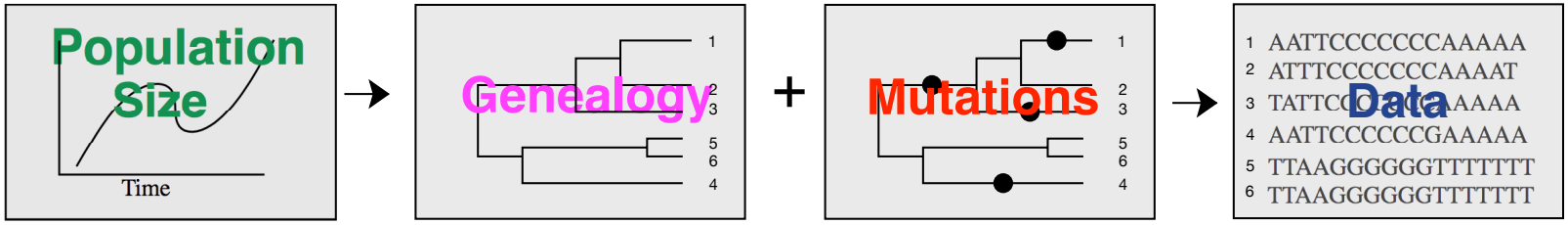
Goal: Estimation of Effective Population Size

Coalescent-based model



- Ancestral process: coalescent process of genealogies.
- Mutation process: Poisson process along the branches of the genealogy.
- Population process: Effective population size trajectory over time.

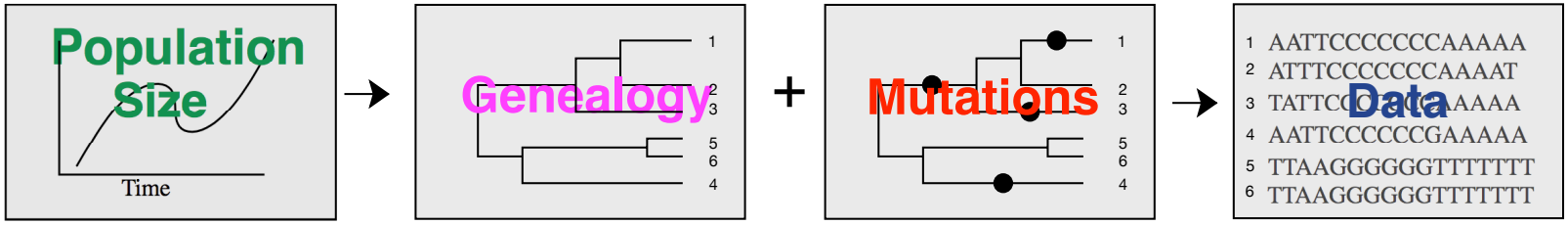
Challenging inference from data



$N_e(t)$ $\mathbf{G}$ $\mathbf{Q}$ $\mathbf{Y}$

$$P\left(N_e(t), \mathbf{G}, \mathbf{Q}, \tau \mid \mathbf{Y}\right) \propto \underbrace{P\left(\mathbf{Y} \mid \mathbf{G}, \mathbf{Q}\right)}_{\text{Likelihood}} \underbrace{P\left(\mathbf{G} \mid N_e(t)\right)}_{\text{Coalescent prior}} P\left(\mathbf{Q}\right) \underbrace{P\left(N_e(t) \mid \tau\right)}_{\log \text{GP}(0, \mathbf{C}(\tau))} P\left(\tau\right)$$

Challenging inference from data

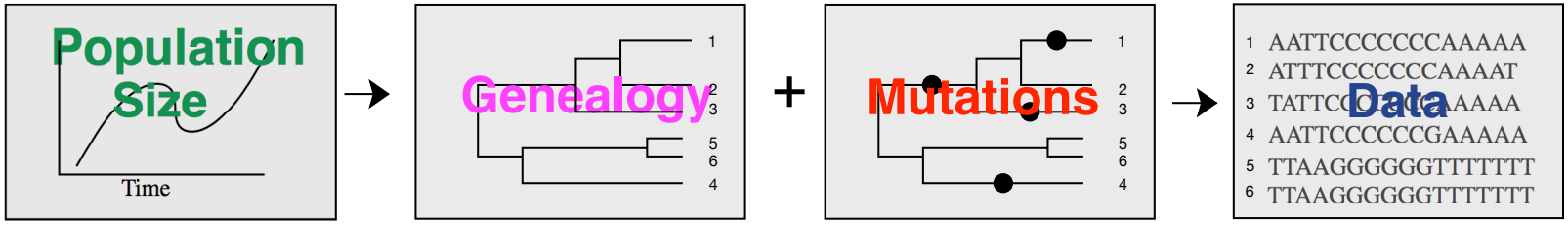


$N_e(t)$ $\quad$ $\mathbf{G}$ $\quad$ $\mathbf{Q}$ $\quad$ $\mathbf{Y}$

$$P\left(N_e(t), \mathbf{G}, \mathbf{Q}, \tau \mid \mathbf{Y}\right) \propto \underbrace{P\left(\mathbf{Y} \mid \mathbf{G}, \mathbf{Q}\right)}_{\text{Likelihood}} \underbrace{P\left(\mathbf{G} \mid N_e(t)\right)}_{\text{Coalescent prior}} P\left(\mathbf{Q}\right) \underbrace{P\left(N_e(t) \mid \tau\right)}_{\log \text{GP}(0, \mathbf{C}(\tau))} P\left(\tau\right)$$

- The likelihood $P\left(\mathbf{Y} \mid \mathbf{G}, \mathbf{Q}\right)$ is tractable.

Challenging inference from data



$N_e(t)$ $\quad$ $\mathbf{G}$ $\quad$ $\mathbf{Q}$ $\quad$ $\mathbf{Y}$

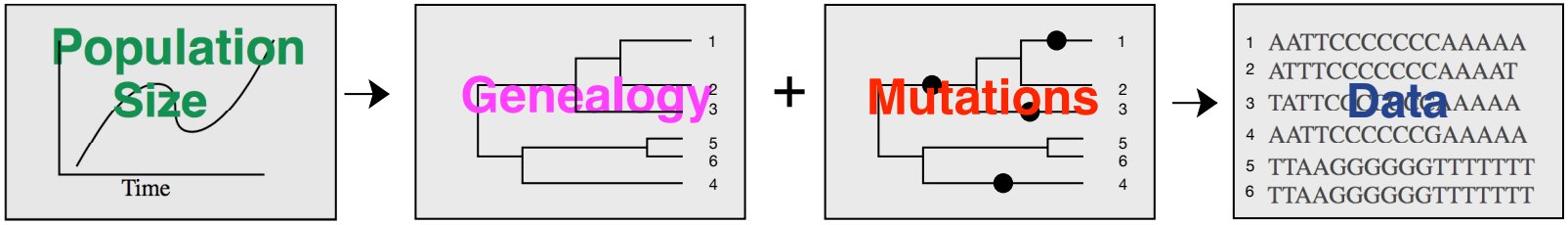
$$P\left(N_e(t), \mathbf{G}, \mathbf{Q}, \tau \mid \mathbf{Y}\right) \propto \underbrace{P\left(\mathbf{Y} \mid \mathbf{G}, \mathbf{Q}\right)}_{\text{Likelihood}} \underbrace{P\left(\mathbf{G} \mid N_e(t)\right)}_{\text{Coalescent prior}} P\left(\mathbf{Q}\right) \underbrace{P\left(N_e(t) \mid \tau\right)}_{\log \text{GP}(0, \mathbf{C}(\tau))} P\left(\tau\right)$$

- The likelihood $P\left(\mathbf{Y} \mid \mathbf{G}, \mathbf{Q}\right)$ is tractable.

The state space of genealogies $\mathcal{G}$

- $\mathcal{G} = \mathcal{T}_n \otimes \mathbb{R}^{+n-1}$

Challenging inference from data



$N_e(t)$ $\quad$ $\mathbf{G}$ $\quad$ $\mathbf{Q}$ $\quad$ $\mathbf{Y}$

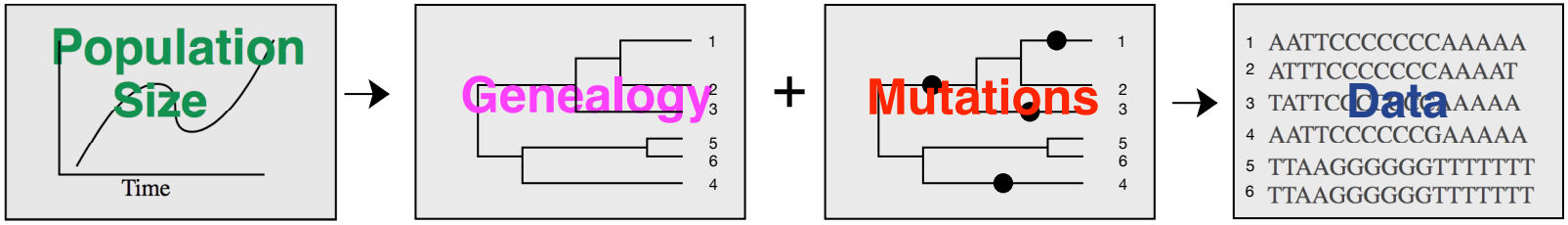
$$P\left(N_e(t), \mathbf{G}, \mathbf{Q}, \tau \mid \mathbf{Y}\right) \propto \underbrace{P\left(\mathbf{Y} \mid \mathbf{G}, \mathbf{Q}\right)}_{\text{Likelihood}} \underbrace{P\left(\mathbf{G} \mid N_e(t)\right)}_{\text{Coalescent prior}} P\left(\mathbf{Q}\right) \underbrace{P\left(N_e(t) \mid \tau\right)}_{\log \mathrm{GP}(0, \mathbf{C}(\tau))} P\left(\tau\right)$$

- The likelihood $P\left(\mathbf{Y} \mid \mathbf{G}, \mathbf{Q}\right)$ is tractable.

The state space of genealogies $\mathcal{G}$

- $\mathcal{G} = \mathcal{T}_n \otimes \mathbb{R}^{+n-1}$
- $|\mathcal{T}_n| = n!(n-1)!/2^{n-1}$
- $|\mathcal{T}_{100}| \approx 10^{284}$

Challenging inference from data



$N_e(t)$ $\quad$ $\mathbf{G}$ $\quad$ $\mathbf{Q}$ $\quad$ $\mathbf{Y}$

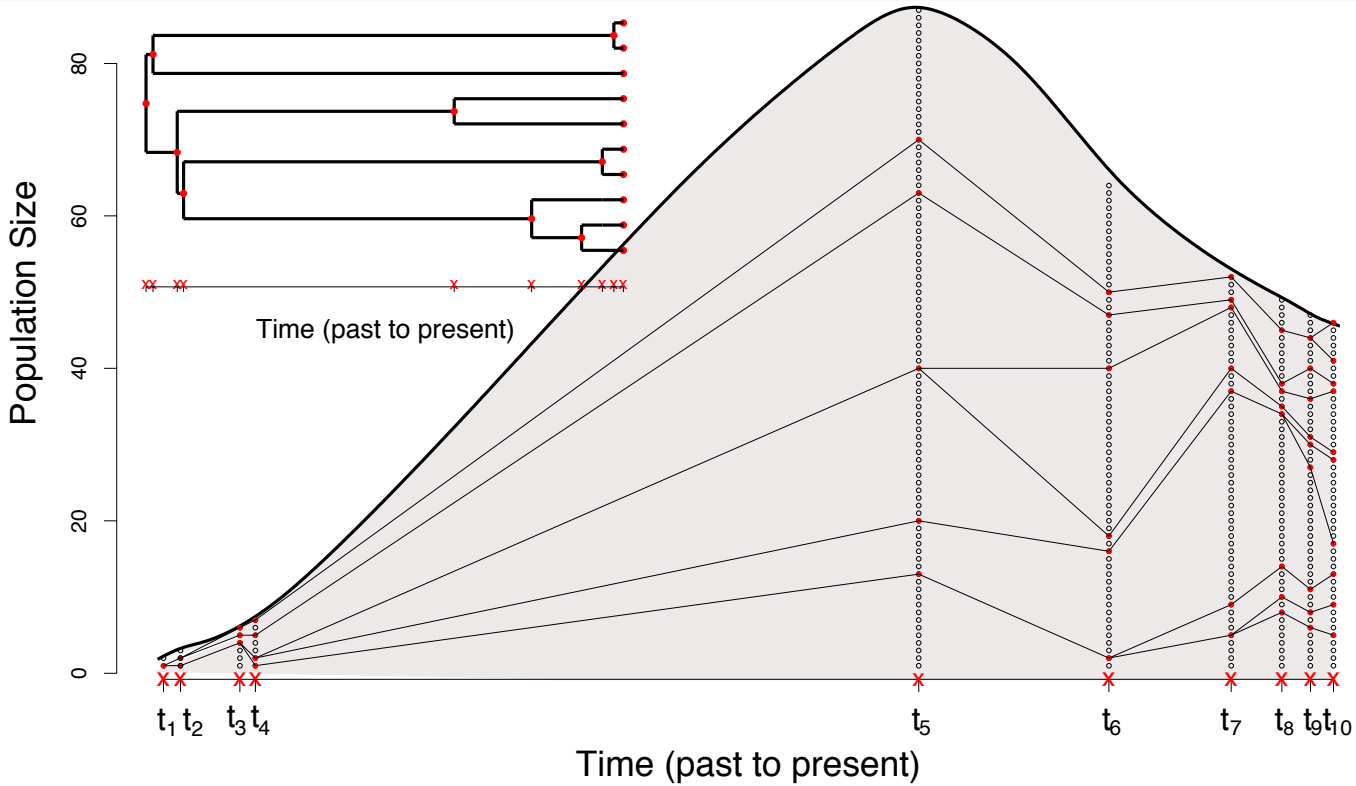
$$P\left(N_e(t), \mathbf{G}, \mathbf{Q}, \tau \mid \mathbf{Y}\right) \propto \underbrace{P\left(\mathbf{Y} \mid \mathbf{G}, \mathbf{Q}\right)}_{\text{Likelihood}} \underbrace{P\left(\mathbf{G} \mid N_e(t)\right)}_{\text{Coalescent prior}} P\left(\mathbf{Q}\right) \underbrace{P\left(N_e(t) \mid \tau\right)}_{\log \mathrm{GP}(0, \mathbf{C}(\tau))} P\left(\tau\right)$$

- The likelihood $P\left(\mathbf{Y} \mid \mathbf{G}, \mathbf{Q}\right)$ is tractable.

The state space of genealogies $\mathcal{G}$

- $\mathcal{G} = \mathcal{T}_n \otimes \mathbb{R}^{+n-1}$
- $|\mathcal{T}_n| = n!(n-1)!/2^{n-1}$
- $|\mathcal{T}_{100}| \approx 10^{284}$
- $\approx 10^{80}$ atoms in the universe
- $\approx 4.4 \times 10^{17}$ seconds since the Big Bang

Coalescent times alone are sufficient statistics for $N(t)$

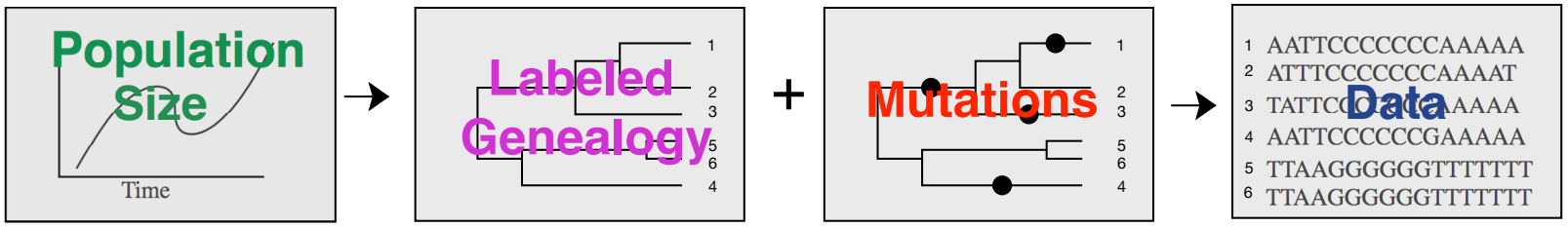


Coalescent Density:

$$P\left(\mathbf{G}|N_e(t)\right) \propto \prod_{k=2}^{n} P[t_{k-1}|t_k, N_e(t)].$$

$\mathbf{t} = (t_2, t_3, \ldots, t_n)$ are sufficient statistics for inferring $N_e(t)$

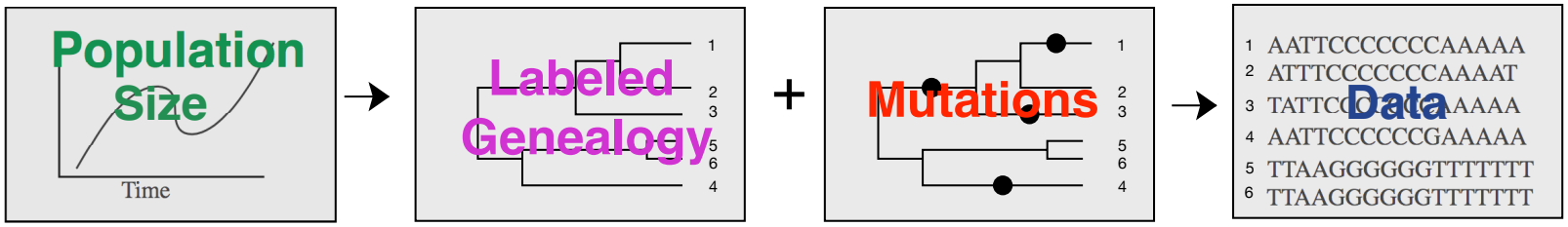
An alternative coalescent model?



What if we replace $\mathbf{G}$ with $\mathbf{t}$, the vector of coalescent times?
Posteriors:

$$P\left(N_e(t), \mathbf{G}, \mathbf{Q}, \tau \mid \mathbf{Y}\right) \text{ vs } P\left(N_e(t), \mathbf{t}, \mathbf{Q}, \tau \mid \mathbf{Y}\right)$$

An alternative coalescent model?



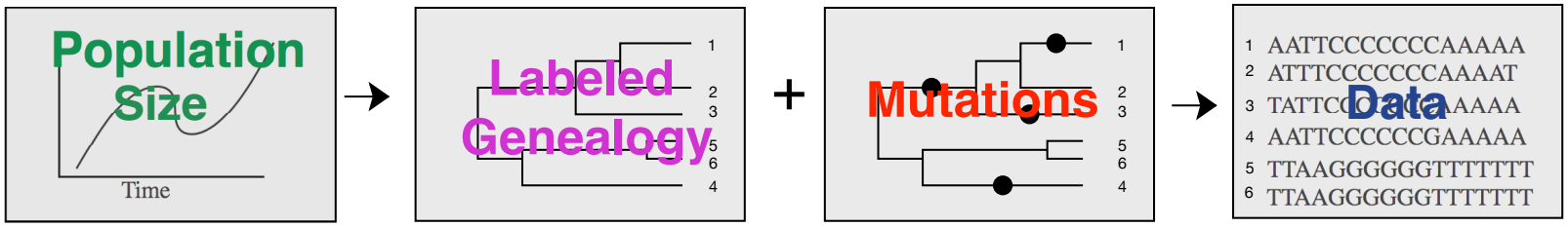
What if we replace $\mathbf{G}$ with $\mathbf{t}$, the vector of coalescent times?
Posteriors:

$$P\left(N_e(t), \mathbf{G}, \mathbf{Q}, \boldsymbol{\tau} \mid \mathbf{Y}\right) \text{ vs } P\left(N_e(t), \mathbf{t}, \mathbf{Q}, \boldsymbol{\tau} \mid \mathbf{Y}\right)$$

Using only coalescent times $\mathbf{t}$:

$$P\left(N_e(t), \mathbf{t}, \mathbf{Q}, \boldsymbol{\tau} \mid \mathbf{Y}\right) \propto \underbrace{P\left(\mathbf{Y} \mid \mathbf{t}, \mathbf{Q}\right)}_{\text{Likelihood}} \underbrace{P\left(\mathbf{t} \mid N_e(t)\right)}_{\text{Coalescent prior}} P\left(\mathbf{Q}, N_e(t), \boldsymbol{\tau}\right)$$

An alternative coalescent model?



What if we replace $\mathbf{G}$ with $\mathbf{t}$, the vector of coalescent times?
Posteriors:

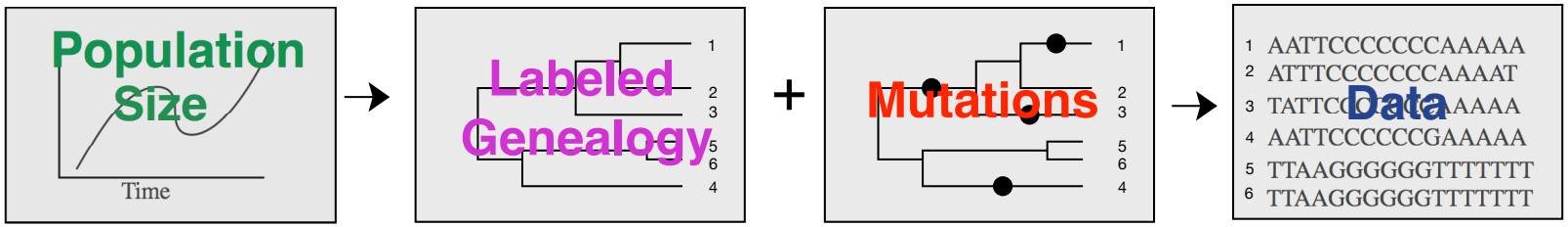
$$P\left(N_e(t), \mathbf{G}, \mathbf{Q}, \boldsymbol{\tau} \mid \mathbf{Y}\right) \text{ vs } P\left(N_e(t), \mathbf{t}, \mathbf{Q}, \boldsymbol{\tau} \mid \mathbf{Y}\right)$$

Using only coalescent times $\mathbf{t}$:

$$P\left(N_e(t), \mathbf{t}, \mathbf{Q}, \boldsymbol{\tau} \mid \mathbf{Y}\right) \propto \underbrace{P\left(\mathbf{Y} \mid \mathbf{t}, \mathbf{Q}\right)}_{\text{Likelihood}} \underbrace{P\left(\mathbf{t} \mid N_e(t)\right)}_{\text{Coalescent prior}} P\left(\mathbf{Q}, N_e(t), \boldsymbol{\tau}\right)$$

- $P\left(\mathbf{Y} \mid \mathbf{t}, \mathbf{Q}\right) = \sum_{\mathcal{T}} P\left(\mathbf{Y}, \mathcal{T} \mid \mathbf{t}, \mathbf{Q}\right)$ is not “practical" since we need to sum over all possible topologies.

An alternative coalescent model?



What if we replace $\mathbf{G}$ with $\mathbf{t}$, the vector of coalescent times?
Posteriors:

$$P\left(N_e(t), \mathbf{G}, \mathbf{Q}, \boldsymbol{\tau} \mid \mathbf{Y}\right) \text{ vs } P\left(N_e(t), \mathbf{t}, \mathbf{Q}, \boldsymbol{\tau} \mid \mathbf{Y}\right)$$

Using only coalescent times $\mathbf{t}$:

$$P\left(N_e(t), \mathbf{t}, \mathbf{Q}, \boldsymbol{\tau} \mid \mathbf{Y}\right) \propto \underbrace{P\left(\mathbf{Y} \mid \mathbf{t}, \mathbf{Q}\right)}_{\text{Likelihood}} \underbrace{P\left(\mathbf{t} \mid N_e(t)\right)}_{\text{Coalescent prior}} P\left(\mathbf{Q}, N_e(t), \boldsymbol{\tau}\right)$$

- $P\left(\mathbf{Y} \mid \mathbf{t}, \mathbf{Q}\right) = \sum_{\mathcal{T}} P\left(\mathbf{Y}, \mathcal{T} \mid \mathbf{t}, \mathbf{Q}\right)$ is not “practical" since we need to sum over all possible topologies.
- The times alone $\mathbf{t}$ are not sufficient for inferring $N_e(t)$ for whole genomes [Palacios et al., 2015]

An alternative coalescent model?

Different resolutions of the coalescent:

Finding the best resolution for the Kingman-Tajima coalescent: theory and applications

R. Sainudiin · T. Stadler · A. Véber

[J. Math. Biol, 2015]
“ considering the optimal resolution with respect to a given statistic can (i) lead to significant computational savings in terms of time complexity by directly sampling from a much smaller hidden space and (ii) help generate samples from the conditional hidden space (given the observed statistics) by controlling the sampling in such a way that only trees or shapes in the hidden space that are compatible with the observed statistics are drawn"

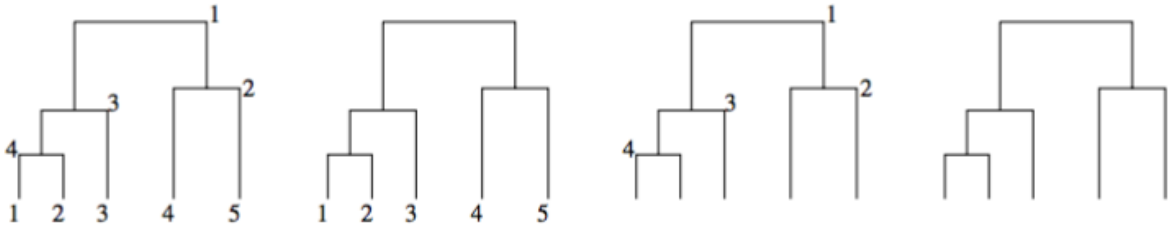
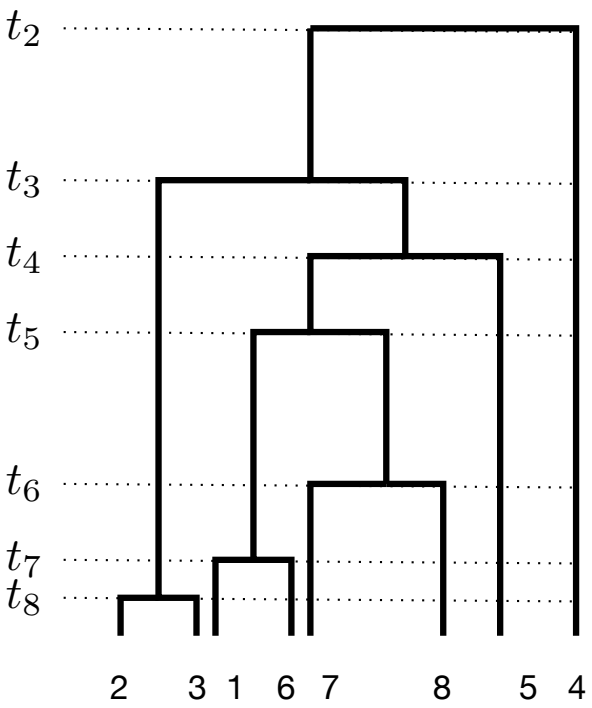


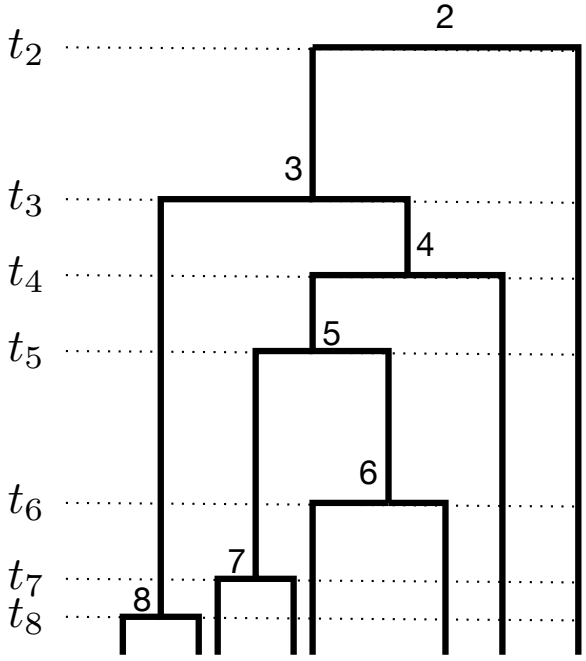
Fig. 2 Example for a ranked labeled tree with leaf label set $\mathfrak{L} = \{1,2,3,4,5\}$, a labeled tree with $\mathfrak{L} = \{1,2,3,4,5\}$, a ranked tree shape and a tree shape (from left to right).

Kingman's genealogies vs Tajima's genealogies

Kingman's genealogy

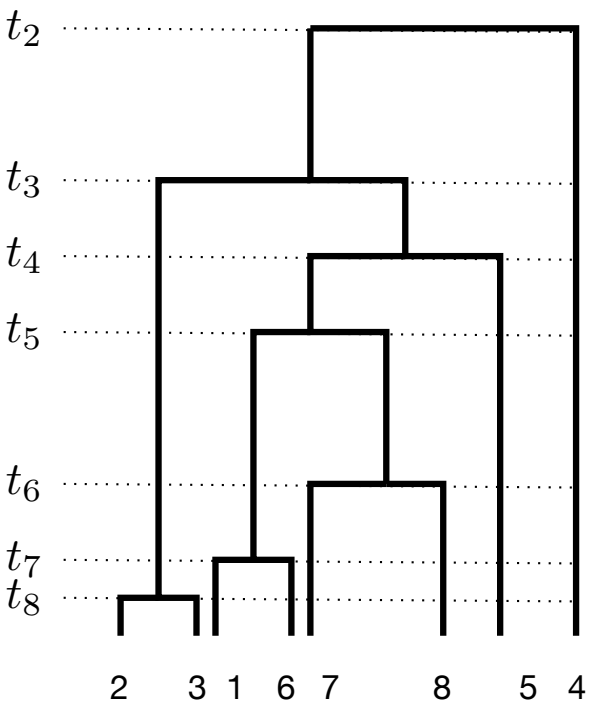


Tajima's genealogy

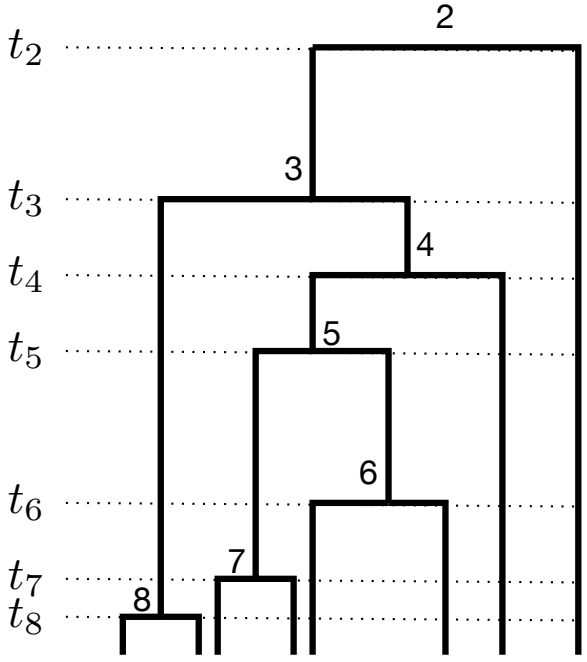


Kingman's genealogies vs Tajima's genealogies

Kingman's genealogy



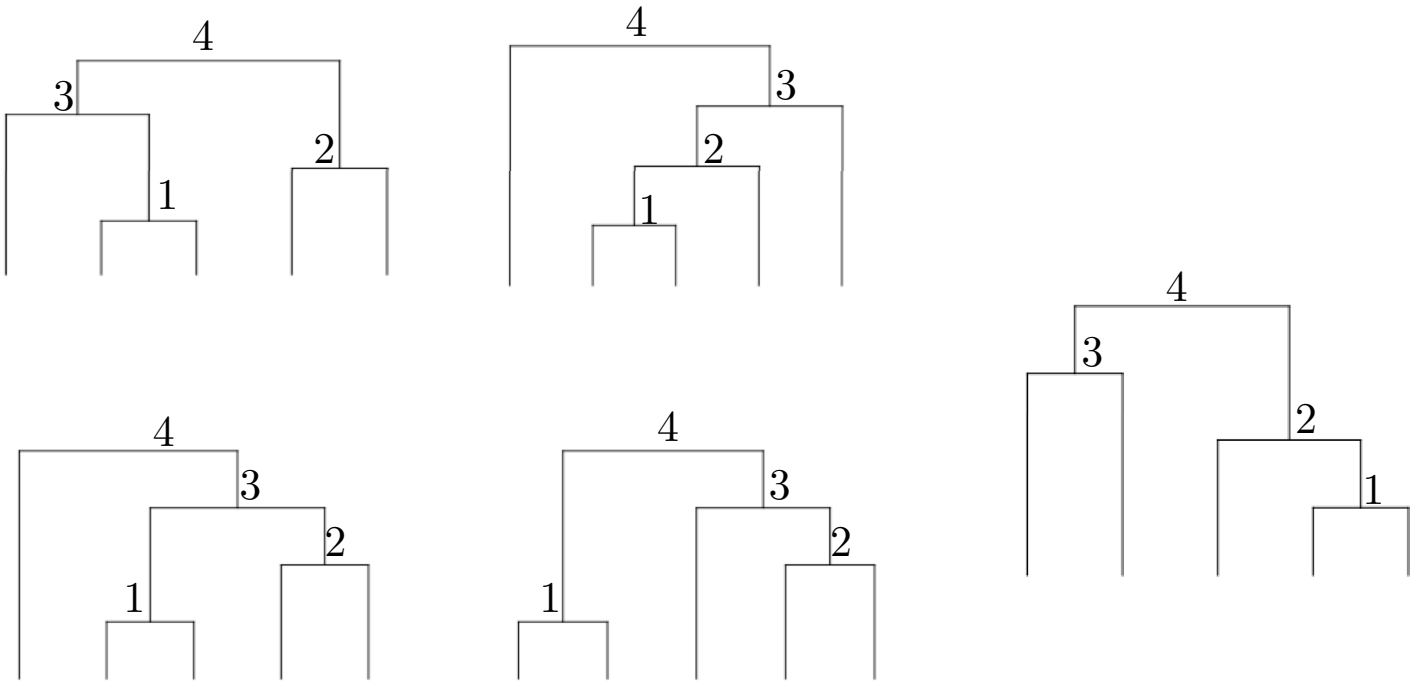
Tajima's genealogy



n	labeled topologies	ranked tree shapes
3	3	1
5	180	5
10	2.5×10^9	7936
20	5.64×10^{29}	2.9×10^{13}
50	3.28×10^{112}	1.9×10^{53}

Ranked tree shapes

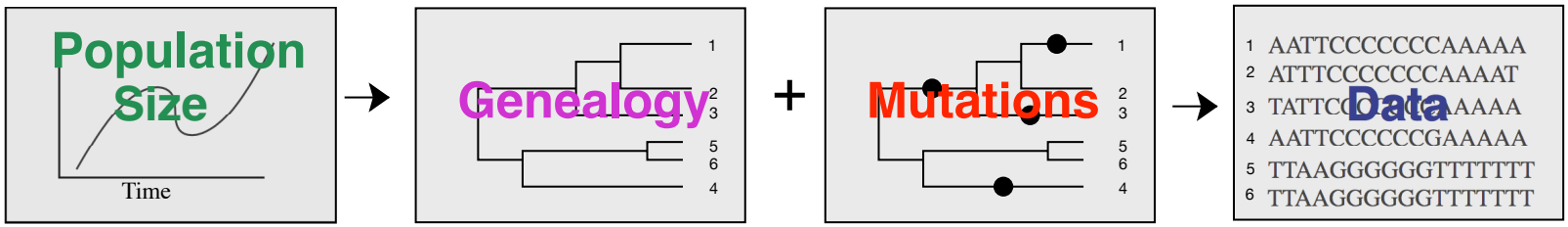
For $n = 5$



In parenthetical notation, the first tree would be represented by

$$4 : (3 : (1 : (,),),2 : (,)) \tag{1}$$

Inference with Tajima's coalescent



With Kingman's coalescent:

- Goal: $P\left(N_e(t), \mathbf{G}, \mathbf{Q}, \tau \mid \mathbf{Y}\right)$
- The state space of genealogies $\mathcal{G}$
  $\mathcal{G} = \mathcal{T}_n \otimes \mathbb{R}^{+n-1}$
- The likelihood $P\left(\mathbf{Y} \mid \mathbf{G}, \mathbf{Q}\right)$ is tractable.

With Tajima's coalescent:

- Goal: $P\left(N_e(t), \mathbf{G}^{\mathbf{T}}, \mathbf{Q}, \tau \mid \mathbf{Y}\right)$
- The state space of Tajima's genealogies $\mathcal{G}^T$
  $\mathcal{G}^T = \mathcal{R}_n \otimes \mathbb{R}^{+n-1}$
- $P\left(\mathbf{Y} \mid \mathbf{G}^{\mathbf{T}}, \mathbf{Q}\right)$ directly with infinite sites mutations model

Outline

Inference with Tajima's coalescent:

- Felsenstein-Tajima conditional likelihood
- Bayesian model for inferring $N(t)$
- MCMC Algorithm for Posterior inference
  - Sampling of ranked tree shapes (F)
  - Sampling of coalescent times (t)
  - Sampling of $N(t)$
- Results
- Summary and future directions

Felsenstein-Tajima conditional likelihood

Goal:

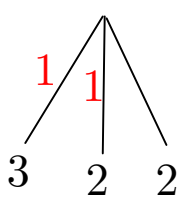
$$P(\mathbf{Y} \mid G^T = \{\mathbf{F}, \mathbf{t}\}, \mu)$$

Assumptions:

- We assume the infinite-sites mutation model
- We know the ancestral state at each polymorphic site
- Our data can be represented as sequences of 0s and 1s
- There is a one-to-one correspondence between data consistent with ISM and a perfect phylogeny (gene tree)

$$P(\mathbf{Y} \mid G^T = \{\mathbf{F}, \mathbf{t}\}, \mu) = P(PP \mid G^T = \{\mathbf{F}, \mathbf{t}\}, \mu)$$

1	A	C
2	A	C
3	A	C
4	T	C
5	T	C
6	T	G
7	T	G
ancestral:	T	C



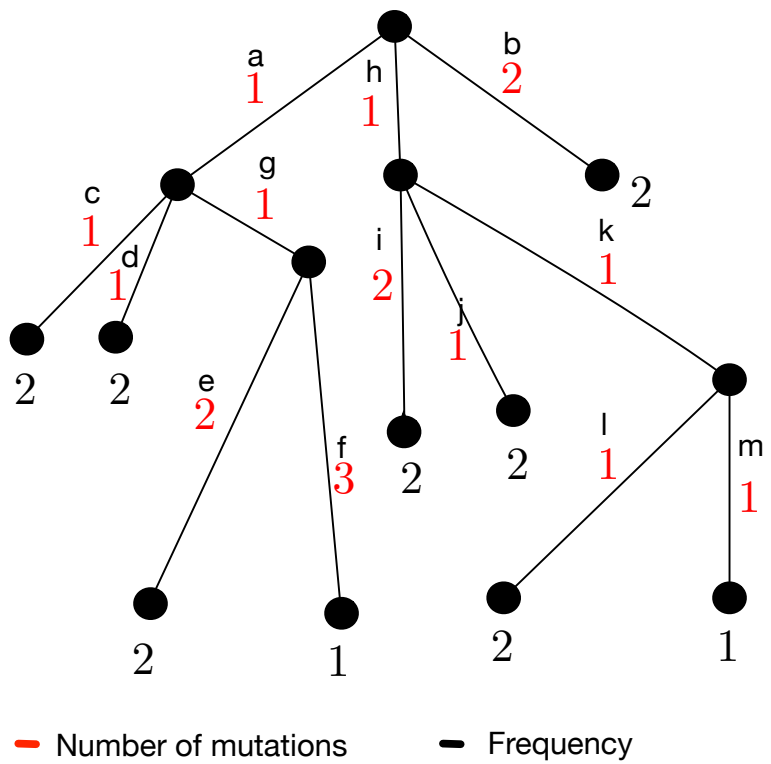
Data as perfect phylogeny (gene tree)

A bigger example (Dan Gusfield, 1991):

A Data $(\mathbf{Y})$

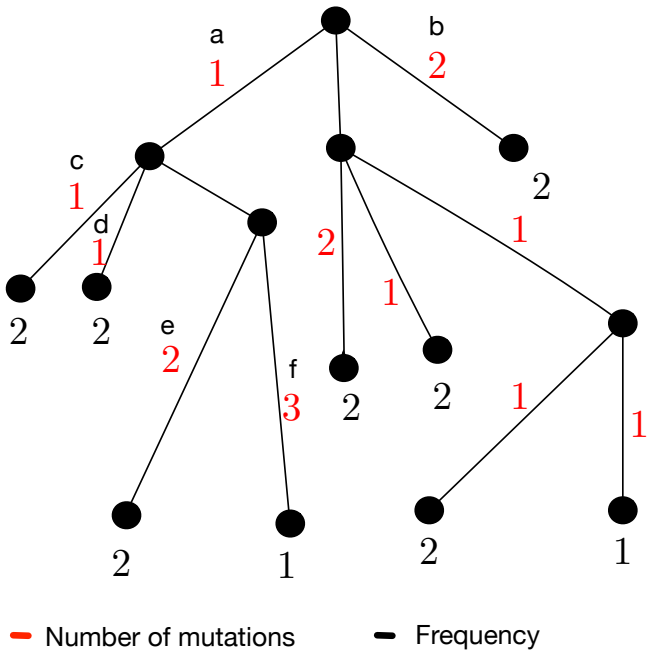
Haplotype	Frequency	a	b	b	c	d	e	e	f	f	f	...
1	2	1	0	0	1	0	0	0	0	0	0	
2	2	1	0	0	0	1	0	0	0	0	0	
3	2	1	0	0	0	0	1	1	0	0	0	
4	1	1	0	0	0	0	0	0	1	1	1	
5	2	0	0	0	0	0	0	0	0	0	0	
6	2	0	0	0	0	0	0	0	0	0	0	
7	2	0	1	1	0	0	0	0	0	0	0	
8	2	0	0	0	0	0	0	0	0	0	0	
9	1	0	0	0	0	0	0	0	0	0	0	
	16											

B Perfect Phylogeny $(\mathcal{T})$

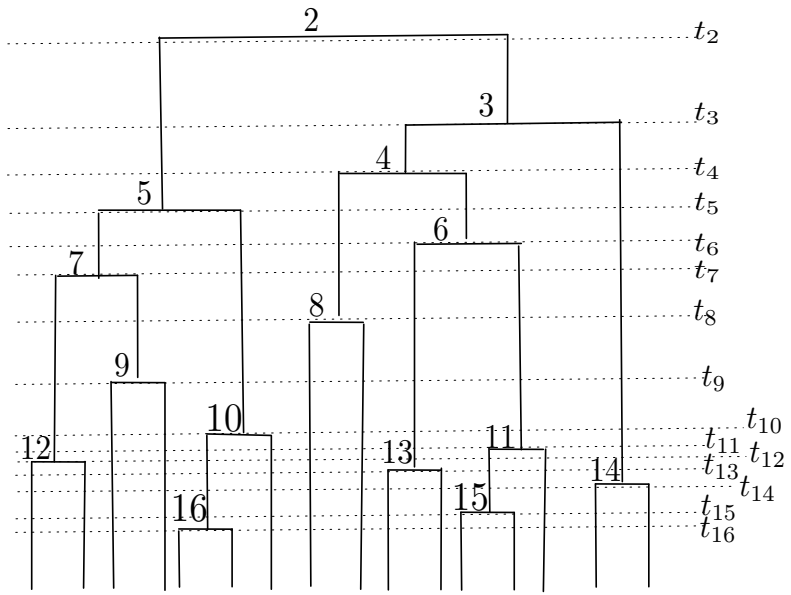


Felsenstein-Tajima conditional likelihood

Perfect Phylogeny $(\mathbf{Y})$



Tajima's genealogy (g^T)



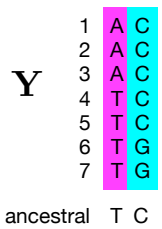
Goal:

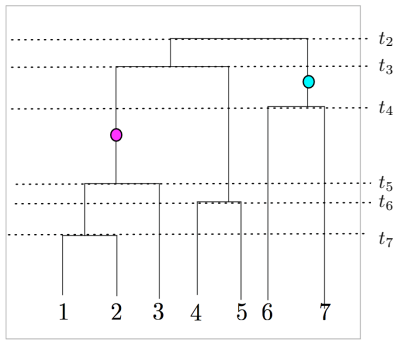
$$P(\mathbf{Y} \mid G^T = \{\mathbf{F}, \mathbf{t}\}, \mu)$$

Felsenstein-Tajima conditional likelihood

The tricky part: once we remove the labels, we can have more than one assignment of mutations to branches

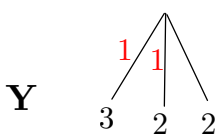
With Kingman's coalescent

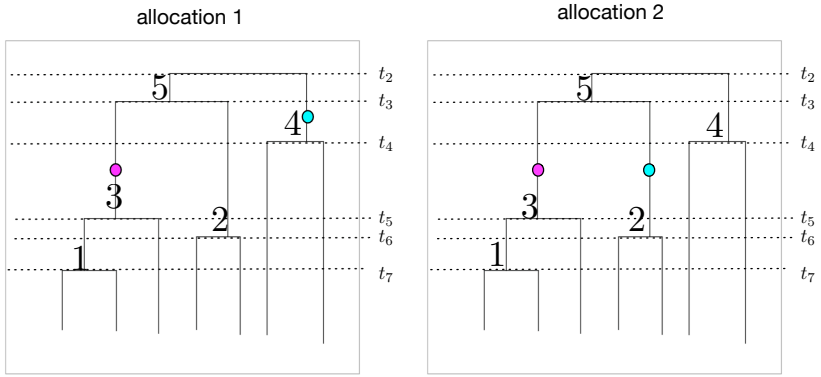




$$\mathrm{P}\left(\mathbf{Y} \mid \mathbf{G}, \mu\right) = \mathrm{P}\left(\text{●},\text{●} \mid \mathbf{G}, \mu\right)$$

With Tajima's coalescent

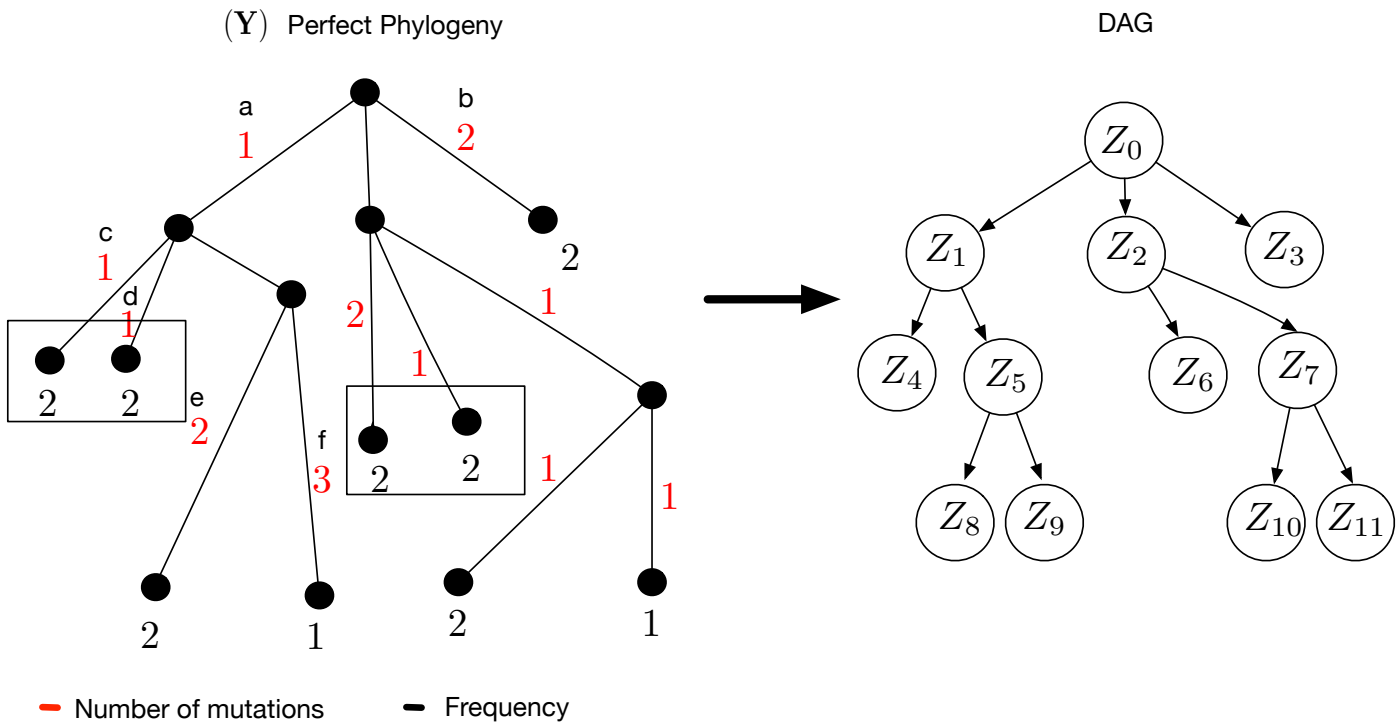




$$\mathrm{P}\left(\mathbf{Y} \mid \mathbf{G}^T, \mu\right) = \mathrm{P}\left(\text{●},\text{●}\text{in allocation 1 or allocation 2} \mid \mathbf{G}^T, \mu\right)$$

Felsenstein-Tajima conditional likelihood

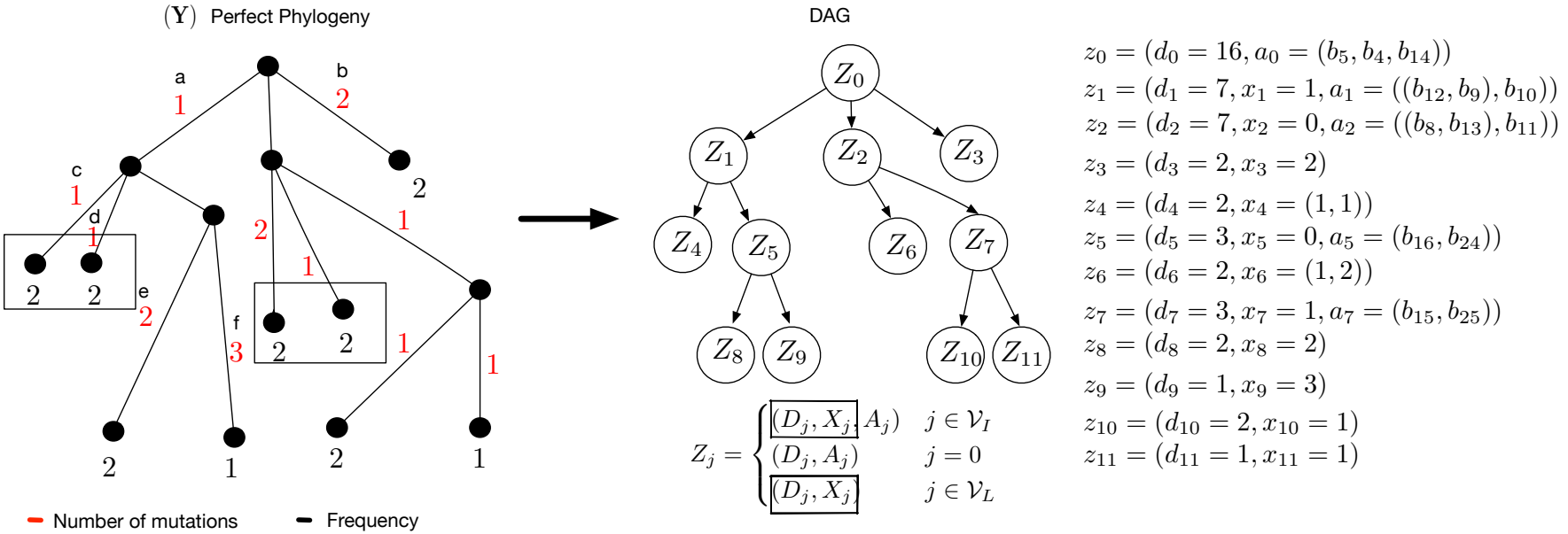
The computational helper: We use the structure of the perfect phylogeny to generate a probabilistic DAG model (Bayes network, Bayes nets, belief networks, Bayesian graphical model)



We merge sister leaves with the same number of descendants into a single leaf in the DAG.

Felsenstein-Tajima conditional likelihood

We equip the DAG with mutations ...

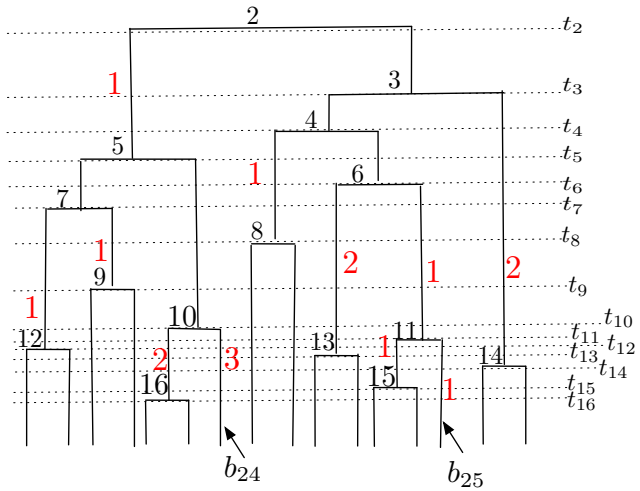


We merge sister leaves with the same number of descendants into a single leaf in the DAG.

Felsenstein-Tajima conditional likelihood

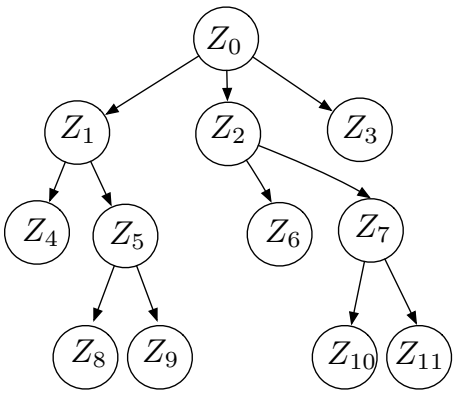
We augment the DAG with allocation of mutations (A_j) along $\mathbf{G}^T$

A Tajima's genealogy and a possible allocation of the mutations observed in the data



Number of mutations

B DAG corresponding to **A**



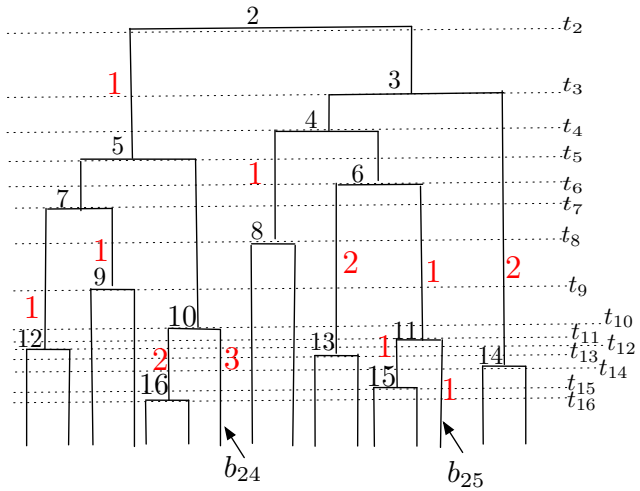
$$Z_j = \begin{cases} (D_j, X_j, A_j) & j \in \mathcal{V}_I \\ (D_j, A_j) & j = 0 \\ (D_j, X_j) & j \in \mathcal{V}_L \end{cases}$$

$$z_0 = (d_0 = 16, a_0 = (b_5, b_4, b_{14}))$$
$$z_1 = (d_1 = 7, x_1 = 1, a_1 = ((b_{12}, b_9), b_{10}))$$
$$z_2 = (d_2 = 7, x_2 = 0, a_2 = ((b_8, b_{13}), b_{11}))$$
$$z_3 = (d_3 = 2, x_3 = 2)$$
$$z_4 = (d_4 = 2, x_4 = (1, 1))$$
$$z_5 = (d_5 = 3, x_5 = 0, a_5 = (b_{16}, b_{24}))$$
$$z_6 = (d_6 = 2, x_6 = (1, 2))$$
$$z_7 = (d_7 = 3, x_7 = 1, a_7 = (b_{15}, b_{25}))$$
$$z_8 = (d_8 = 2, x_8 = 2)$$
$$z_9 = (d_9 = 1, x_9 = 3)$$
$$z_{10} = (d_{10} = 2, x_{10} = 1)$$
$$z_{11} = (d_{11} = 1, x_{11} = 1)$$

Felsenstein-Tajima conditional likelihood

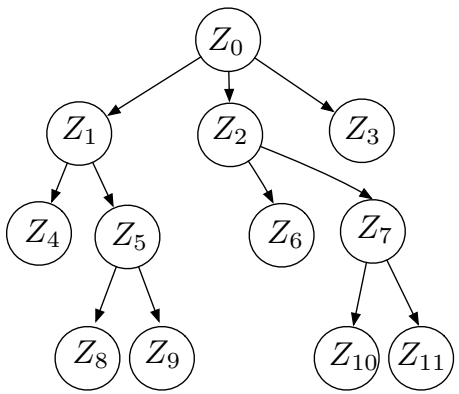
We augment the DAG with allocation of mutations (A_j) along $\mathbf{G}^T$

A Tajima's genealogy and a possible allocation of the mutations observed in the data



Number of mutations

B DAG corresponding to **A**

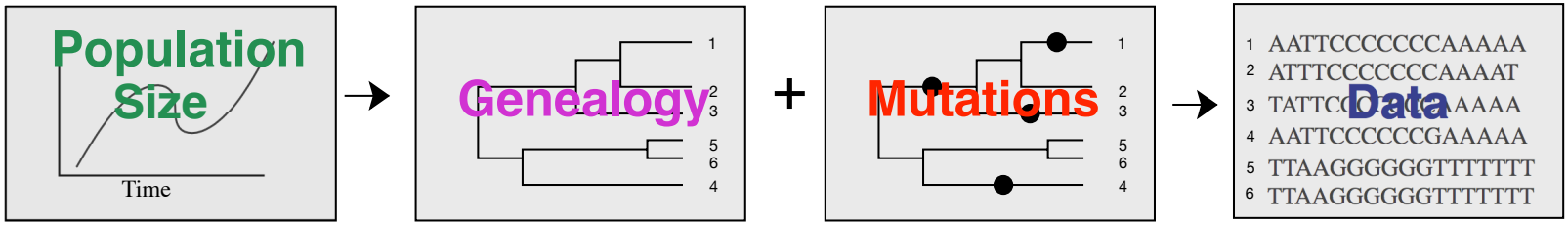


$$Z_j = \begin{cases} (D_j, X_j, A_j) & j \in \mathcal{V}_I \\ (D_j, A_j) & j = 0 \\ (D_j, X_j) & j \in \mathcal{V}_L \end{cases}$$

$z_0 = (d_0 = 16, a_0 = (b_5, b_4, b_{14}))$
$z_1 = (d_1 = 7, x_1 = 1, a_1 = ((b_{12}, b_9), b_{10}))$
$z_2 = (d_2 = 7, x_2 = 0, a_2 = ((b_8, b_{13}), b_{11}))$
$z_3 = (d_3 = 2, x_3 = 2)$
$z_4 = (d_4 = 2, x_4 = (1, 1))$
$z_5 = (d_5 = 3, x_5 = 0, a_5 = (b_{16}, b_{24}))$
$z_6 = (d_6 = 2, x_6 = (1, 2))$
$z_7 = (d_7 = 3, x_7 = 1, a_7 = (b_{15}, b_{25}))$
$z_8 = (d_8 = 2, x_8 = 2)$
$z_9 = (d_9 = 1, x_9 = 3)$
$z_{10} = (d_{10} = 2, x_{10} = 1)$
$z_{11} = (d_{11} = 1, x_{11} = 1)$

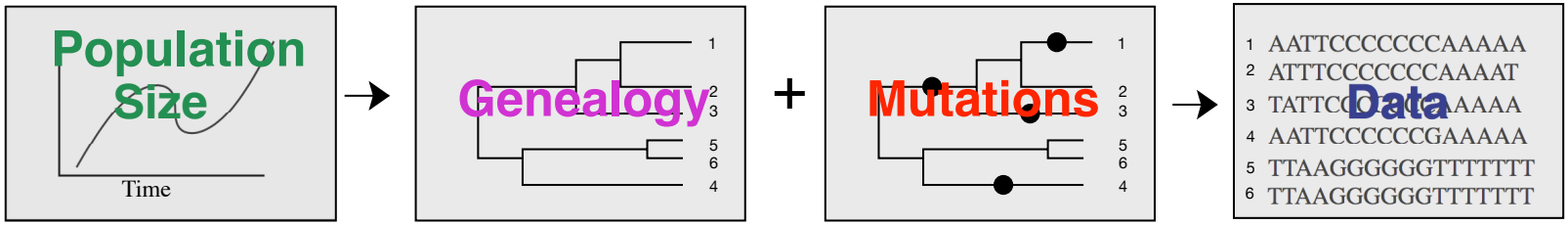
$$\begin{aligned} P(\mathbf{Y} \mid G^T, \mu) \propto & \sum_{A_0} \sum_{A_1} \cdots \sum_{A_{n_I}} P(\mathbf{D} \mid G^T, \mu) e^{-\mu \mathcal{L}} \\ = & \; e^{-\mu \mathcal{L}} \sum_{A_0} \sum_{A_1} \cdots \sum_{A_{n_I}} P(Z_0, \ldots, Z_{n_I+n_L} \mid G^T, \mu) \\ = & \; e^{-\mu \mathcal{L}} \sum_{A_0} \sum_{A_1} \cdots \sum_{A_{n_I}} \prod_{i=1}^{n_I+n_L} P(Z_i \mid Z_{pa(i)}, G^T, \mu) \end{aligned}$$

Bayesian model for inferring $N(t)$



- $\gamma = \log N(t) \sim GP(0, C(\tau)), \tau \sim Gamma(\alpha, \beta)$

Bayesian model for inferring $N(t)$



- $\gamma = \log N(t) \sim GP(0, C(\tau))$, $\tau \sim Gamma(\alpha, \beta)$
- Tajima coalescent prior

$$P[G^T = \{\mathbf{F}, \mathbf{t}\} \mid N_e(t)] = P(\mathbf{F}) \prod_{k=2}^{n} P[t_k \mid t_{k+1}, N_e(t)], \quad (2)$$

and

$$P(\mathbf{F}) = \frac{2^{n-c-1}}{(n-1)!}, \quad (3)$$

$$P[t_{k-1} | t_k, N_e(t)] = \frac{C_{k-1}}{N_e(t_{k-1})} \exp\left[- \int_{t_k}^{t_{k-1}} \frac{C_{k-1} dt}{N_e(t)} \right], \quad (4)$$

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

MCMC Algorithm for Posterior inference

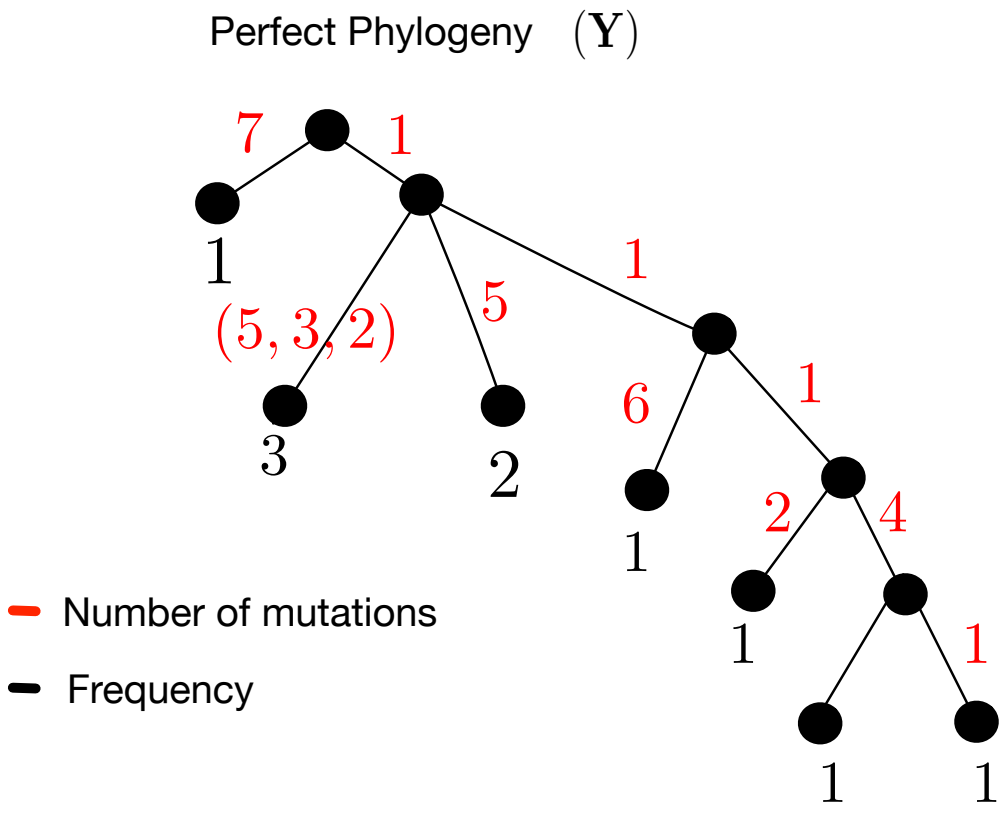
Goal:

$$P[\boldsymbol{\gamma}, G^T, \tau \mid \mathbf{Y}, \mu] \propto P(\mathbf{Y} \mid G^T, \mu) P[G^T \mid \boldsymbol{\gamma}] P[\boldsymbol{\gamma} \mid \tau] P(\tau) \qquad (5)$$

Metropolis-Hastings

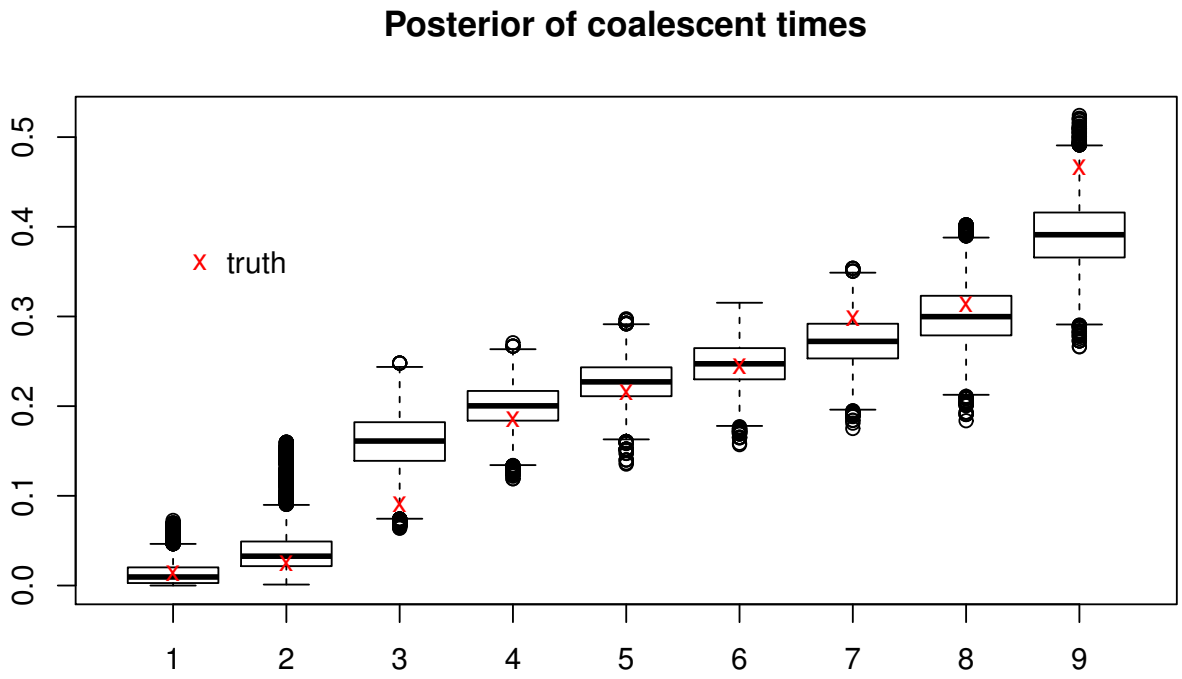
- splitHMC [Lan et al., Bioinformatics 2015] to sample $\boldsymbol{\gamma}, \tau$
- HMC to sample $\mathbf{t}$
- $q(F \mid \mathbf{Y})$ proposal for ranked tree shapes exploiting the DAG representation of perfect phylogeny

Simulation

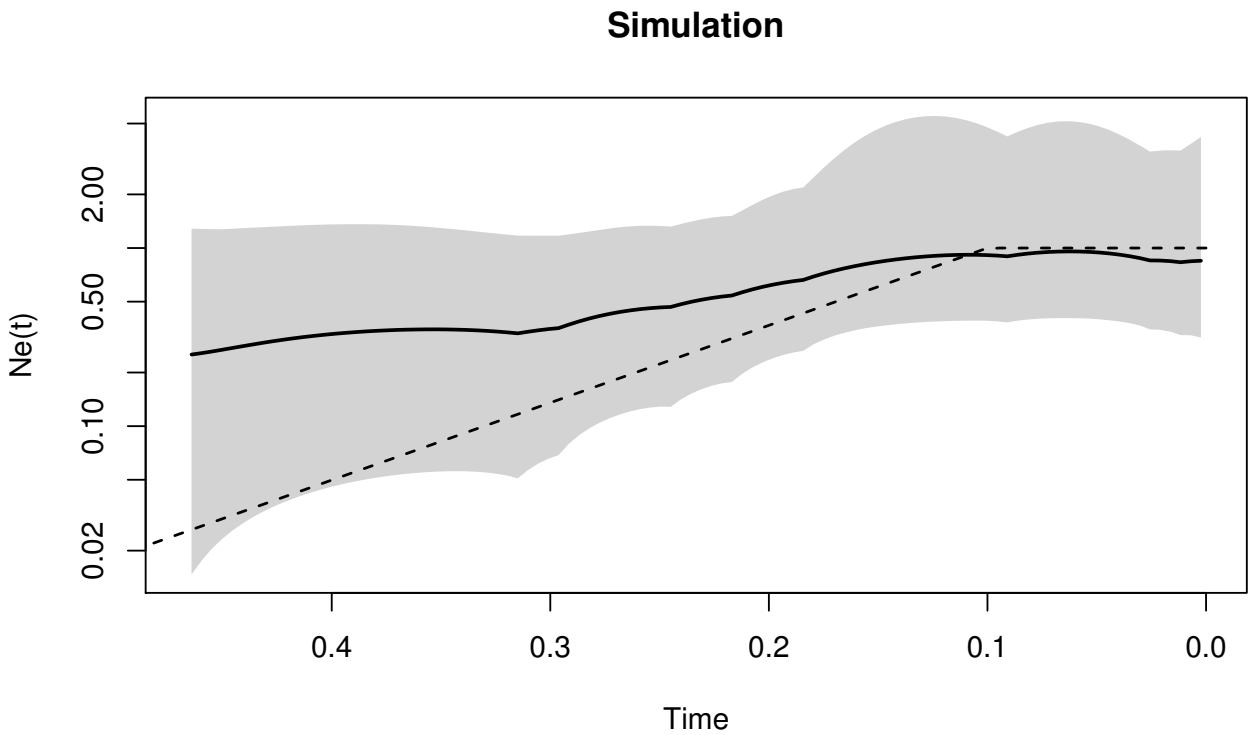


- Only 548 ranked tree shapes are compatible with the data

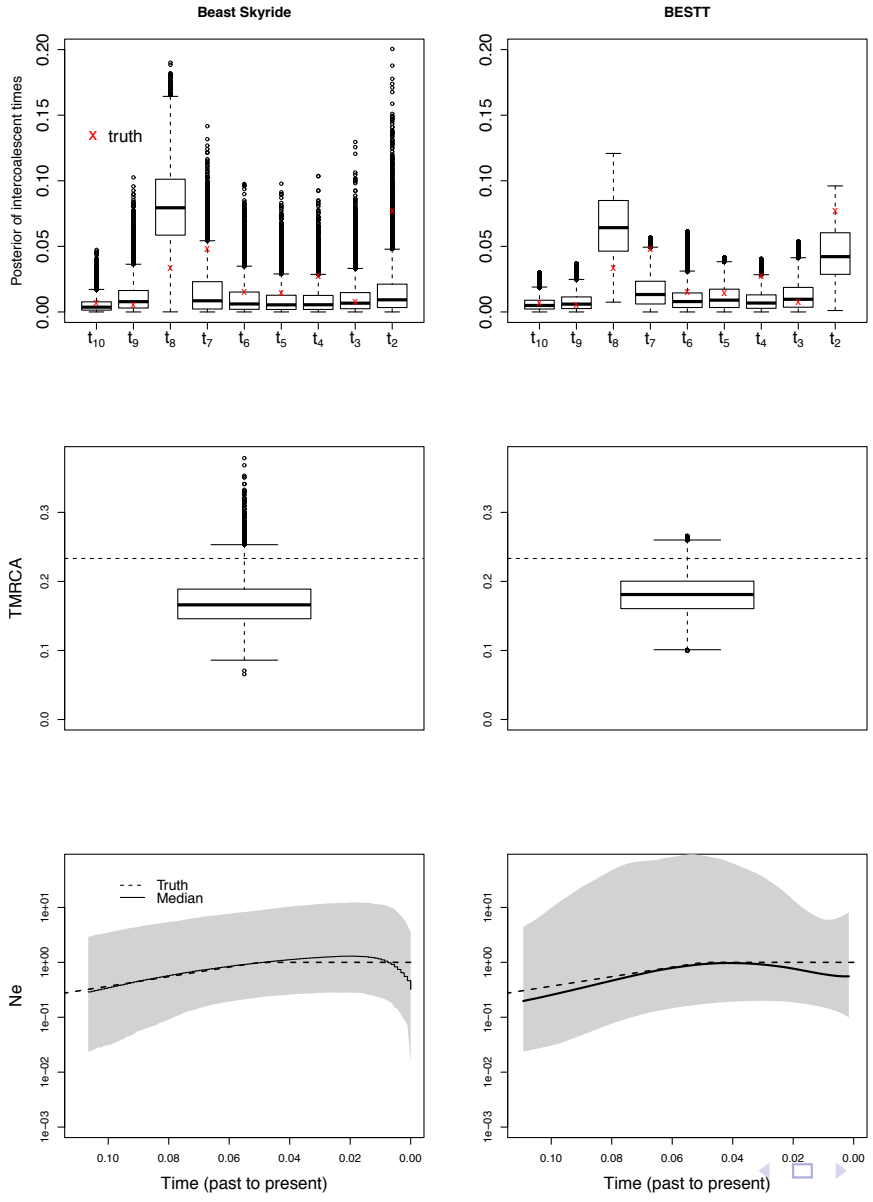
Simulation: Sampling coalescent times



Simulation: Sampling Ne



Comparison with BEAST [Suchard et al (2018)]



Summary

- Tajima's coalescent is a more **efficient** lower resolution model for inference of effective population sizes.
- A priori, the hidden state space of ranked tree shapes is much **smaller** than the **space** of labeled topologies.
- Current implementation (in R `phylodyn`) is **limited to the infinite-sites** mutation model.
- Tajima's inference can be extended for modeling recombination under the sequential Markovian coalescent.
- The Felsenstein-Tajima conditional likelihood calculation can be extended for tree shapes.

Size of ranked tree shapes

- The cardinality of the hidden space of ranked tree shapes depends on your observed data.
- For our simulation example ($n = 10$ samples)
  - $|\mathcal{T}_{10}| = 2.5 \times 10^9$ labeled topologies

Size of ranked tree shapes

- The cardinality of the hidden space of ranked tree shapes depends on your observed data.
- For our simulation example ($n = 10$ samples)
  - $|\mathcal{T}_{10}| = 2.5 \times 10^9$ labeled topologies
  - $|\mathcal{R}_{10}| = 7936$

Size of ranked tree shapes

- The cardinality of the hidden space of ranked tree shapes depends on your observed data.
- For our simulation example ($n = 10$ samples)
  - $|\mathcal{T}_{10}| = 2.5 \times 10^9$ labeled topologies
  - $|\mathcal{R}_{10}| = 7936$
  - Only $|\mathcal{R}_{10} \mid \mathbf{Y}| = 548$ compatible with the data

Size of ranked tree shapes

- The cardinality of the hidden space of ranked tree shapes depends on your observed data.
- For our simulation example ($n = 10$ samples)
  - $|\mathcal{T}_{10}| = 2.5 \times 10^9$ labeled topologies
  - $|\mathcal{R}_{10}| = 7936$
  - Only $|\mathcal{R}_{10} \mid \mathbf{Y}| = 548$ compatible with the data

Size of ranked tree shapes

- The cardinality of the hidden space of ranked tree shapes depends on your observed data.
- For our simulation example ($n = 10$ samples)
  - $|\mathcal{T}_{10}| = 2.5 \times 10^9$ labeled topologies
  - $|\mathcal{R}_{10}| = 7936$
  - Only $|\mathcal{R}_{10} \mid \mathbf{Y}| = 548$ compatible with the data
- Let $\beta_x = (x+1, n-x+1)$ be a binary perfect phylogeny with two leaves with $x+1$ leaves on one side and $n-x+1$ on the other side.

Size of ranked tree shapes

- The cardinality of the hidden space of ranked tree shapes depends on your observed data.
- For our simulation example ($n = 10$ samples)
  - $|\mathcal{T}_{10}| = 2.5 \times 10^9$ labeled topologies
  - $|\mathcal{R}_{10}| = 7936$
  - Only $|\mathcal{R}_{10} \mid \mathbf{Y}| =$ 548 compatible with the data
- Let $\beta_x = (x+1, n-x+1)$ be a binary perfect phylogeny with two leaves with $x+1$ leaves on one side and $n-x+1$ on the other side.
- The set of unlabeled histories (ranked tree shapes) compatible with β_x, assuming $x < n/2$ is

$$|R_{n+2} \mid \beta_x| = e_x \cdot e_{n-x} \cdot \binom{n}{x}, \tag{6}$$

where e_i is the number of unlabeled histories with i internal nodes. The integer e_i is the ith Euler number:

$$\sum_{i=0}^{\infty} \frac{e_i z^i}{i!} = \frac{1}{\cos(z)} + \tan(z) \tag{7}$$

Size of ranked tree shapes

Theorem

The maximum number of ranked tree shapes of n leaves compatible with the perfect phylogeny occurs when the perfect phylogeny is a multifurcating phylogeny of degree n or $n-1$.

Size of ranked tree shapes

Theorem

The maximum number of ranked tree shapes of n leaves compatible with the perfect phylogeny occurs when the perfect phylogeny is a multifurcating phylogeny of degree n or $n-1$.

Theorem

The maximum number of ranked tree shapes of n leaves compatible with a binary perfect phylogeny occurs when there is a single bifurcation event dividing the samples in two groups of sizes $(n-2)$ and (2).

Acknowledgments

- Amandine Veber (Ecole Polytechnique)
- John Wakeley (Harvard University)
- Sohini Ramachandran (Brown University)
- Zhangyuan Wang (Stanford University)
- Noah Rosenberg (Stanford University)
- Filippo Disanto (Stanford University)
- Anand Bhaskar (Facebook)