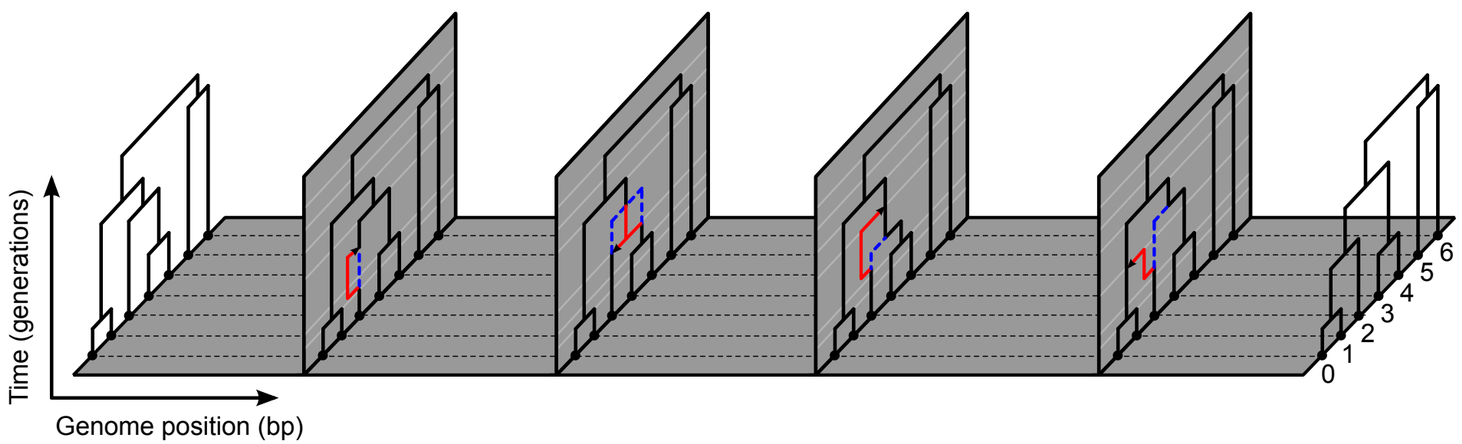
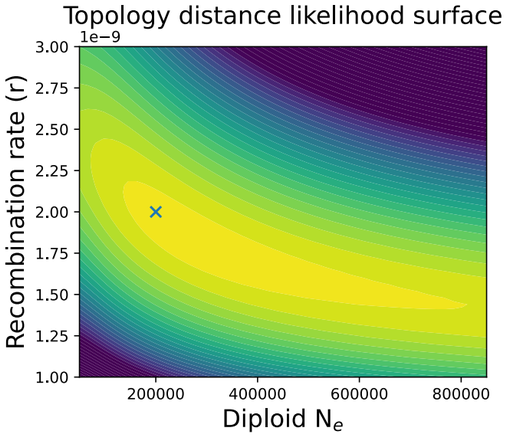
Linking Phylogenetic Inference at Genome-wide and Genealogical Scales



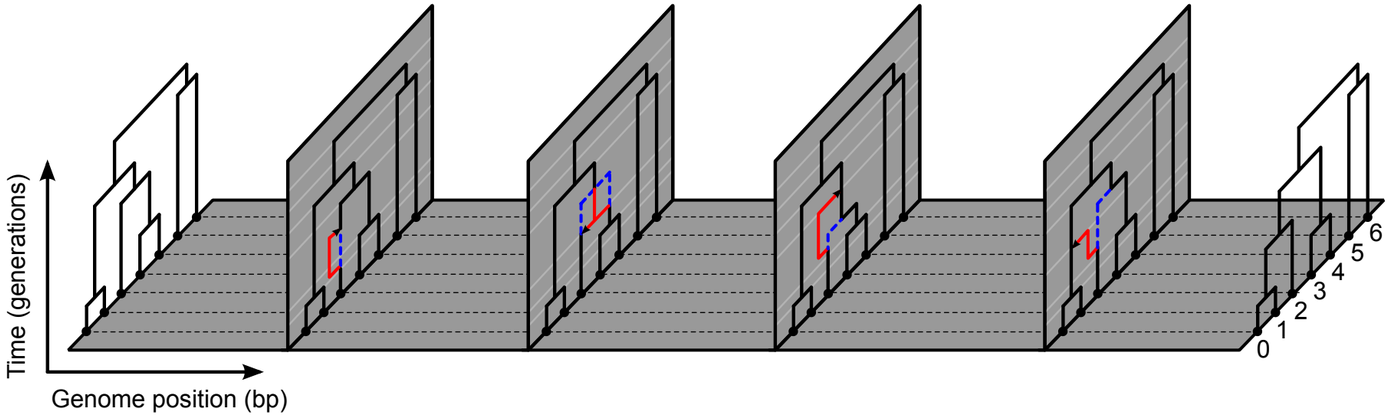


Deren Eaton and Patrick McKenzie

Ecology, Evolution, and Environmental Biology, Columbia University

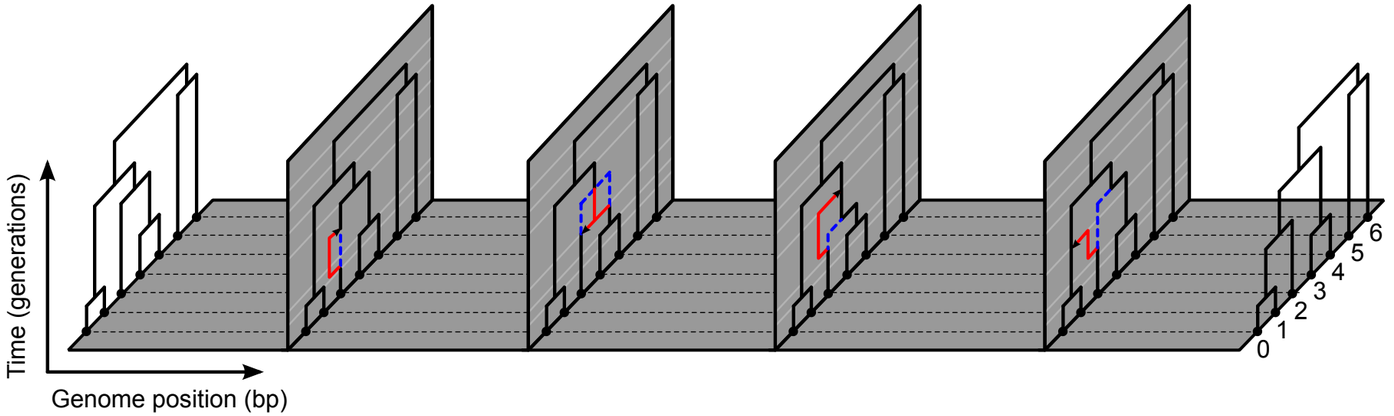
Genealogical variation

Genomes are composed of a mosaic of segments inherited from different ancestors, each separated by past recombination events.



Genealogical variation

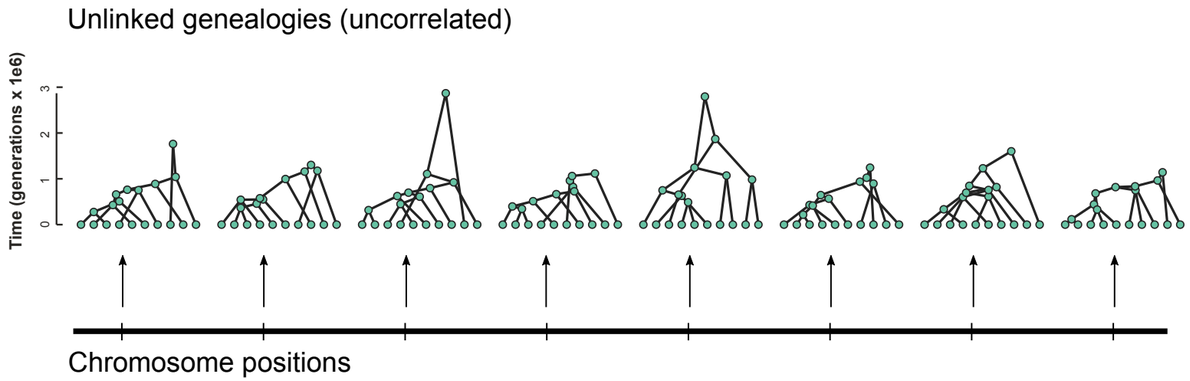
Genomes are composed of a mosaic of segments inherited from different ancestors, each separated by past recombination events.



Consequently, genealogical relationships vary spatially across genomes.

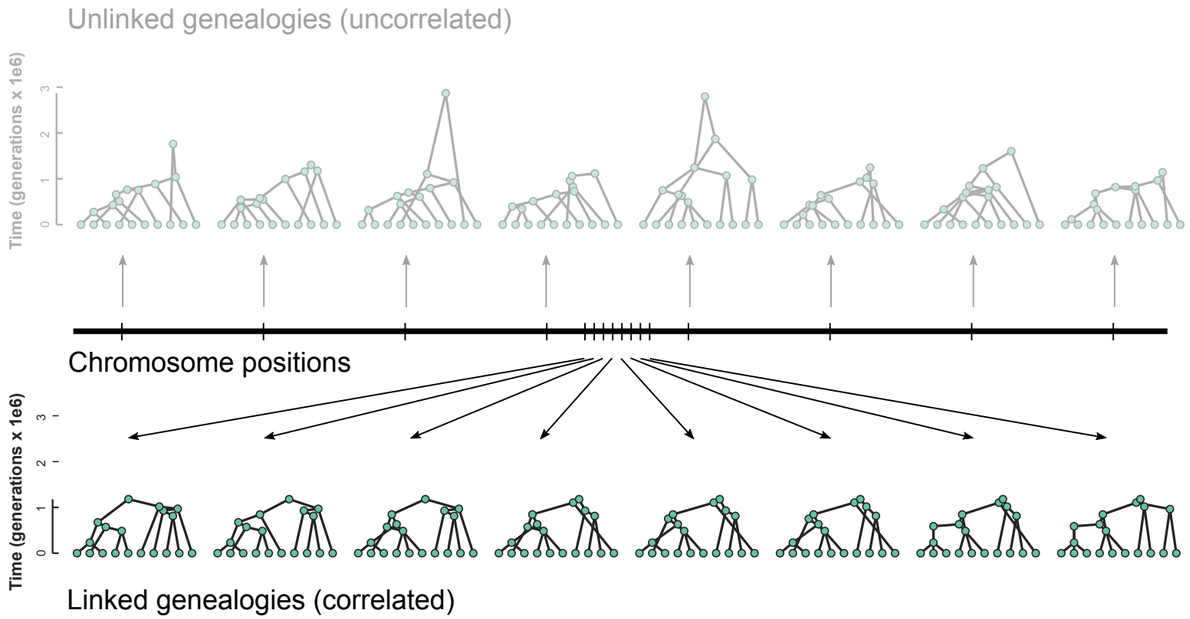
Multispecies coalescent assumptions

The multispecies coalescent (MSC) describes the expected distribution of ***unlinked*** genealogies, as a function of demographic model parameters (N_e, τ, topology).



Multispecies coalescent assumptions

The multispecies coalescent (MSC) describes the expected distribution of ***unlinked*** genealogies, as a function of demographic model parameters (N_e, τ, topology).



The expected distribution of ***linked*** genealogical variation is poorly characterized.

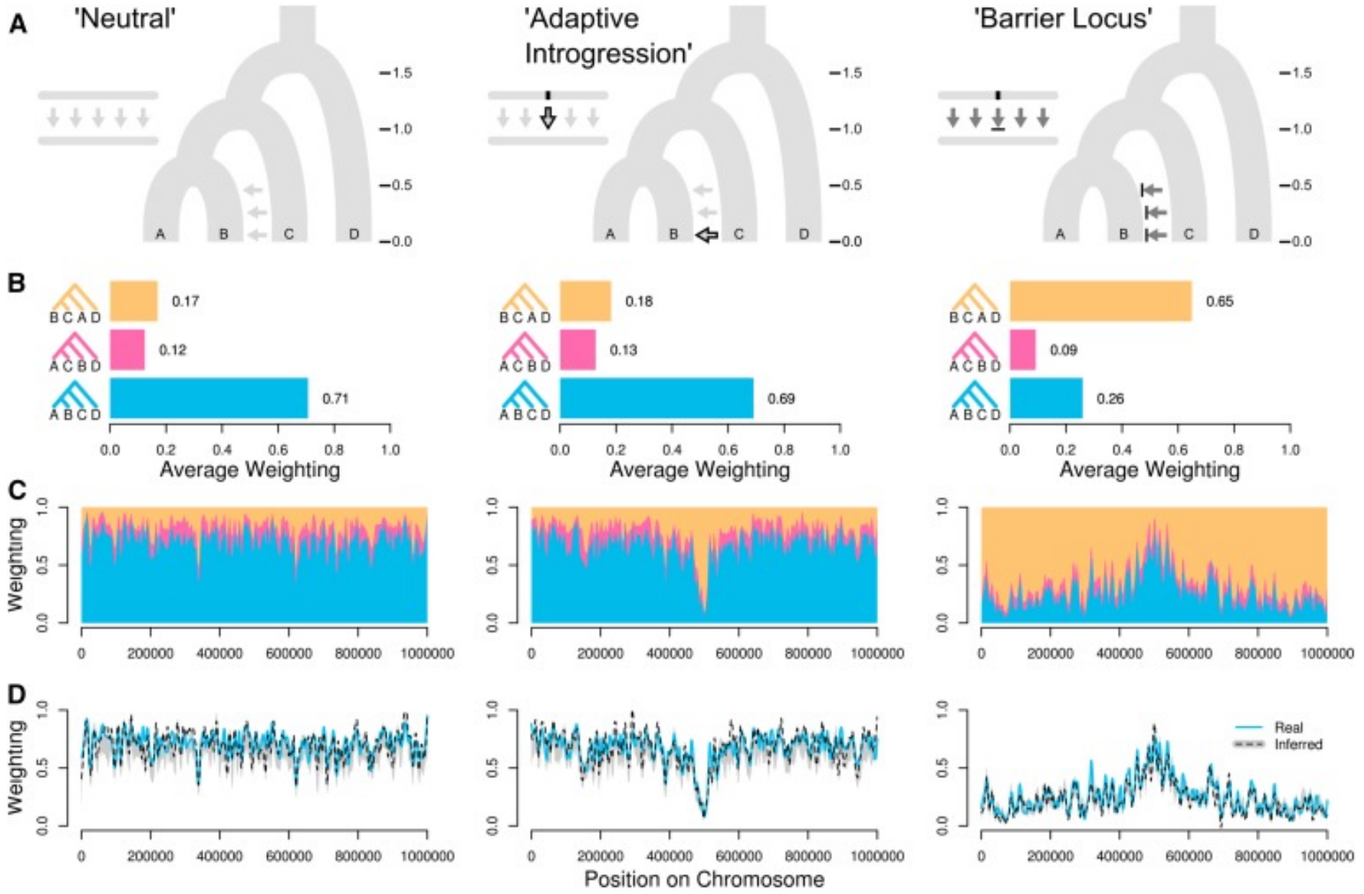
What is the expectation for the distribution of *linked* genealogical variation?

How does it relate to demographic model parameters?

Why care about local genealogical variation?

- Subsampling unlinked loci effectively discards >99% genomic info.
- Ignoring linkage introduces bias (*concatalescence*; Gatesy 2013).
- Local ancestry is informative about selection and introgression.

Why care about local genealogical variation?



(Martin & Belleghem 2017)

Why care about local genealogical variation?

- Subsampling unlinked loci effectively discards >99% genomic info.
- Ignoring linkage introduces bias (*concatalescence*; Gatesy 2013).
- Local ancestry is informative about selection and introgression.
- **We lack a null expectation for spatial genealogical variation.**

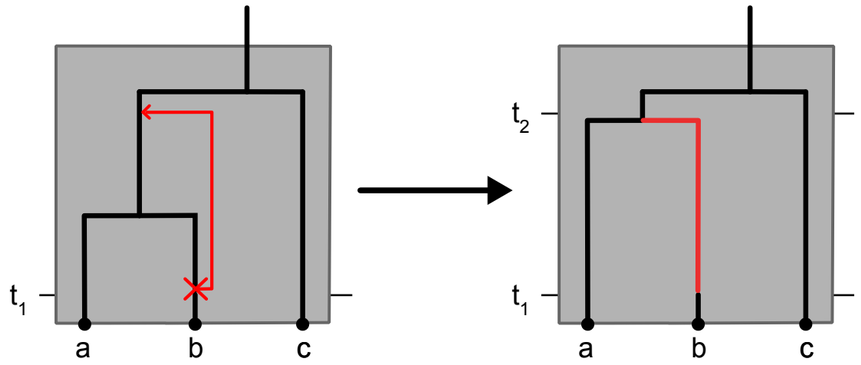
Outline: *Multispecies Sequentially Markov Coalescent*

- Background: *SMC' model.*
- *SMC' waiting distances (Deng et al. 2021) in a single population.*
- *Introduce our new model for MS-SMC' waiting distances.*
- *Validate solutions against stochastic coalescent simulations.*
- *Demonstrate likelihood framework to use waiting distances to fit models.*

Sequentially Markov Coalescent (McVean and Cardin, 2005)

An approximation of the coalescent with recombination

Given a starting genealogy a change to the next genealogy is modeled as a Markov process — a single transition — which enables a tractable likelihood framework.

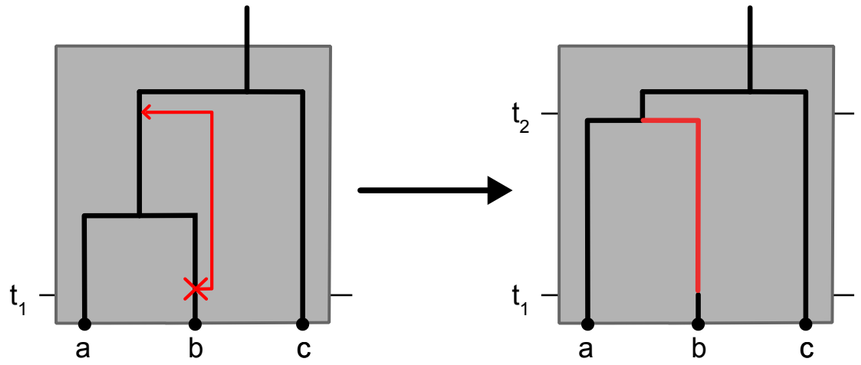


Sequentially Markov Coalescent (McVean and Cardin, 2005)

An approximation of the coalescent with recombination

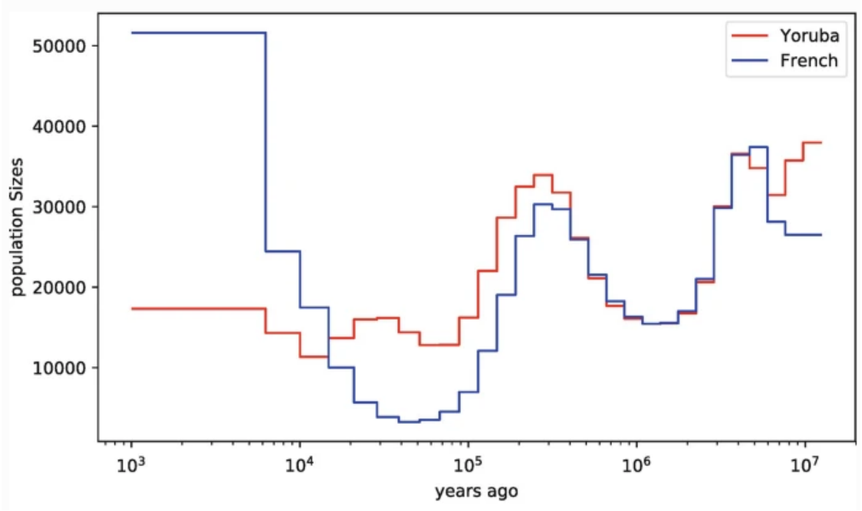
Given a starting genealogy a change to the next genealogy is modeled as a Markov process — a single transition — which enables a tractable likelihood framework.

Process: recombination occurs w/ uniform probability anywhere on a tree (t_1), creating a detached subtree, which re-coalesces above t_1 with an ancestral lineage.



SMC' is widely used in HMM methods

PSMC (Li & Durbin 2011), MSMC (Schiffels & Durbin 2014), use pairwise coalescent times between sequential genealogies to infer changes in N_e through time.

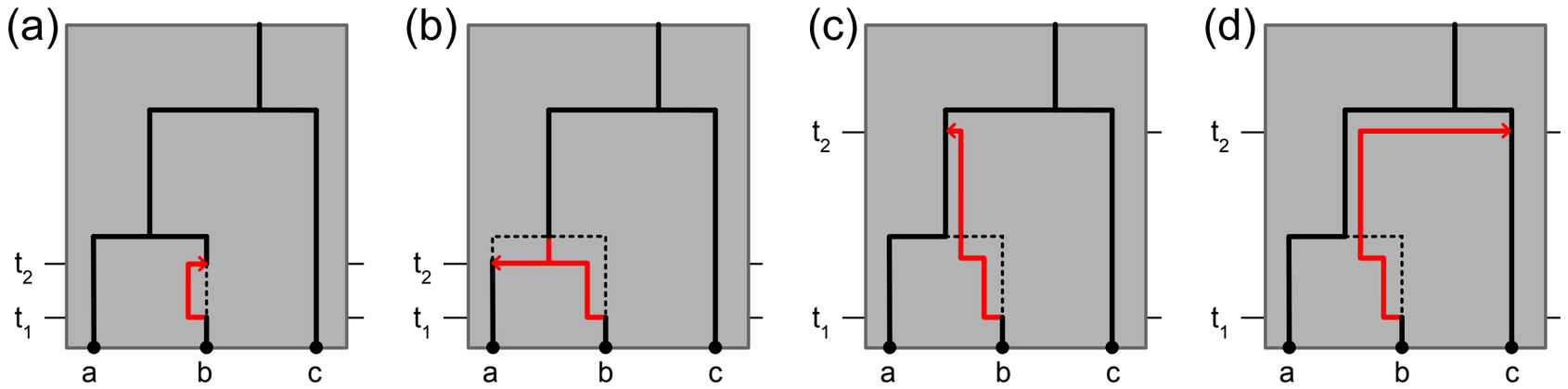


ARGweaver (Rasmussen et al. 2014) and ARGweaver-D (Hubisz & Siepel 2020) use an SMC'-based conditional sampling method to infer ARGs from sequence data.

Currently, we extract a fairly limited amount of spatial information from genomes.

Categorical event outcomes under the SMC'

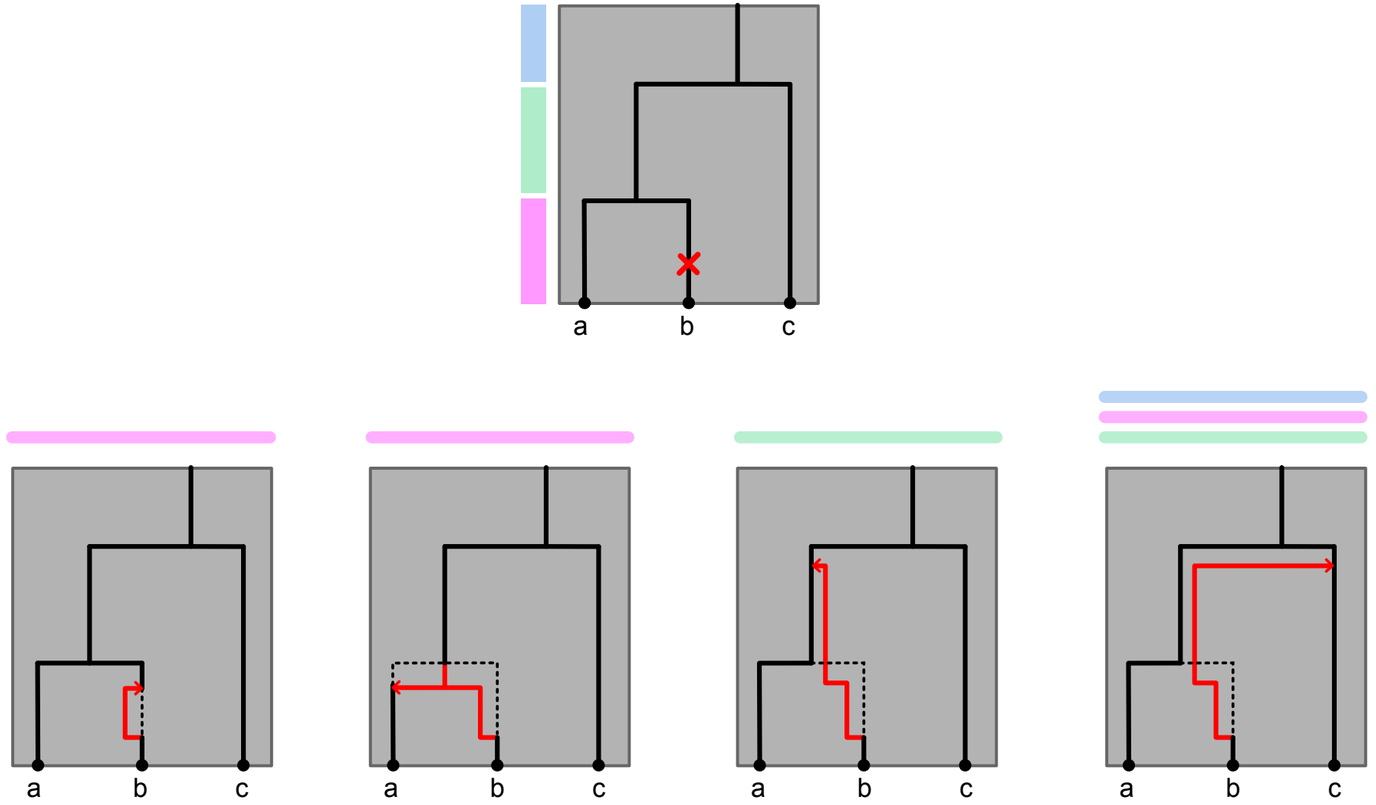
(a) no-change; (b-c) tree-change; and (d) topology-change.



(Deng et al. 2021)

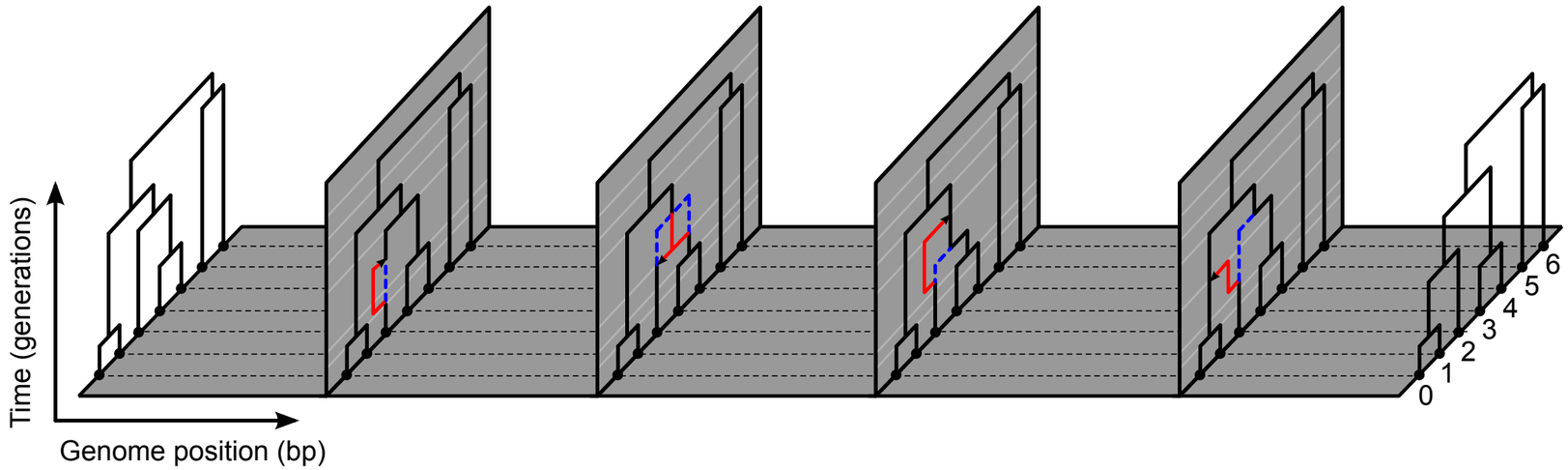
The distribution of waiting distances in ancestral recombination graphs

Yun Deng [a,*], Yun S. Song [b,c,d], Rasmus Nielsen [b,e]



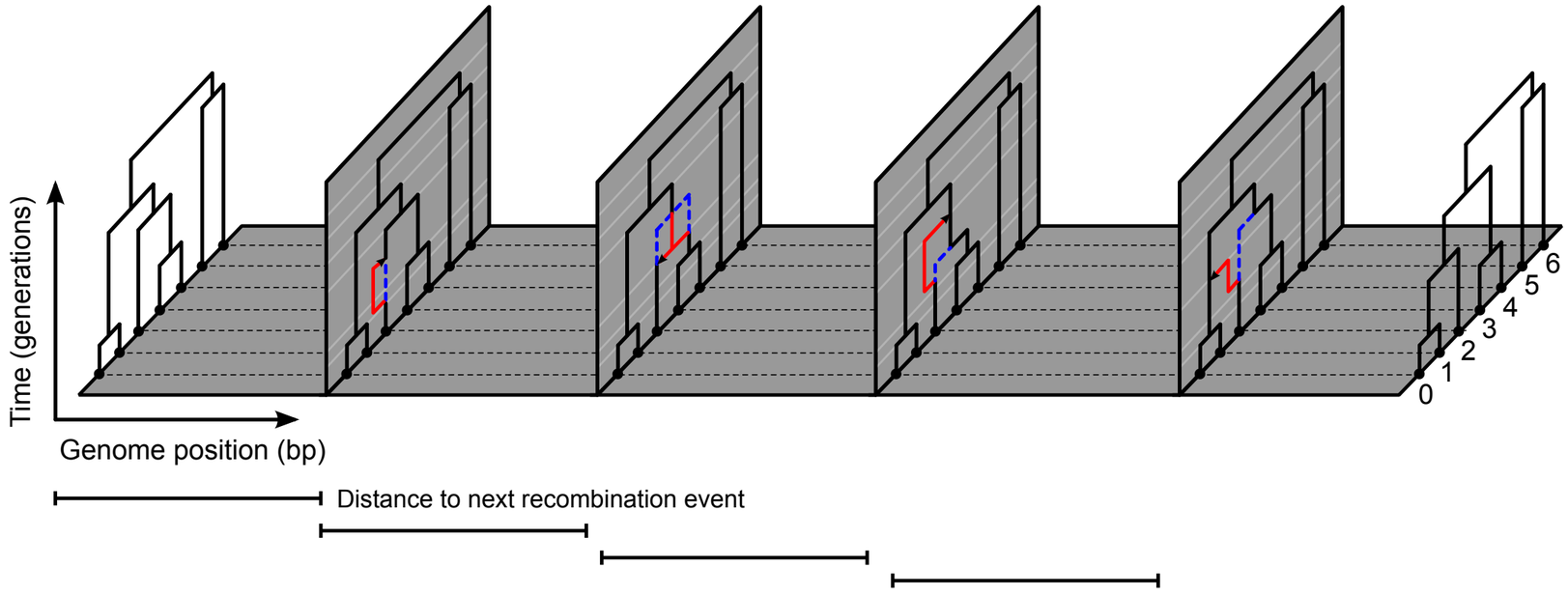
Estimating waiting distances under the SMC'

Expected Tree and Topology Distances represent new spatial genetic information.



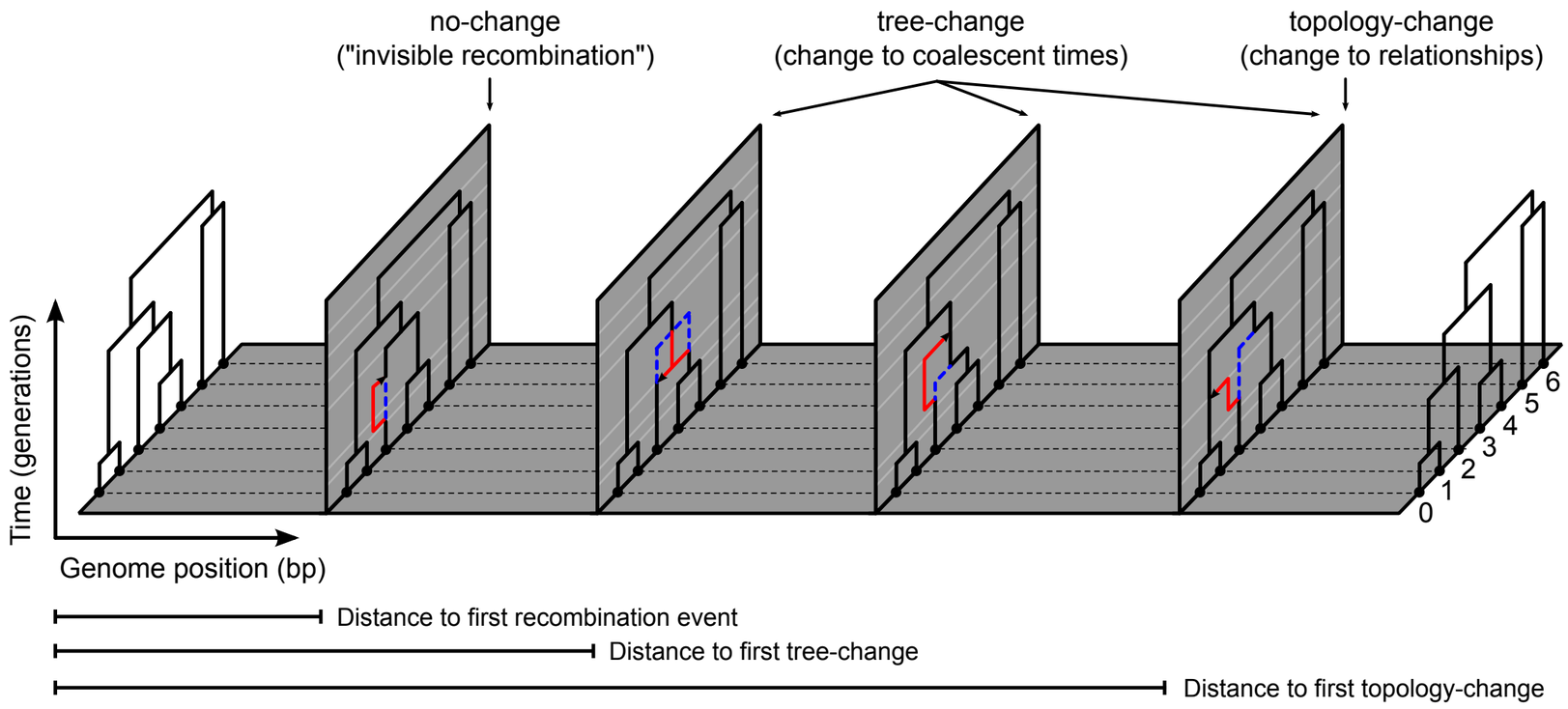
Estimating waiting distances under the SMC'

Expected Tree and Topology Distances represent new spatial genetic information.



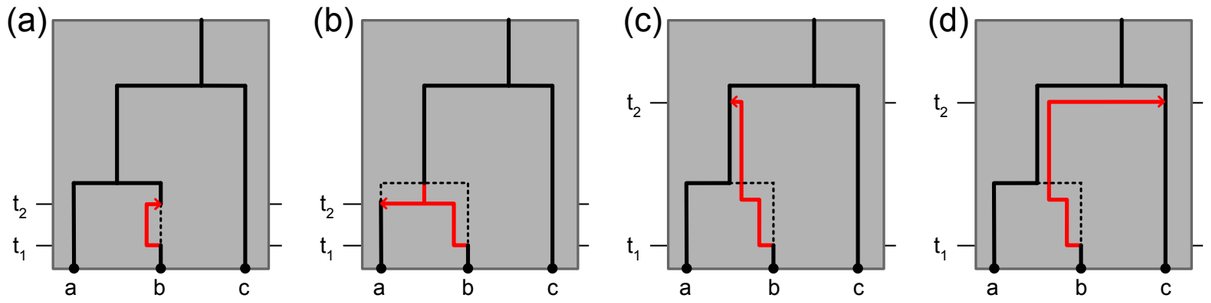
Estimating waiting distances under the SMC'

Expected Tree and Topology Distances represent new spatial genetic information.



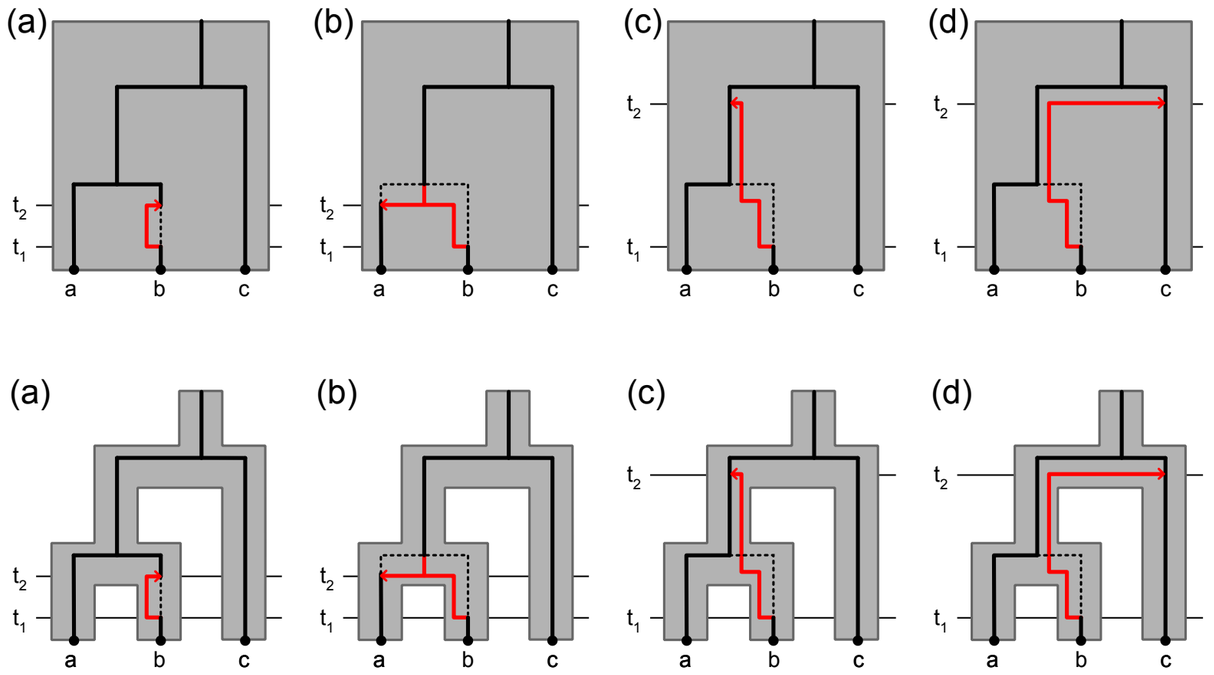
A multispecies extension to estimating waiting distances

Barriers to coalescence and variable N_e among species tree intervals.



A multispecies extension to estimating waiting distances

Barriers to coalescence and variable N_e among species tree intervals.

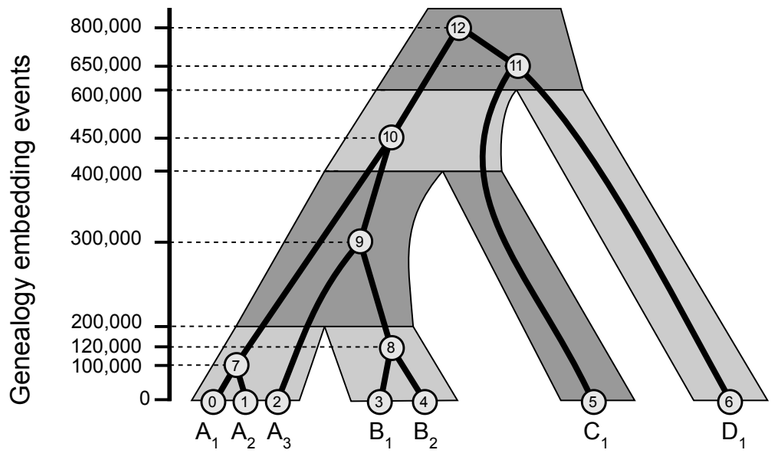


Patrick McKenzie

PhD student

Extending SMC' waiting distance estimation

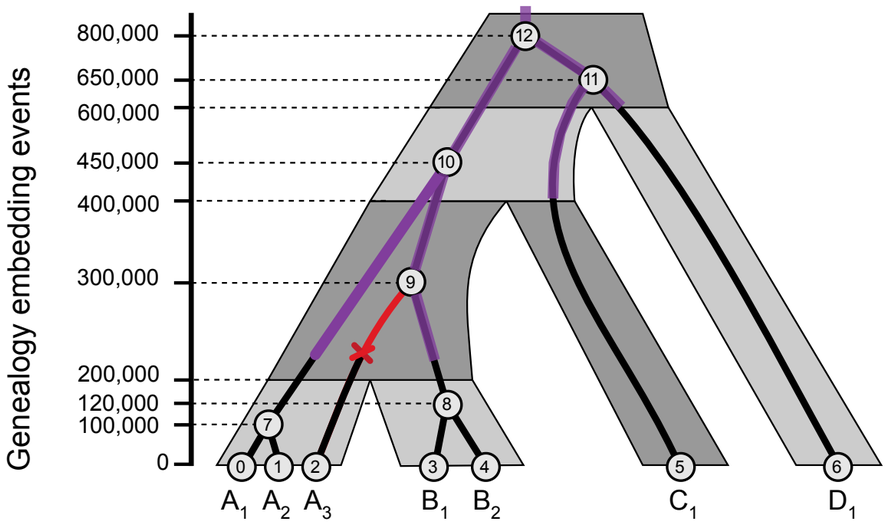
Genealogy embedding table with piecewise constant coal rates in all intervals between coal events or population intervals.



	start	stop	st_node	neff	nedges	coal	edges	dist
0	0	100000	0	uniform (see plots)	3	7	[0, 1, 2]	100000
1	100000	200000	0	uniform (see plots)	2	<NA>	[2, 7]	100000
2	0	120000	1	uniform (see plots)	2	8	[3, 4]	120000
3	120000	200000	1	uniform (see plots)	1	<NA>	[8]	80000
4	200000	300000	4	uniform (see plots)	3	9	[2, 7, 8]	100000
5	300000	400000	4	uniform (see plots)	2	<NA>	[7, 9]	100000
6	0	400000	2	uniform (see plots)	1	<NA>	[5]	400000
7	400000	450000	5	uniform (see plots)	3	10	[5, 7, 9]	50000
8	450000	600000	5	uniform (see plots)	2	<NA>	[5, 10]	150000
9	0	600000	3	uniform (see plots)	1	<NA>	[6]	600000
10	600000	650000	6	uniform (see plots)	3	11	[5, 6, 10]	50000
11	650000	800000	6	uniform (see plots)	2	12	[10, 11]	150000
12	800000	<NA>	6	uniform (see plots)	1	<NA>	[12]	<NA>

MS-SMC' analytical solutions

$$\mathbb{P}(\textit{tree-unchanged}|\mathcal{S},\mathcal{G},b,t_r) = \int_{t_r}^{t_b^u} \frac{1}{2N(\tau)} e^{-\int_{t_r}^{\tau} \frac{A(s)}{2N(s)} ds} d\tau$$



MS-SMC' analytical solutions

$$\mathbb{P}(\textit{tree-unchanged}|\mathcal{S},\mathcal{G},b,t_r) = \int_{t_r}^{t_b^u} \frac{1}{2N(\tau)} e^{-\int_{t_r}^{\tau} \frac{A(s)}{2N(s)} ds} d\tau$$

MS-SMC' analytical solutions

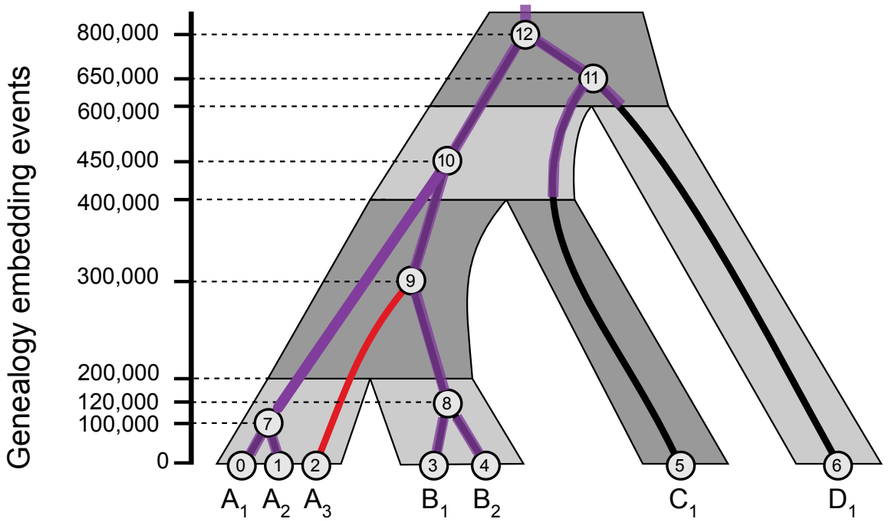
$$\mathbb{P}(\textit{tree-unchanged}|\mathcal{S},\mathcal{G},b,t_r) = \int_{t_r}^{t_b^u} \frac{1}{2N(\tau)} e^{-\int_{t_r}^{\tau} \frac{A(s)}{2N(s)} ds} d\tau$$

$$\mathbb{P}(\textit{tree-unchanged}|\mathcal{S},\mathcal{G},b) = \frac{1}{t_b^u - t_b^l} \int_{t_b^l}^{t_b^u} \mathbb{P}(\textit{tree-unchanged}|\mathcal{S},\mathcal{G},b,t)dt$$

MS-SMC' analytical solutions

$$\mathbb{P}(\textit{tree-unchanged}|\mathcal{S},\mathcal{G},b,t_r) = \int_{t_r}^{t_b^u} \frac{1}{2N(\tau)} e^{-\int_{t_r}^{\tau} \frac{A(s)}{2N(s)} ds} d\tau$$

$$\mathbb{P}(\textit{tree-unchanged}|\mathcal{S},\mathcal{G},b) = \frac{1}{t_b^u - t_b^l} \int_{t_b^l}^{t_b^u} \mathbb{P}(\textit{tree-unchanged}|\mathcal{S},\mathcal{G},b,t)dt$$



MS-SMC' analytical solutions

$$\mathbb{P}(\textit{tree-unchanged}|\mathcal{S},\mathcal{G},b,t_r) = \int_{t_r}^{t_b^u} \frac{1}{2N(\tau)} e^{-\int_{t_r}^{\tau} \frac{A(s)}{2N(s)} ds} d\tau$$

$$\mathbb{P}(\textit{tree-unchanged}|\mathcal{S},\mathcal{G},b) = \frac{1}{t_b^u - t_b^l} \int_{t_b^l}^{t_b^u} \mathbb{P}(\textit{tree-unchanged}|\mathcal{S},\mathcal{G},b,t)dt$$

$$\mathbb{P}(\textit{tree-unchanged}|\mathcal{S},\mathcal{G}) = \sum_{b \in \mathcal{G}} \left\lceil \frac{t_b^u - t_b^l}{L(\mathcal{G})} \right\rceil \mathbb{P}(\textit{tree-unchanged}|\mathcal{S},\mathcal{G},b)$$

Exponentially distributed waiting distances

Expected number of sites until a recombination event is observed.

$$\lambda_r = L(\mathcal{G}) \times r$$

Exponentially distributed waiting distances

Expected number of sites until a recombination event is observed.

$$\lambda_r = L(\mathcal{G}) \times r$$

$$\lambda_n = L(\mathcal{G}) \times r \times \mathbb{P}(\textit{tree-unchanged}|\mathcal{S}, \mathcal{G})$$

Exponentially distributed waiting distances

Expected number of sites until a recombination event is observed.

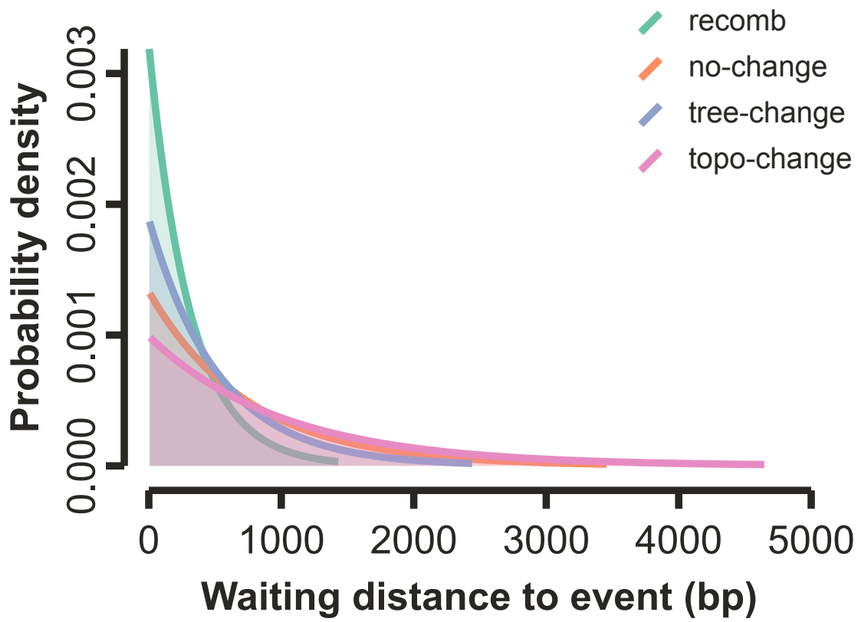
$$\lambda_r = L(\mathcal{G}) \times r$$

$$\lambda_n = L(\mathcal{G}) \times r \times \mathbb{P}(\textit{tree-unchanged}|\mathcal{S}, \mathcal{G})$$

$$\lambda_g = L(\mathcal{G}) \times r \times \mathbb{P}(\textit{tree-changed}|\mathcal{S}, \mathcal{G})$$

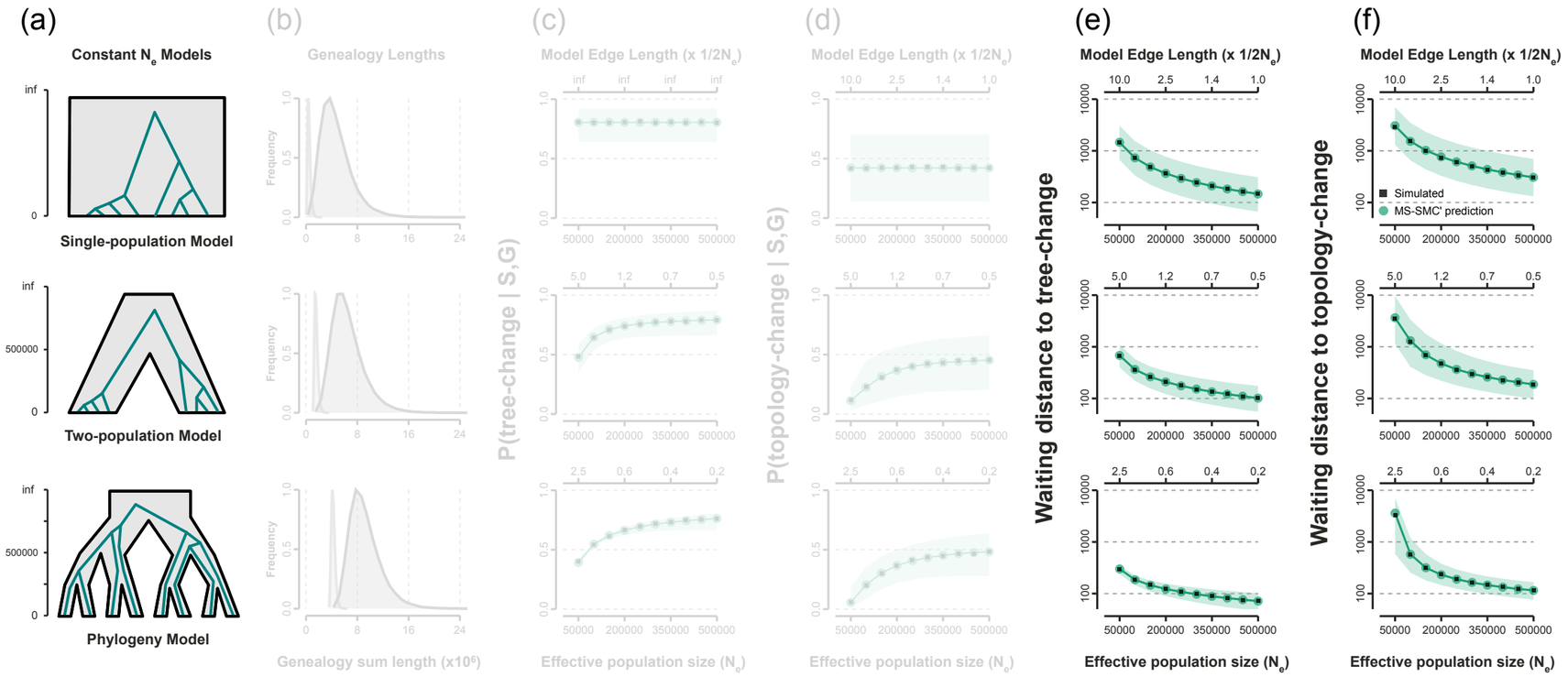
$$\lambda_t = L(\mathcal{G}) \times r \times \mathbb{P}(\textit{topology-changed}|\mathcal{S}, \mathcal{G})$$

Exponentially distributed waiting distances



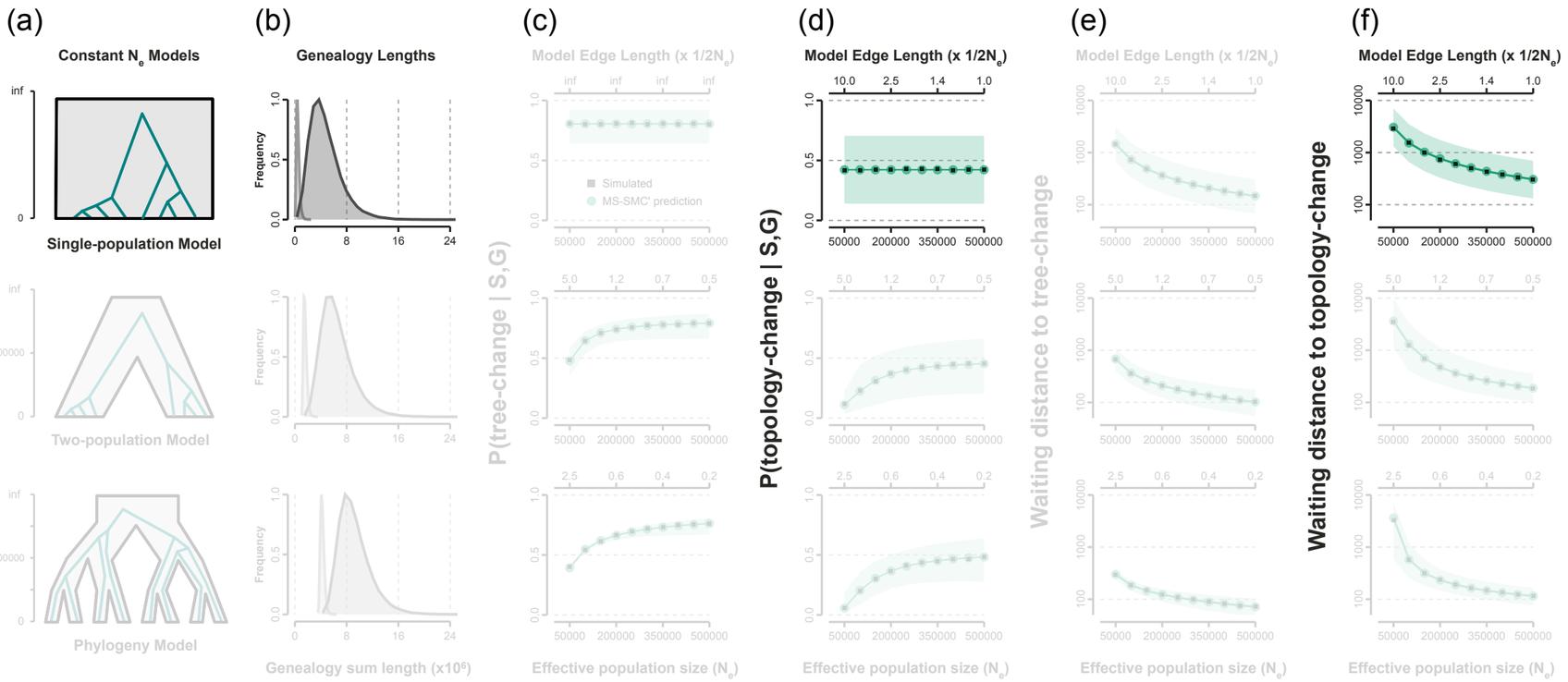
Validation:

Analytical results match expectation of stochastic coalescent simulations.



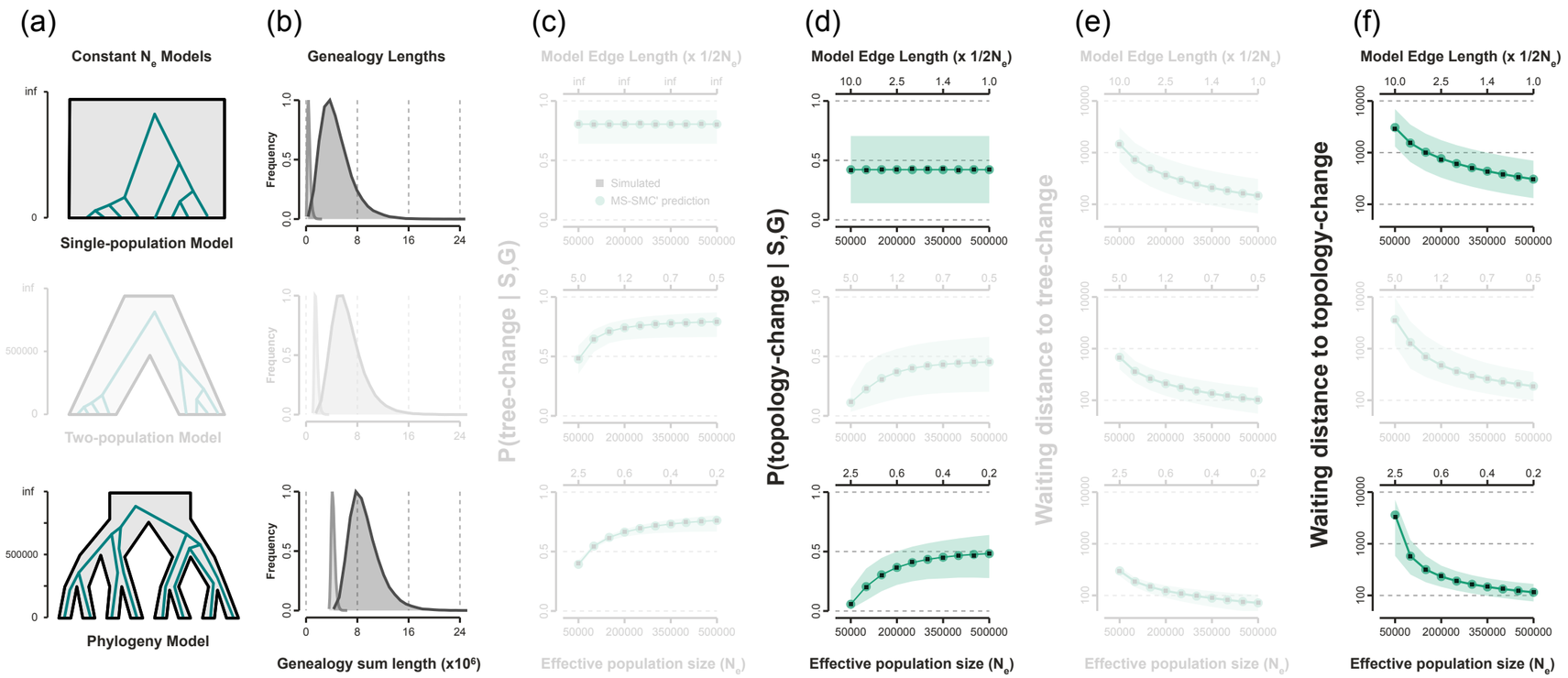
Validation:

In single population model (Deng et al.) N_e only affects edge lengths.



Validation:

In an MSC model N_e affects probability of tree/topology change as well as edge lengths.



Likelihood framework

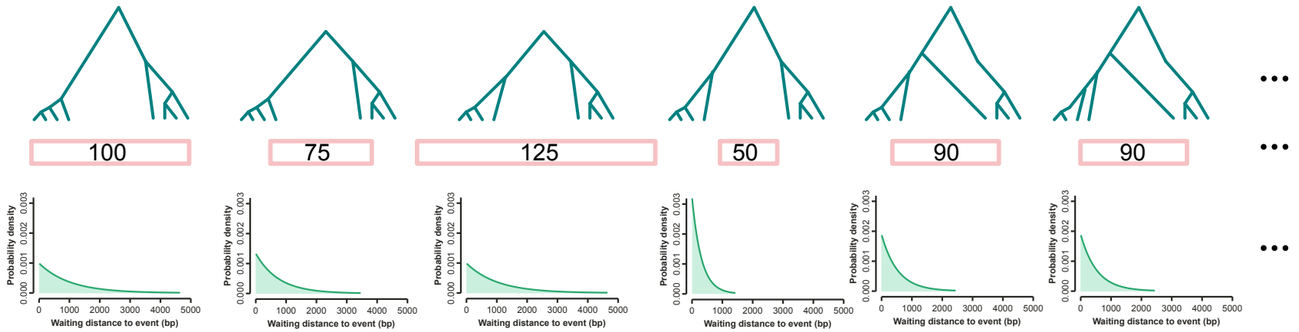
Given an observed/proposed ARG (genealogies and interval lengths)

get expected waiting distance for each (λ_i)...

$G = [\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3, \ldots]$

$X_g = [x_1, x_2, x_3, \ldots]$

$\Lambda_g = [\lambda_{g1}, \lambda_{g2}, \lambda_{g3}, \ldots]$



where: $\lambda_g = L(\mathcal{G}) \times r \times \mathbb{P}(\text{tree-unchanged}|\mathcal{S}, \mathcal{G})$

Likelihood framework

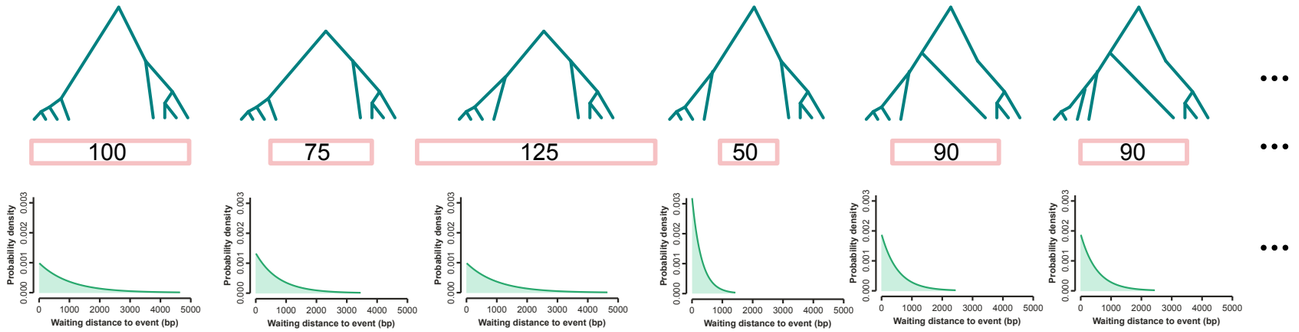
Given an observed/proposed ARG (genealogies and interval lengths)

get expected waiting distance for each (λ_i)...

$G = [\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3, \ldots]$

$X_g = [x_1, x_2, x_3, \ldots]$

$\Lambda_g = [\lambda_{g1}, \lambda_{g2}, \lambda_{g3}, \ldots]$



where: $\lambda_g = L(\mathcal{G}) \times r \times \mathbb{P}(\text{tree-unchanged}|\mathcal{S}, \mathcal{G})$

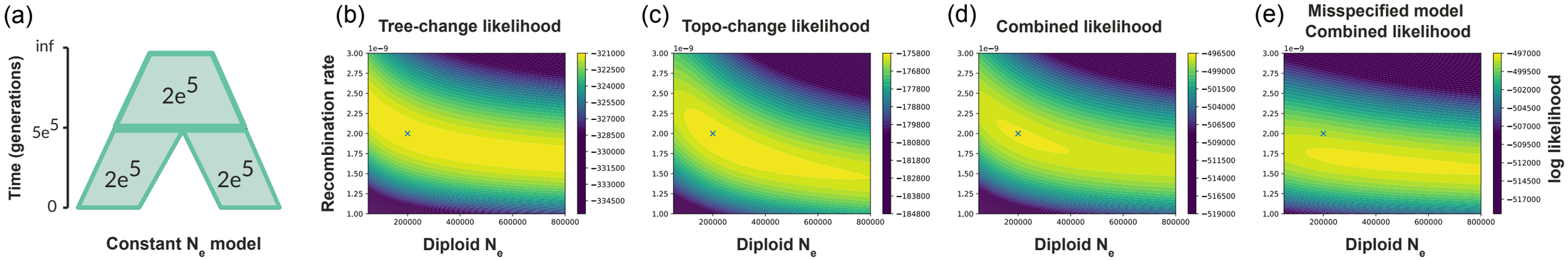
... and calculate likelihood of MSC model ($\mathcal{S}$) from exponential probability densities.

$$\mathcal{L}(\mathcal{S}|\Lambda_g, X_g) = \sum_i log(\lambda_i e^{-\lambda_i x_i})$$

Likelihood surface: single N_e

Topology-changes are more informative than tree-changes; optima at true sim. values.

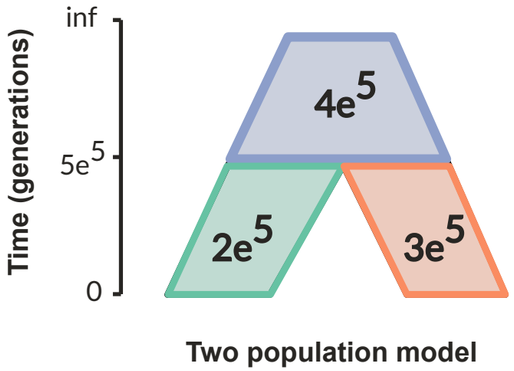
Example: loci=50, length=0.1Mb, recomb=2e-9, samples-per-lineage=4.

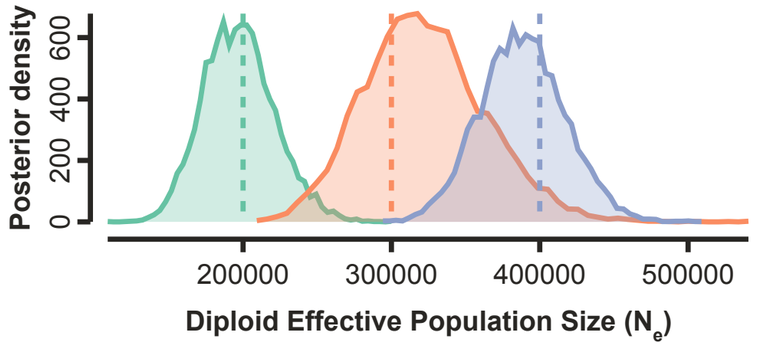


Joint inference of multiple MSC model parameters

Metropolis Hastings MCMC converges on correct w/ increasing data.

Example: loci=50, length=0.1Mb, recomb=2e-9, samples-per-lineage=4.





Summary: *Multispecies Sequentially Markov Coalescent*

- We extended method of Deng et al. (2021) to MSC models
- Analytical solutions for E[waiting distance] to tree or topology change

Summary: *Multispecies Sequentially Markov Coalescent*

- We extended method of Deng et al. (2021) to MSC models
- Analytical solutions for E[waiting distance] to tree or topology change
- Validated: accurate against stochastic coalescent simulations

Summary: *Multispecies Sequentially Markov Coalescent*

- We extended method of Deng et al. (2021) to MSC models
- Analytical solutions for E[waiting distance] to tree or topology change
- Validated: accurate against stochastic coalescent simulations
- MSC models predict more informative statistics (waiting distances) about linked genealogical variation than a single population model.

Summary: *Multispecies Sequentially Markov Coalescent*

- We extended method of Deng et al. (2021) to MSC models
- Analytical solutions for E[waiting distance] to tree or topology change
- Validated: accurate against stochastic coalescent simulations
- MSC models predict more informative statistics (waiting distances) about linked genealogical variation than a single population model.
- A big step towards estimating MSC models from *linked genome data!*

Summary: *Multispecies Sequentially Markov Coalescent*

- We extended method of Deng et al. (2021) to MSC models
- Analytical solutions for E[waiting distance] to tree or topology change
- Validated: accurate against stochastic coalescent simulations
- MSC models predict more informative statistics (waiting distances) about linked genealogical variation than a single population model.
- A big step towards estimating MSC models from *linked genome data!*
- Topology-changes are easily detactable in sequence data.

Future directions

- Manuscript on biorxiv (hopefully soon also in print)
- Implemented at https://github.com/eaton-lab/ipcoal/
- Software to analyze real genetic data is a future development
- Extensions to Multispecies Network Coalescent.
- and more...

Acknowledgements

Thanks to the symposium organizers and to:

Patrick McKenzie
PhD student

Eaton lab at Columbia University in New York City.
Contact us: @dereneaton

NSF DEB-2046813