

# CONTROLLED LUMPED COALESCENT MARKOV CHAINS FOR POPULATION GENOMIC INFERENCE

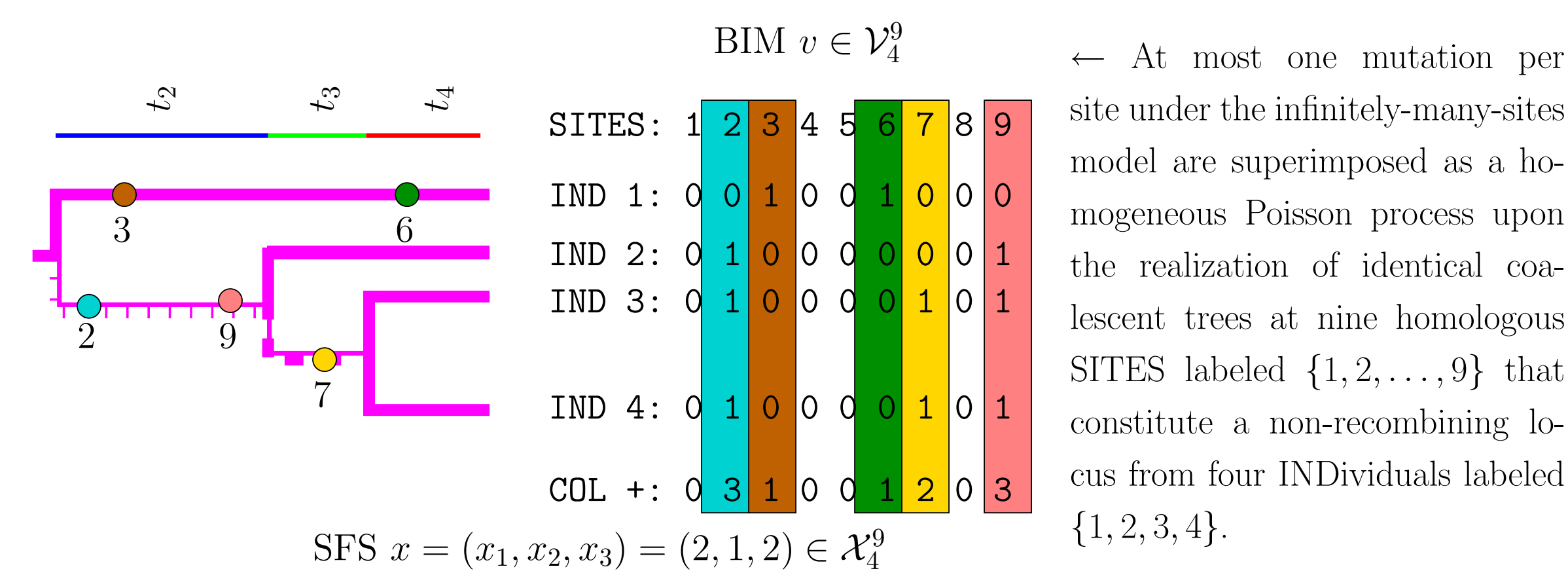
Raazesh Sainudiin

Biomathematics Research Centre, Department of Mathematics and Statistics, University of Canterbury, Christchurch, NZ

## Abstract

We derive the transition structure of a Markovian lumping of Kingman's  $n$ -coalescent [Kin82]. Lumping a Markov chain is meant in the sense of [KS60, def. 6.3.1]. The lumped Markov process, referred as the unlabeled  $n$ -coalescent, is a continuous-time Markov chain on the set of all integer partitions of the sample size  $n$ . We derive the backward-transition, forward-transition, state-specific, and sequence-specific probabilities of this chain. We show that the likelihood of any given site-frequency-spectrum (SFS), a commonly used statistics in genome scans, from a locus free of intra-locus recombination, can be directly obtained by integrating conditional realizations of the unlabeled  $n$ -coalescent. We develop a controlled Markov chain for importance sampling such integrals from an augmented unlabeled  $n$ -coalescent forward in time. We apply the methods to population-genetic data to conduct demographic inference at the empirical resolution of the site-frequency-spectra. We also extend a family of classical hypothesis tests of standard neutrality at a non-recombining locus based on any statistics of the SFS to a more powerful version that conditions on the topological information contained in the SFS. We formalize a graph of coalescent experiments to set a decision-theoretic stage for population genetic inference across different empirical resolutions.

## Data, Statistics and Likelihood

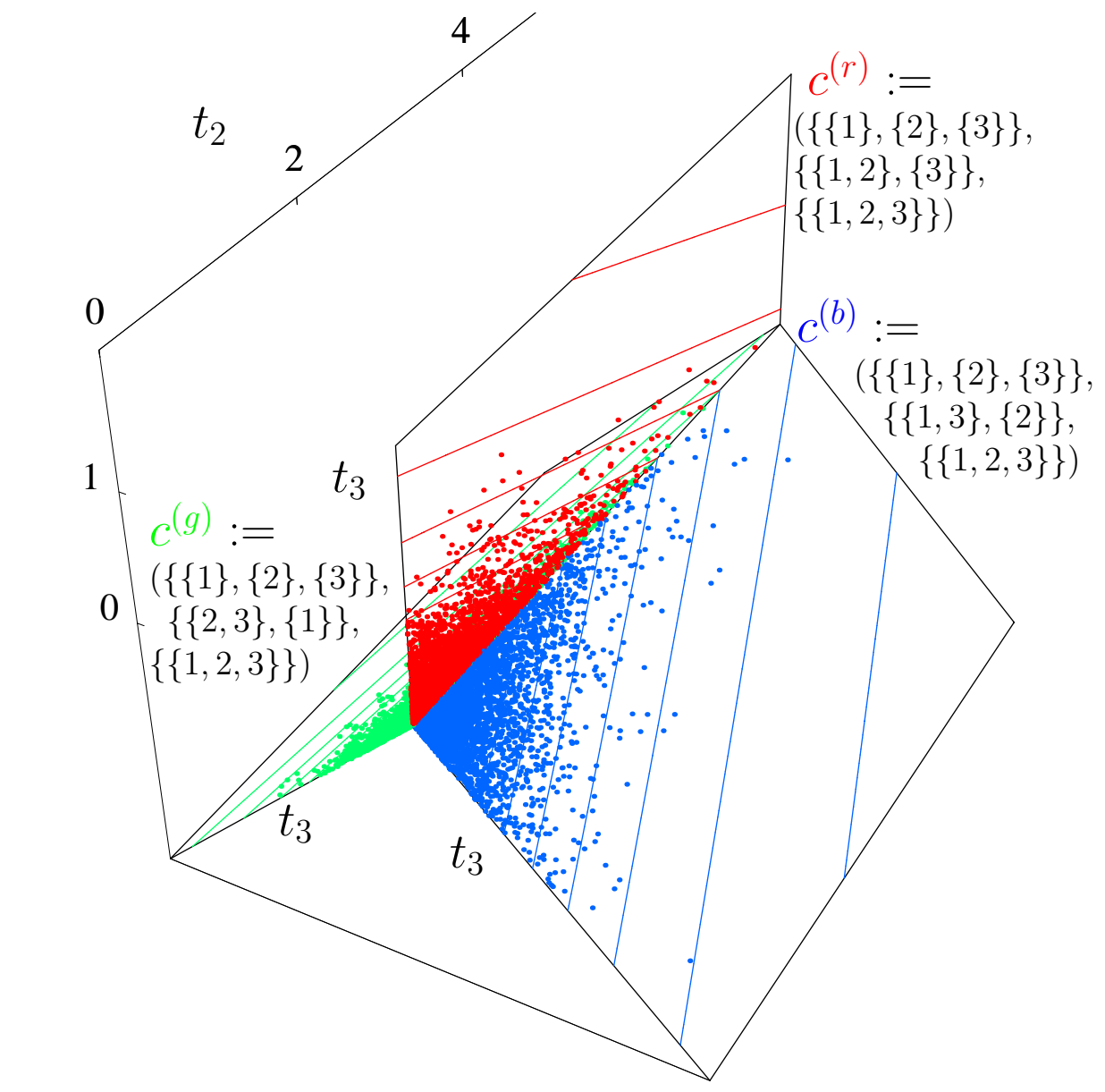
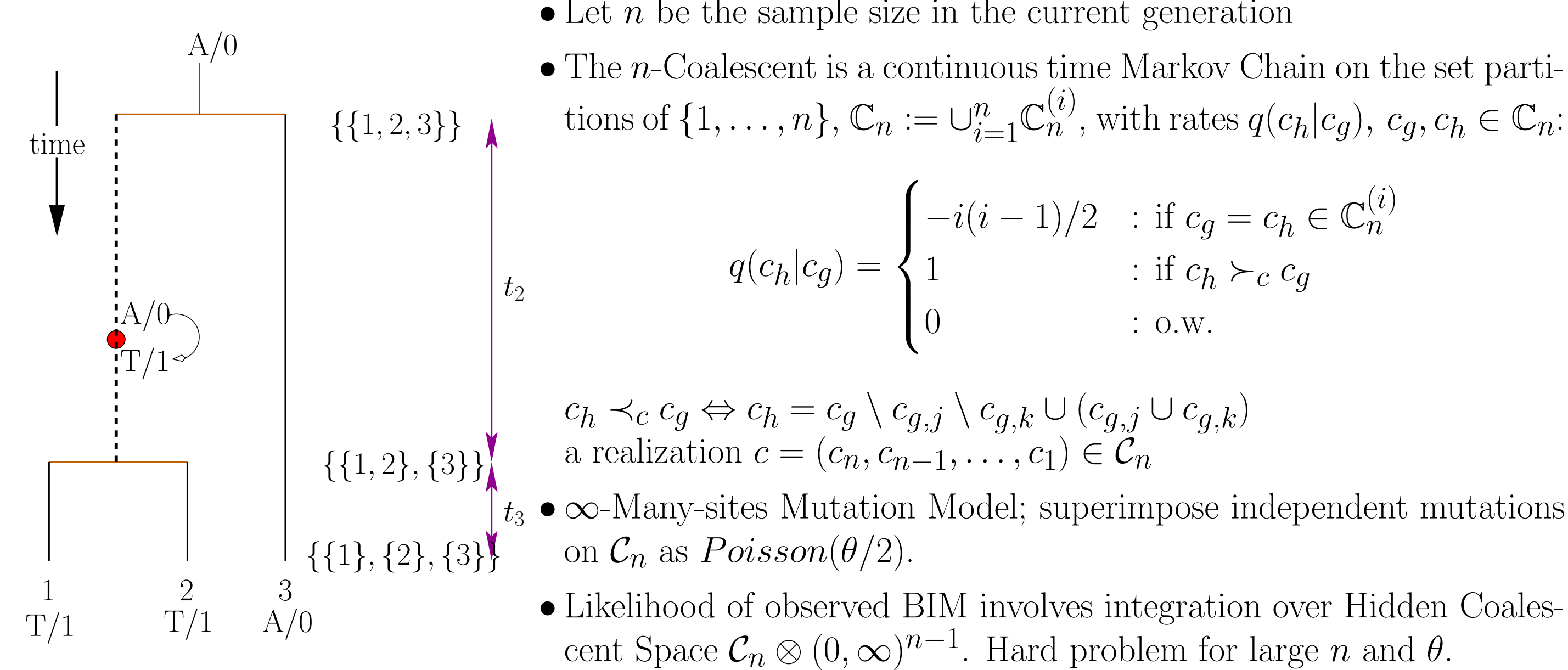


## Likelihood is computationally prohibitive at BIM

- Exact Methods: Complete Recursion in Griffiths' **PTREE** (1980). Even 1 Locus,  $\theta = 10$  ( $\rightarrow$  out of stack for  $n > 4$ ).
- Approximate Methods: Seq. Imp. Sampling in Griffiths' **GENETREE** (1994):  $L(\theta|v) \approx 4$  CPU hrs /  $\theta$
- The **Bottom Line**: Exact Genome Scanning at fine DNA resolution is currently impractical for  $n > 4$
- A **Solution**: Inference at coarser empirical resolutions, eg. **SFS** and its sub-experiments – **novel**
- Unlike "Likelihood-free" Methods called Approximate Bayesian Computations we do it exactly with controlled lumped coalescent Markov chains.

## $n$ -Coalescent Probability Models

**Kingman's Labeled  $n$ -Coalescent** [Kin82] with parameter  $\theta = 4N_e\mu$  (scaled mutation rate) is an approximation of the Wright-Fisher (W-F) Model.



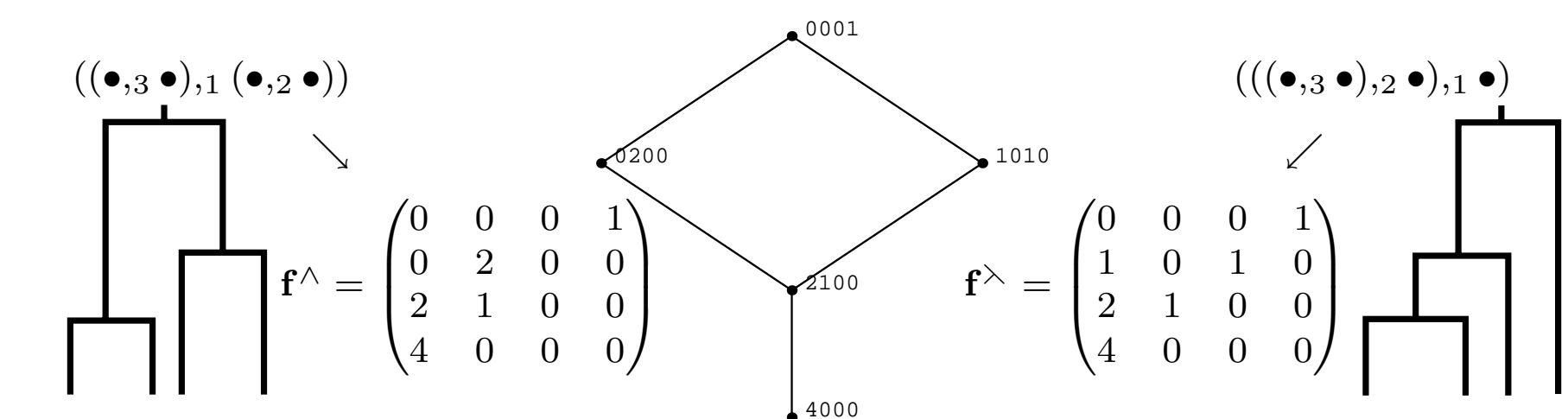
**Kingman's Unlabeled  $n$ -Coalescent** with parameter  $\theta = 4N_e\mu$  (scaled mutation rate) is an approximation of a Lumped Wright-Fisher (W-F) Model.

- The Unlabeled  $n$ -Coalescent is the continuous time Markov chain  $\{F^\uparrow(t)\}_{t \in \mathbb{R}_+}$  on  $\mathbb{F}_n$  the set of integer partitions of  $n$ ,  $\mathbb{F}_n := \cup_{i=1}^n \mathbb{F}_n^{(i)}$ , [SS09] whose rate matrix  $Q = q(f_{i'}|f_i)$  for any two states  $f_i, f_{i'} \in \mathbb{F}_n$  is:

$$q(f_{i'}|f_i) = \begin{cases} -i(i-1)/2 & \text{if } \mathbb{F}_n^{(i)} \ni f_i = f_{i'}, \\ f_{i,j}f_{i,k} & \text{if } \mathbb{F}_n^{(i-1)} \ni f_{i'} = f_i - e_j - e_k + e_{j+k}, j \neq k, f_i \in \mathbb{F}_n^{(i)}, \\ (f_{i,j})(f_{i,j} - 1)/2 & \text{if } \mathbb{F}_n^{(i-1)} \ni f_{i'} = f_i - e_j - e_k + e_{j+k}, j = k, f_i \in \mathbb{F}_n^{(i)}, \\ 0 & \text{otherwise} \end{cases}$$

The initial state is  $f_n = (n, 0, 0, \dots, 0)$  and the final absorbing state is  $f_1 = (0, 0, \dots, 1)$ . A realization  $f = (f_n, f_{n-1}, \dots, f_1) \in \mathcal{F}_n$  is an  $f$ -sequence.

- Likelihood of observed SFS under  $\infty$ -Many-sites Mutation Model involves integration over the Hidden Unlabeled Coalescent Space  $\mathcal{F}_n \otimes (0, \infty)^{n-1}$ . Easier problem for large  $n$  and  $\theta$ .



↑ The two  $f$ -sequences  $F^\wedge$  and  $F^\lambda$  the state transition diagrams of  $\{F^\uparrow(k)\}_{k \in [n]_-}$  and  $\{F^\downarrow(k)\}_{k \in [n]_+}$  on  $\mathbb{F}_n$  (middle panel).

## Likelihood of a Site Frequency Spectrum

Let  $c_t \in \mathcal{C}_n \mathbb{T}_n$  be a given coalescent tree,  $c$  be its  $c$ -sequence,  $f = F(c)$  be its  $f$ -sequence,  $t := (t_2, t_3, t_n) \in (0, \infty)^{n-1}$  be its epoch times and let

$$L^c(t) = l := (l_1, \dots, l_{n-1}) = t^\top f = \left( \sum_{i=2}^n t_i f_{i,1}, \dots, \sum_{i=2}^n t_i f_{i,n-1} \right), \quad l_\bullet \equiv \sum_{i=2}^n l_i, \quad \bar{l}_i \equiv \frac{l_i}{l_\bullet}$$

be its lineage lengths subtending  $1, 2, \dots, n-1$  leaves, the total tree-size, and relative lineage lengths respectively.

$$P(x|\phi, a) = P(x|\phi, l = t^\top f) = e^{-\theta l_\bullet} (\theta l_\bullet)^S \prod_{i=1}^{n-1} \bar{l}_i^{x_i} / \prod_{i=1}^{n-1} x_i!$$

$$P(x|\phi, a) = P(x|\phi, l = t^\top f) = e^{-\theta l_\bullet} (\theta l_\bullet)^S \prod_{i=1}^{n-1} \bar{l}_i^{x_i} / \prod_{i=1}^{n-1} x_i!$$

$$P(x|\phi) = \frac{1}{\prod_{i=1}^{n-1} x_i!} \sum_{f \in \mathcal{F}_n^*(x^*)} P(f) \left( \int_{t \in (0, \infty)^{n-1}} \left( e^{-\theta l_\bullet} (\theta l_\bullet)^S \prod_{i=1}^{n-1} \bar{l}_i^{x_i} \right) P(t|\phi) \right)$$

where,  $F_n(x^*) \equiv \bigcup_{\{k: x_k^* = 1\}} \{f \in \mathcal{F}_n : \sum_{i=1}^n f_{i,h} = 0\}$

$$X^*(x) = x^* \equiv (x_1^*, \dots, x_{n-1}^*) \equiv (\mathbf{1}_N(x_1), \dots, \mathbf{1}_N(x_{n-1})) \in \{0, 1\}^{n-1}$$

$$P((f_{i'}, z_{i'})|(f_i, z_i)) = \begin{cases} P(f_{i'}|f_i)/\Sigma(f_i, z_i) & \text{if } (f_i, z_i) \prec_{f,z} (f_{i'}, z_{i'}) \\ 0 & \text{otherwise} \end{cases}$$

where,

$$\Sigma(f_i, z_i) = \sum_{(j,k) \in \Xi(f_i, z_i)} P(f_i - e_{j+k} + e_j + e_k | f_i),$$

$$\Xi(f_i, z_i) := \{(j, k) : f_{i,j+k} > 0, 1 \leq j \leq \hat{j} \leq k \leq j+k-1\},$$

$$\hat{j} := \max\{\min\{\max\{\ell : z_{i,\ell} = 1\}, j+k-1\}, \lfloor \frac{j+k}{2} \rfloor\},$$

$$(f_i, z_i) \prec_{f,z} (f_{i'}, z_{i'}) \Leftrightarrow \begin{cases} f_{i'} = f_i + e_j + e_k - e_{j+k}, (j, k) \in \Xi(f_i, z_i) \text{ and} \\ z_{i'} = z_i - \mathbf{1}_{\{j\}}(z_{i,j}) e_j - \mathbf{1}_{\{k\}}(z_{i,k}) e_k \end{cases}$$

and with  $(f_n, (0, 0, \dots, 0)) = ((n, 0, \dots, 0), (0, 0, \dots, 0))$  as the final absorbing state.

Let  $\mathcal{F}_n^{x^*}$  be the set of sequential realizations of the first component of the ordered pairs of states visited by  $\{F^\uparrow(x^*(k))\}_{k \in [n]_+}$ , i.e.

$$\mathcal{F}_n^{x^*} := \{f = (f_n, f_{n-1}, \dots, f_1) : f_i \in \mathbb{F}_n^{(i)}, (f_i, z_i) \prec_{f,z} (f_{i+1}, z_{i+1}), z_1 = x^*\}.$$

Then  $\mathcal{F}_n^{x^*} = \mathcal{C}F_n(x^*)$ .

## SFS Proposal by an $x^*$ -controlled unlabeled $n$ -coalescent

1: **input**:

- scaled mutation rate  $\phi_1$  of the locus
- observed  $x^*$  (note that sample size  $n = |x^*| + 1$ )

2: **output**: an SFS sample  $x$  such that the underlying  $f$ -sequence  $f \in \mathcal{C}F_n(x^*)$

- generate an  $f$ -sequence  $f$  under  $\{F^\uparrow(x^*(k))\}_{k \in [n]_+}$
- draw  $t \sim T = (T_2, T_3, \dots, T_n) \sim \otimes_{i=2}^n \binom{2}{j} e^{-\binom{j}{2} t_i}$ , or as desired
- $l = t^\top \cdot f$ , where  $f = \mathbf{F}(f)$
- draw  $x$  from Poisson-Multinomial distribution  $e^{-\phi_1 l_\bullet} (\phi_1 l_\bullet)^S \prod_{i=1}^{n-1} \bar{l}_i^{x_i} / \prod_{i=1}^{n-1} x_i!$
- return**:  $x$

**SFS-based Estimator of  $\phi_1^* = \theta^*$  and  $\phi_2^*$  (exp. growth rate)**

| $n$ | $\hat{\phi}_2$ | $\hat{\phi}_1$ | $C_{99\%}$ | $C_{99\%}$ | $Qrt(K)$ |
|-----|----------------|----------------|------------|------------|----------|
| 4   | 46             | 30             | 42         | 43         | 30       |
| 5   | 32             | 19             | 42         | 31         | 22       |
| 6   | 31             | 18             | 41         | 35         | 23       |
| 7   | 34             | 19             | 48         | 32         | 20       |
| 8   | 26             | 12             | 66         | 21         | 11       |
| 9   | 27             | 12             | 65         | 18         | 10       |
| 10  | 23             | 11             | 64         | 17         | 10       |

**The Forward Markov chain  $\{F^\downarrow(k)\}_{k \in [n]_+}$  on  $\mathbb{F}_n$ ,  $[n]_+ := \{1, 2, \dots, n\}$ :**

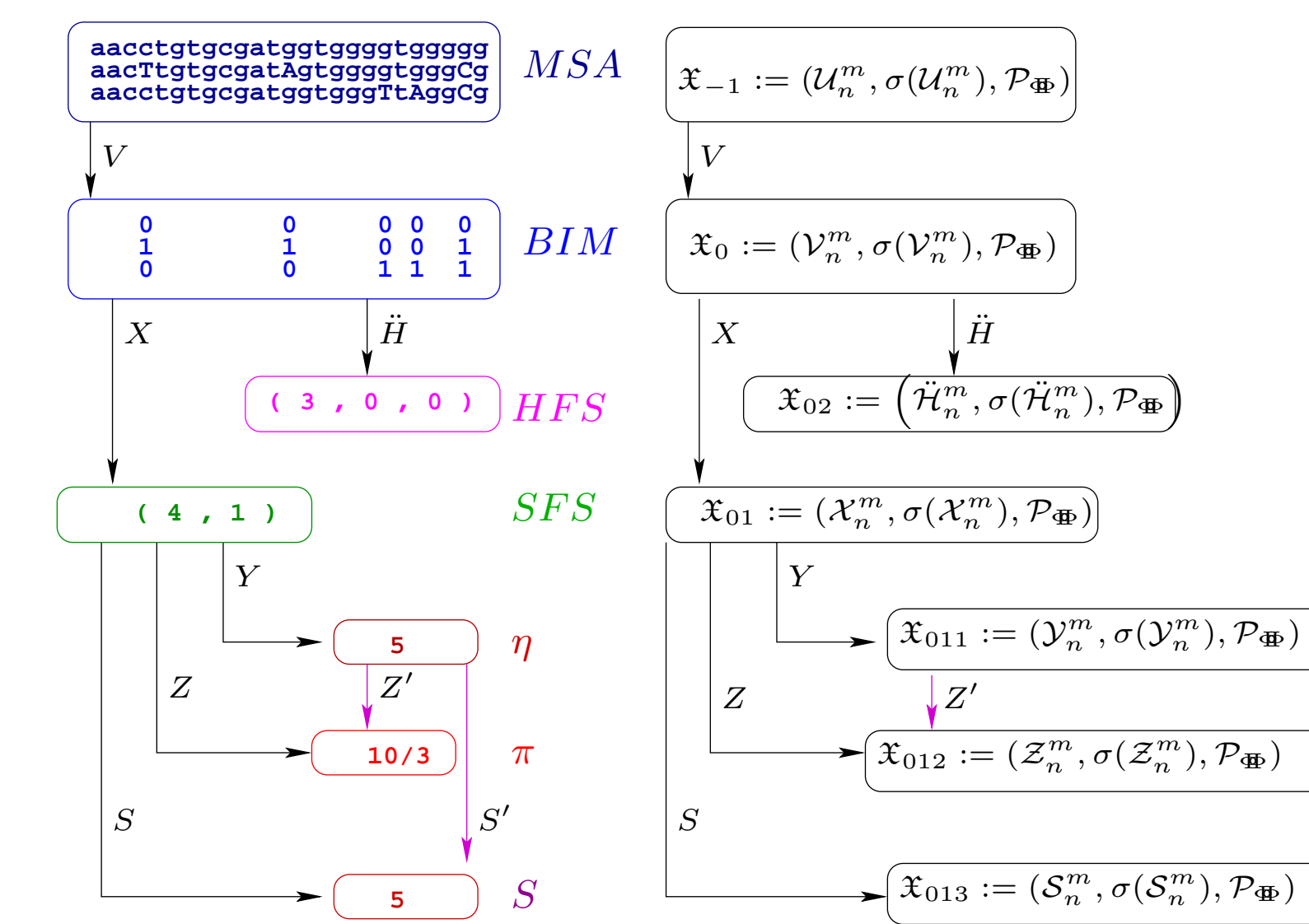
$$P(f_i|f_{i-1}) = \begin{cases} 2f_{i-1,j+k}(n-i+1)^{-1} & \text{if } f_i = f_{i-1} + e_j + e_k - e_{j+k}, j \neq k, \\ & j+k > 1, f_i \in \mathbb{F}_n^{(i)}, f_{i-1} \in \mathbb{F}_n^{(i-1)} \\ f_{i-1,j+k}(n-i+1)^{-1} & \text{if } f_i = f_{i-1} + e_j + e_k - e_{j+k}, j = k, \\ & j+k > 1, f_i \in \mathbb{F}_n^{(i)}, f_{i-1} \in \mathbb{F}_n^{(i-1)} \\ 0 & \text{otherwise} \end{cases}$$

with initial state  $f_1 = (0, 0, \dots, 1)$  and final absorbing state  $f_n = (n, 0, \dots, 0)$ .

## Controlled Lumped Coalescent Proposal Chain over $\mathcal{C}F_n(x^*)$

For a given SFS  $x \in \mathcal{X}_n^m$  and  $X^*(x) = x^* \in \{0, 1\}^{n-1}$ , consider the discrete time Markov chain  $\{F^\downarrow(x^*(k))\}_{k \in [n]_+}$  over the state space of ordered pairs  $(f_{i'}, z_{i'}) \in \mathbb{F}_n^{x^*} \subset \mathbb{F}_n \times \{0, 1\}^{n-1}$ , with the initial state given by  $(f_1, z_1^*) = ((0, 0, \dots, 1), x^*)$ , the transition probabilities obtained by a controlled reweighing of the the transition probabilities of  $\{F^\downarrow(k)\}_{k \in [n]_+}$  over  $\mathbb{F}_n$  as follows:

## An $n$ -Coalescent Experiments Graph



## Definition 1 (The Experiments Graph)

Consider an  $\mathfrak{A}$ -indexed set of experiments  $\{X_\alpha, \alpha \in \mathfrak{A}\}$ . Let,  $T_{\alpha,\beta} : Z_\alpha \rightarrow Z_\beta$ , for some  $\alpha, \beta \in \mathfrak{A}$  with  $\sigma(Z_\alpha) \supset \sigma(Z_\beta)$  be a statistic (measurable map). Let  $\mathfrak{M}$  be a set of such maps as well as the identity map. Then, the directed graph of experiments  $\mathfrak{G}_{\mathfrak{A}, \mathfrak{M}}$  with nodes  $\{X_\alpha, \alpha \in \mathfrak{A}\}$  and directed edges from a node  $X_\alpha$  to a node  $X_\beta$ , provided there exists an  $T_{\alpha,\beta} \in \mathfrak{M}$ , is the experiments graph. Consider the partial ordering  $\succ_x$  induced on the experiments in  $\{X_\alpha, \alpha \in \mathfrak{A}\}$  by the maps in  $\mathfrak{M}$ , i.e.,  $X_\alpha \succ_x X_\beta$  if and only if there exists a composition of maps from  $\mathfrak{M}$  given by  $T_{\alpha,\beta} := T_{\alpha,\beta} \circ T_{i,j} \circ \dots \circ T_{i',j'} : Z_\alpha \rightarrow Z_\beta$ , such that  $\sigma(Z_\alpha) \supset \sigma(Z_\beta)$ . Then, by construction, (1) the random variables  $\{X_\alpha, \alpha \in \mathfrak{A}\}$  that are adapted to this partially ordered filtration, i.e., for each  $\alpha \in \mathfrak{A}$ ,  $X_\alpha$  is  $\sigma(X_\alpha)$ -measurable, such that (2)  $E(\{X_\alpha\}) < \infty$  for all  $\alpha \in \mathfrak{A}$ , form a martingale relative to  $\mathcal{P}_\mathfrak{M}$  and the partially ordered filtration on  $\mathfrak{G}_{\mathfrak{A}, \mathfrak{M}}$ , i.e.,  $E(X_\alpha | \sigma(X_\beta)) = X_\beta$ , provided  $X_\alpha \succ_x X_\beta$ .

This allows for theory of approximate sufficiency [Cam64] over a partially-ordered family of statistical  $n$ -coalescent experiments whose hidden space is given by a corresponding graph of lumped  $n$ -coalescent Markov chains [SS09].

## Acknowledgments

This work is supported by a research fellowship from the Royal Commission for the Exhibition of 1851 to R.S. This is joint work with James Booth, Peter Donnelly, Robert Griffiths, Gilean McVean, Tanja Stadler, Michael Stillman and Kevin Thornton.

## References

- [Cam64] L. Le Cam. Sufficiency and approximate sufficiency. *Ann. Math. Stats.*, 35:1419–1455, 1964.
- [Kin82] JFC Kingman. The coalescent. *Stochastic Processes and their Applications*, 13:235–248, 1982.
- [KS60] Kemeny and Snell. *Finite Markov Chains*. D. Van Nostrand Co., 1960.
- [SS09] R Sainudiin and T Stadler. A unified multi-resolution coalescent: Markov lumpings of the Kingman-Tajima  $n$ -coalescent – UCDS Research Report 2009/4, april 5, 2009 (submitted). Available at <http://www.math.canterbury.ac.nz/~r.sainudiin/preprints/SixCoal.pdf>, 2009.