ABCDE Approximate Bayesian Computations Done Exactly: Experiments with the Site Frequency Spectrum

Raazesh Sainudiin[†,′]

[†]Biomathematics Research Centre and
[′]Department of Mathematics and Statistics
University of Canterbury, Christchurch, New Zealand
http://www.math.canterbury.ac.nz/~r.sainudiin/

Joint with: Jim Booth[•], Peter Donnelly[‡], Bob Griffiths[‡], Jenny Harlow[′], Gil McVean[‡], Mike Stillman[⋆] and Kevin Thornton[®]

[®]Biological Sciences, University of California, Irvine, USA
Department of [•]Biological Statistics and Computational Biology and
[⋆]Department of Mathematics, Cornell University, Ithaca, USA
[‡]Department of Statistics, University of Oxford, Oxford, UK

ANR MANEGE Stochastic Models in Ecology, Genetics and Evolution Research Meeting, École Polytechnique, Palaiseau, France, Tuesday February 8, 2011

- **Tradition**: Modern European Empiricism (English Roots)

- **Tradition**: Modern European Empiricism (English Roots)
- **Internal Consistency** : Aristotelean Logic(s)

- **Tradition**: Modern European Empiricism (English Roots)
- **Internal Consistency** : Aristotelean Logic(s)
- **Universe of Hypotheses**: Popper's Falsifiability

The Dualistic Context ("The Bigger Picture")

- **Tradition**: Modern European Empiricism (English Roots)
- **Internal Consistency** : Aristotelean Logic(s)
- **Universe of Hypotheses**: Popper's Falsifiability
- **Empirical Resolution** : n Human DNA Seqns – Data $\mathcal{D}_o$

The Dualistic Context ("The Bigger Picture")

- **Tradition**: Modern European Empiricism (English Roots)
- **Internal Consistency** : Aristotelean Logic(s)
- **Universe of Hypotheses**: Popper's Falsifiability
- **Empirical Resolution** : n Human DNA Seqns – Data $\mathcal{D}_o$
- **Objective** : Pop. Genet. Parameter Inference

The Dualistic Context ("The Bigger Picture")

- **Tradition**: Modern European Empiricism (English Roots)
- **Internal Consistency** : Aristotelean Logic(s)
- **Universe of Hypotheses**: Popper's Falsifiability
- **Empirical Resolution** : n Human DNA Seqns – Data $\mathcal{D}_o$
- **Objective** : Pop. Genet. Parameter Inference
- **Approach**: Statistical Decision Theory

The Dualistic Context ("The Bigger Picture")

- **Tradition**: Modern European Empiricism (English Roots)
- **Internal Consistency** : Aristotelean Logic(s)
- **Universe of Hypotheses**: Popper's Falsifiability
- **Empirical Resolution** : n Human DNA Seqns – Data $\mathcal{D}_o$
- **Objective** : Pop. Genet. Parameter Inference
- **Approach**: Statistical Decision Theory
- **Prob. Models** : Neutral Coalescent w/ Demography

The Dualistic Context ("The Bigger Picture")

- **Tradition**: Modern European Empiricism (English Roots)
- **Internal Consistency** : Aristotelean Logic(s)
- **Universe of Hypotheses**: Popper's Falsifiability
- **Empirical Resolution** : n Human DNA Seqns – Data $\mathcal{D}_o$
- **Objective** : Pop. Genet. Parameter Inference
- **Approach**: Statistical Decision Theory
- **Prob. Models** : Neutral Coalescent w/ Demography
- **Engineering Constraints**: Resource-limited Info. Proc.

The Dualistic Context ("The Bigger Picture")

- **Tradition**: Modern European Empiricism (English Roots)
- **Internal Consistency** : Aristotelean Logic(s)
- **Universe of Hypotheses**: Popper's Falsifiability
- **Empirical Resolution** : n Human DNA Seqns – Data $\mathcal{D}_o$
- **Objective** : Pop. Genet. Parameter Inference
- **Approach**: Statistical Decision Theory
- **Prob. Models** : Neutral Coalescent w/ Demography
- **Engineering Constraints**: Resource-limited Info. Proc.
- **Solution**: Approximate Inference from Summaries of $\mathcal{D}_o$

Outline

- Wright-Fisher Model – vanilla version
- The n-Coalescent Approximation
- Computationally Intensive Likelihoods
- A Paritllay-ordered Coalescent Experiments Graph
- Unlabeled n-Coalescent
- Likelihood of SFS
- Controlled Lumped Coalescent
- Results
- Summary
- Acknowledgments

A Population of $N = 10$ homologous DNA seqns. of length m and the Population History of site i

```
  : 1 2 3 4 5 6 7 8 9 10
1 : A A A A A A A A A C
2 : G G G G G G G G G G
...
i : T T A A A A A A A A
...
k : ...
```
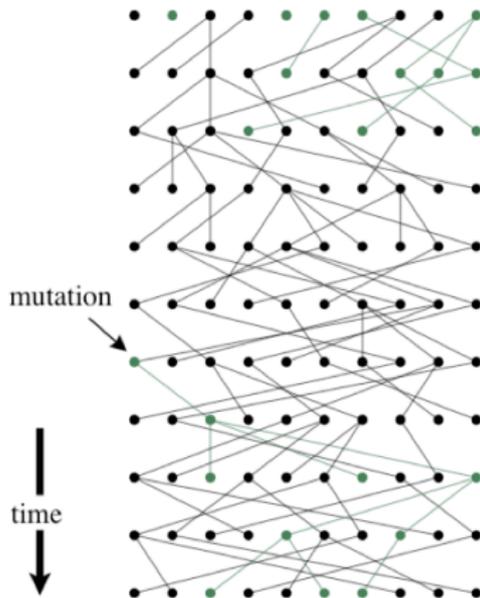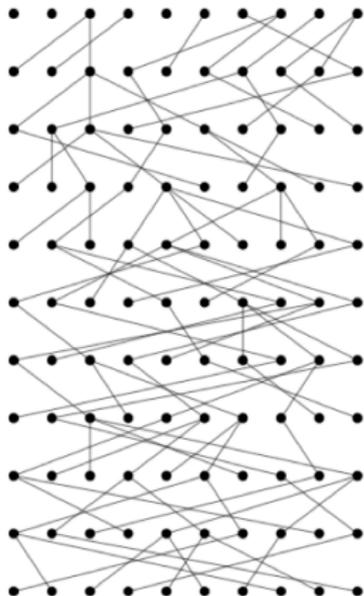

mutation

time

Ex: Data of 3 homologous DNA sequences at site i, its Population History and the Sample History of sampled individuals 1,2, and 3.

```
  : 1 2 3
i : T T A
```
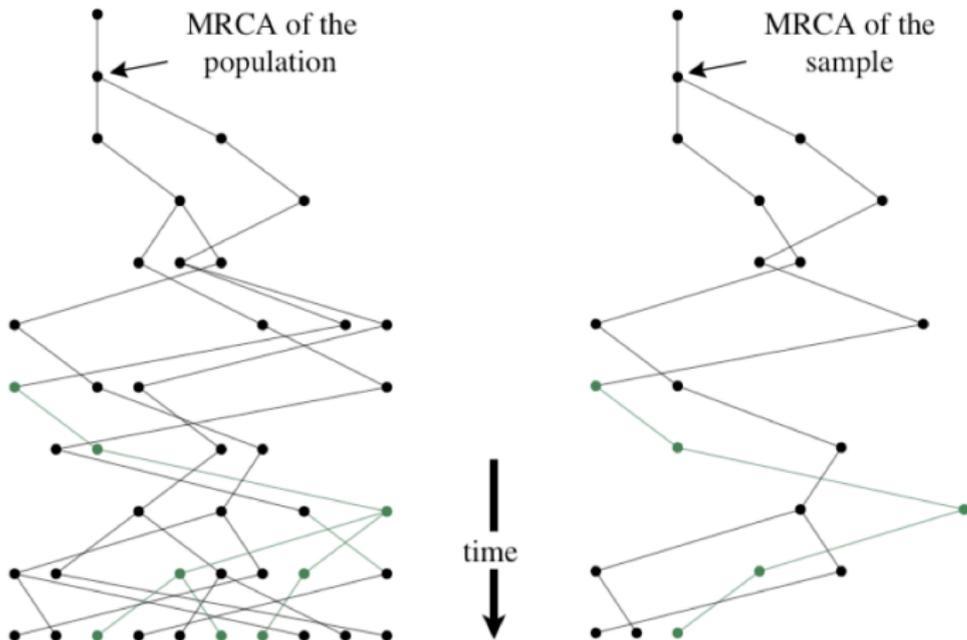

MRCA of the population

MRCA of the sample

time

A Sample Coalescent Sequence or c-sequence ($\{\{1\}, \{2\}, \{3\}\}, \{\{1,2\}, \{3\}\}, \{\{1,2,3\}\}$)

and coalescent times or epoch times $t_i, i \in \{3, 2\}$.

- Offspring "choose" parents uniformly and independently in W-F model

- $\Pr(2 \text{ lineages coalesce in } 1 \text{ generation}) = 1/N$

- $\Pr(2 \text{ lins. are distinct} > g \text{ gens.}) = (1 - 1/N)^g$

- Rescaled time t is g in units of N gens. Then, $\Pr(2 \text{ lins. remain distinct} > t)$ is

$$(1 - 1/N)^{\lfloor Nt \rfloor} \stackrel{N \to \infty}{\longrightarrow} e^{-t}$$

- Lineage Death Process: In general, the R.V. T_i that any pair of i lineages coalesce is approximately exponentially distributed for large N.

$$T_i \sim \text{Exponential}\left(\binom{i}{2}\right)$$

- Uniform Binary Fusion of two extant lineages.
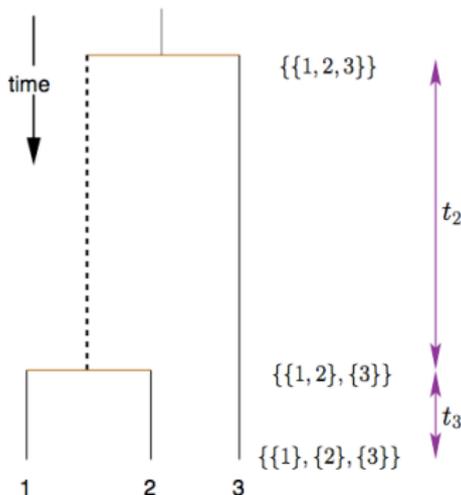
The Coalescent Approximation of the Wright-Fisher (W-F) Model (Kingman, 1982)

- The n-Coalescent is a continuous time Markov Chain on $\mathbb{C}_n \equiv \cup_{i=1}^n \mathbb{C}_n^i$, the set partitions of $\{1, \dots, n\}$, with rates $q(c_h | c_g)$, $c_g, c_h \in \mathbb{C}_n$:

$$q(c_h | c_g) = \begin{cases} -i(i-1)/2 & : \text{if } c_g = c_h \in \mathbb{C}_n^i \\ 1 & : \text{if } c_h \succ_c c_g \\ 0 & : \text{o.w.} \end{cases}$$

$$c_h \prec_c c_g \Leftrightarrow c_h = c_g \setminus c_{g,j} \setminus c_{g,k} \cup (c_{g,j} \cup c_{g,k})$$

a realization $c = (c_n, c_{n-1}, \dots, c_1) \in \mathcal{C}_n$

- Superimpose indep. mutations $\sim Poisson(\theta/2 \equiv 2N\mu)$

∞-many-sites mutation model
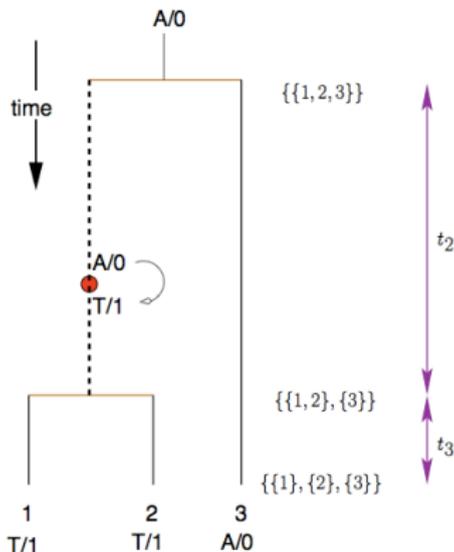
One Parameter: $\phi := (\theta) \in \Phi$, $\theta = 4N_e\mu$

$\mathscr{C}_3\mathbb{T}_3 := \mathscr{C}_3 \otimes \mathbb{T}_3 = \mathscr{C}_3 \otimes (0, \infty)^2$

Two Parameters: $\phi := (\theta, \nu) \in \Phi$, $\theta = 4N_e\mu$
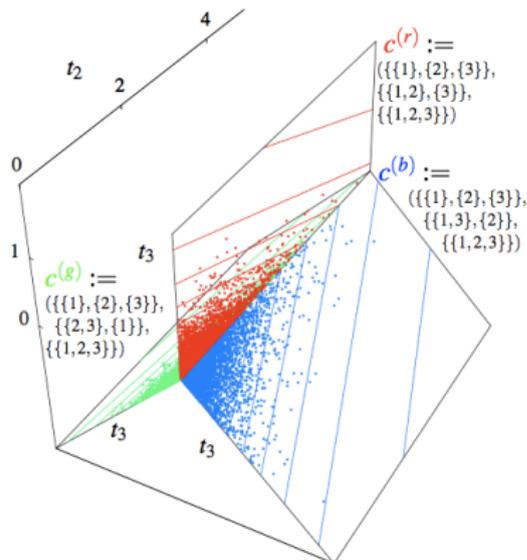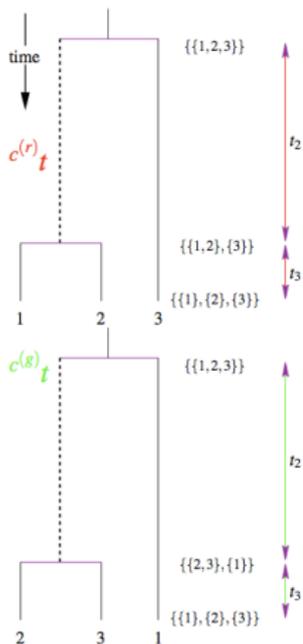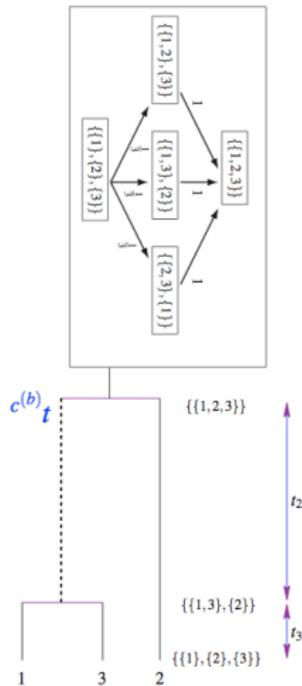
constant size
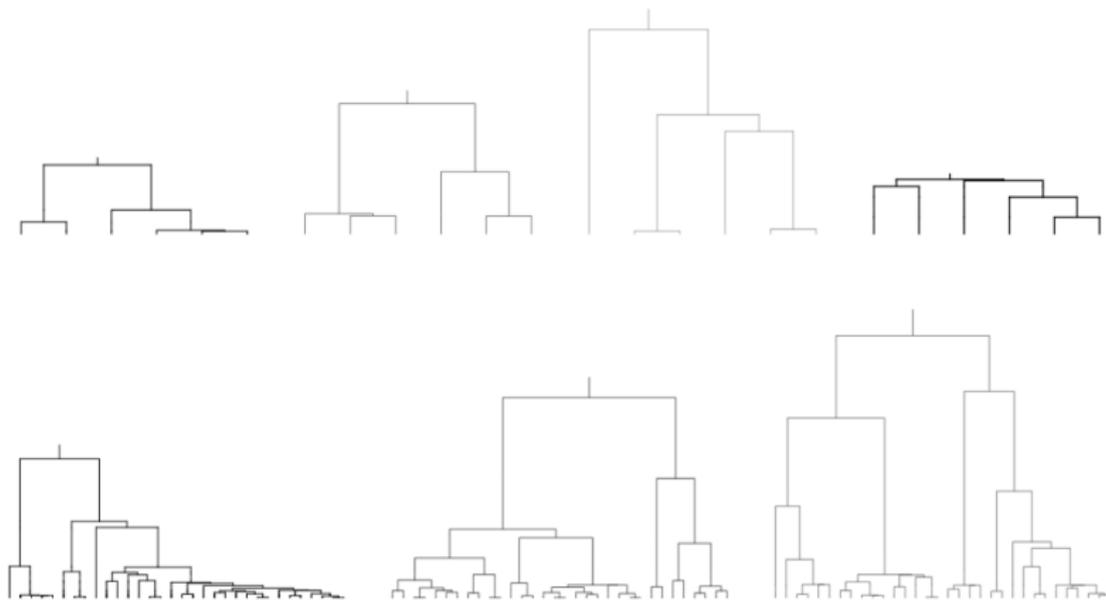
exponential growth
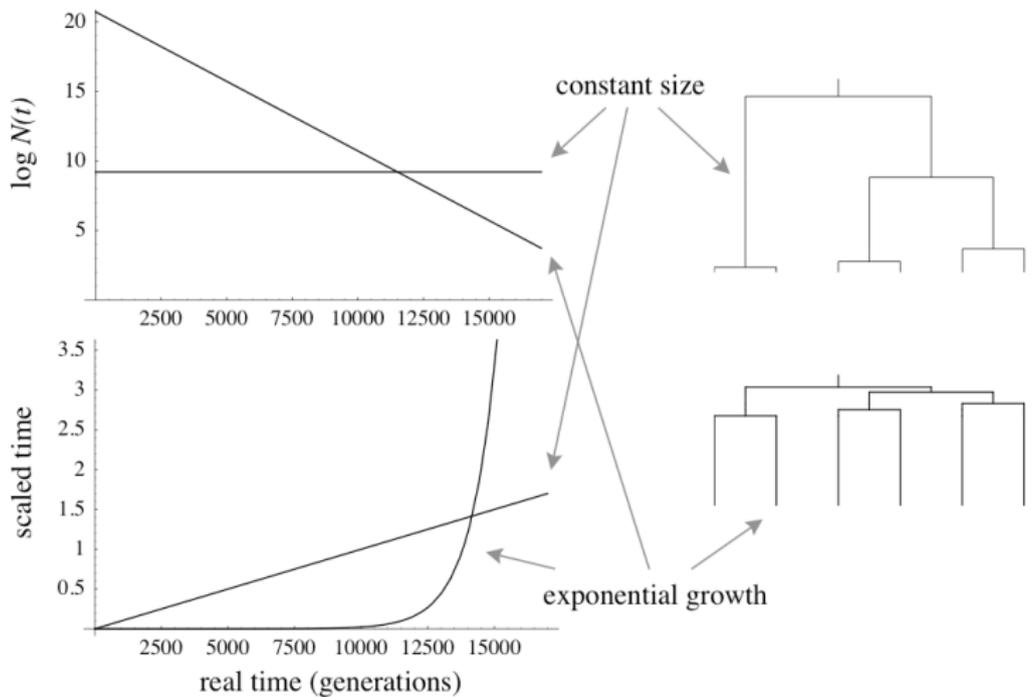
Figures 1-6 of M. Nordburg, Coalescent Theory, 2000

Partially Ordered Coalescent Experiments Graph
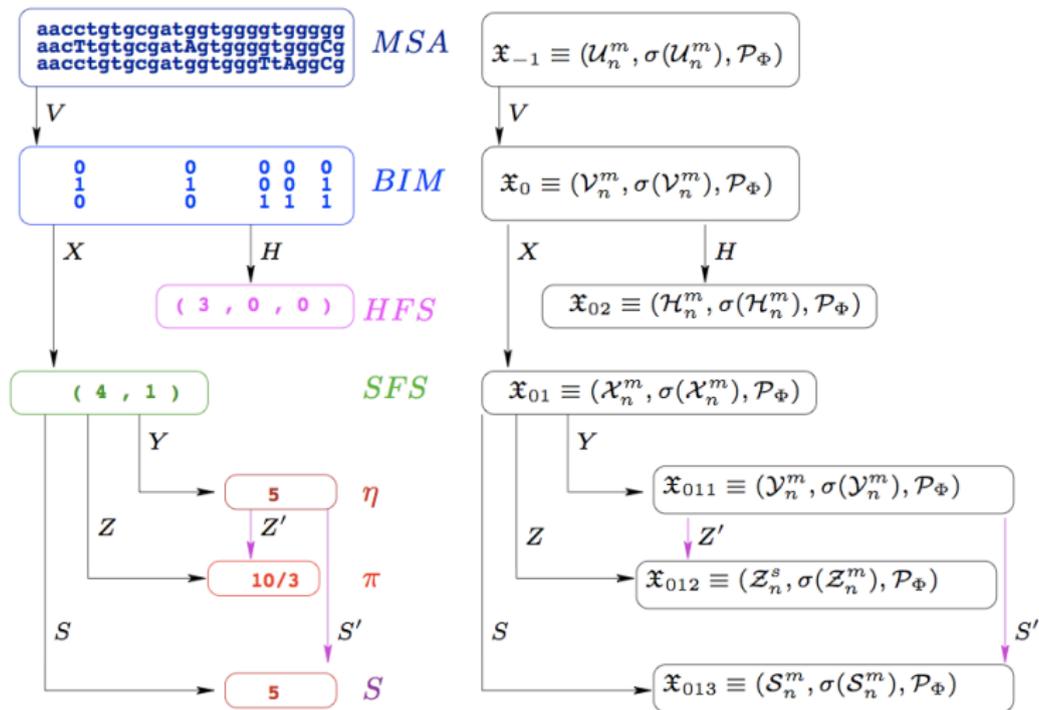
(1) Every directed acyclic subgraph of the POEG indexes a Martingale

(2) Each node of the POEG is a tri-sequential asymptotic family of Experiments

Likelihood, $P(D|\phi)$, is computed by Integrating Missing-Data:

$$\sum_{c \in \mathcal{C}_n} \int_{t \in (0,\infty)^{n-1}} P(D|c,t,\phi) P(c,t|\phi) dt \; dc$$

Cardinalities of the state spaces of the standard n-coalescent on $\mathbb{C}_n$ and the unlabeled n-coalescent on $\mathbb{F}_n$ (to be seen in the sequel).

n	4	10	30	60	90
$|\mathbb{C}_n|$	15	1.2×10^5	8.5×10^{23}	9.8×10^{59}	1.4×10^{101}
$|\mathbb{F}_n|$	5	42	5.6×10^3	9.7×10^5	5.7×10^7
$|\mathbb{F}_n|/|\mathbb{C}_n|$	0.33	3.6×10^{-4}	6.6×10^{-21}	9.9×10^{-55}	4.0×10^{-94}

Likelihood is computationally prohibitive for MSA/BIM
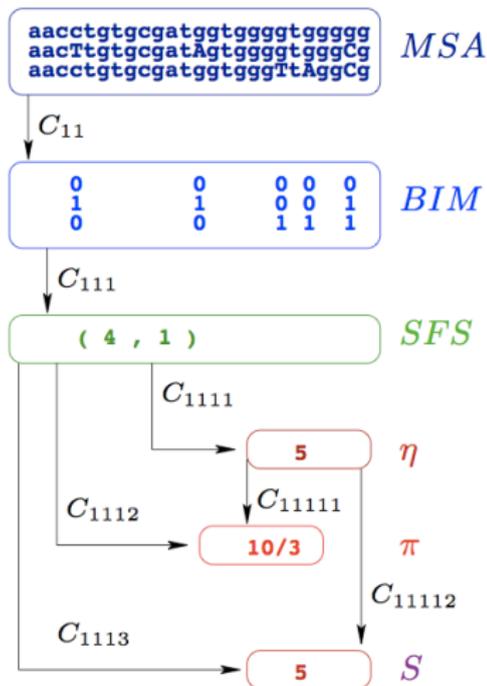
- Exact Methods :
  - MSA $10,000$ Auto-validating i.i.d. Posterior Samples in MRS SY2006 – novel
    (3/4 leaved phylogenetic tree spaces)
    ≈ 200 CPU sec for $n <= 3$,
    :-($\rightarrow$ impractical for $n > 4$
  - BIM Complete Recursion in PTREE G1980
    (1 Locus, $\theta = 10$, C-Model 1)
    :-($\rightarrow$ out of stack for $n > 4$

- Approximate Methods :
  - MSA MCMC in COALESCE KYF1998 : $n < 200$ & heuristic
  - BIM SIS in GENETREE GT1994 : $L(\theta|v) \cong 4$ CPU hrs / θ

- 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

Algorithm 3 A Simple ABC/ALC Algorithm

1: **input:**

    1. a samplable distribution $P(v|\phi)$ over $\mathscr{V}_n^m$ indexed by $\phi \in \boldsymbol{\Phi}$

    2. a samplable prior $P(\phi)$

    3. observed data $v_o \in \mathscr{V}(v)_n^m$ and summaries $r_o = R(v_o) \in \mathscr{R}_n^m$

    4. tolerance $\varepsilon \geq 0$

    5. a map $\mathfrak{m} : \mathscr{R}_n^m \times \mathscr{R}_n^m \to \mathbb{R}_+$

    6. a large positive integer `MAXTRIALS` $\in \mathbb{N}$

2: **output:** a sample $U \sim P(\phi | \mathbf{r}_\varepsilon(r_o)) \cong P(\phi|r_o) \cong P(\phi|v_o)$ or $\{\}$,

    where, $\mathbf{r}_\varepsilon(r_o) := \{r : \mathfrak{m}(r,r_o) \leq \varepsilon\}$.

3: **initialize:** `TRIALS` $\leftarrow 0$, `SUCCESS` $\leftarrow$ `false`, $U \leftarrow \{\}$

4: **repeat**

5:     $\phi \leftarrow P(\phi)$ {DRAW from Prior}

6:     $v \leftarrow P(v|\phi)$ {SIMULATE data}

7:     $r \leftarrow R(v)$ {SUMMARIZE data}

8:     **if** $\mathfrak{m}(r,r_o) \leq \varepsilon$ **then** {COMPARE summaries and ACCEPT/REJECT parameter}

9:         $U \leftarrow \phi$, `SUCCESS` $\leftarrow$ `true`

10:     **end if**

11:     `TRIALS` $\leftarrow$ `TRIALS` $+1$

12: **until** `TRIALS` $\geq$ `MAXTRIALS` or `SUCCESS` $\leftarrow$ `true`

13: **return:** U

A Currently Popular Alternative is ABC

Algorithm 3 A Simple ABC/ALC Algorithm

1: **input:**

    1. a samplable distribution $P(v|\phi)$ over $\mathcal{V}_n^m$ indexed by $\phi \in \Phi$

    2. a samplable prior $P(\phi)$

    3. observed data $v_o \in \mathcal{V}(v)_n^m$ and summaries $r_o = R(v_o) \in \mathcal{R}_n^m$

    4. tolerance $\varepsilon \geq 0$

    5. a map $\mathfrak{m} : \mathcal{R}_n^m \times \mathcal{R}_n^m \to \mathbb{R}_+$

    6. a large positive integer $\texttt{MAXTRIALS} \in \mathbb{N}$

2: **output:** a sample $U \sim P(\phi|\mathbf{r}_\varepsilon(r_o)) \cong P(\phi|r_o) \cong P(\phi|v_o)$ or $\{\}$,
    where, $\mathbf{r}_\varepsilon(r_o) := \{r : \mathfrak{m}(r, r_o) \leq \varepsilon\}$.

3: **initialize:** $\texttt{TRIALS} \leftarrow 0$, $\texttt{SUCCESS} \leftarrow \texttt{false}$, $U \leftarrow \{\}$

4: **repeat**

5:     $\phi \leftarrow P(\phi)$ {DRAW from Prior}

6:     $v \leftarrow P(v|\phi)$ {SIMULATE data}

7:     $r \leftarrow R(v)$ {SUMMARIZE data}

8:     **if** $\mathfrak{m}(r, r_o) \leq \varepsilon$ **then** {COMPARE summaries and ACCEPT/REJECT parameter}

9:         $U \leftarrow \phi$, $\texttt{SUCCESS} \leftarrow \texttt{true}$

10:    **end if**

11:    $\texttt{TRIALS} \leftarrow \texttt{TRIALS} + 1$

12: **until** $\texttt{TRIALS} \geq \texttt{MAXTRIALS}$ or $\texttt{SUCCESS} \leftarrow \texttt{true}$

13: **return:** U

- PROBLEM 1: but what is approximately sufficient?

Algorithm 3 A Simple ABC/ALC Algorithm

1: **input:**

    1. a samplable distribution $P(v|\phi)$ over $\mathscr{V}_n^m$ indexed by $\phi \in \boldsymbol{\Phi}$

    2. a samplable prior $P(\phi)$

    3. observed data $v_o \in \mathscr{V}(v)_n^m$ and summaries $r_o = R(v_o) \in \mathscr{R}_n^m$

    4. tolerance $\varepsilon \geq 0$

    5. a map $\mathfrak{m} : \mathscr{R}_n^m \times \mathscr{R}_n^m \to \mathbb{R}_+$

    6. a large positive integer MAXTRIALS $\in \mathbb{N}$

2: **output:** a sample $U \sim P(\phi|\mathbf{r}_\varepsilon(r_o)) \cong P(\phi|r_o) \cong P(\phi|v_o)$ or $\{\}$,

    where, $\mathbf{r}_\varepsilon(r_o) := \{r : \mathfrak{m}(r, r_o) \leq \varepsilon\}$.

3: **initialize:** TRIALS $\leftarrow 0$, SUCCESS $\leftarrow$ false, $U \leftarrow \{\}$

4: **repeat**

5:     $\phi \leftarrow P(\phi)$ {DRAW from Prior}

6:     $v \leftarrow P(v|\phi)$ {SIMULATE data}

7:     $r \leftarrow R(v)$ {SUMMARIZE data}

8:     **if** $\mathfrak{m}(r, r_o) \leq \varepsilon$ **then** {COMPARE summaries and ACCEPT/REJECT parameter}

9:         $U \leftarrow \phi$, SUCCESS $\leftarrow$ true

10:    **end if**

11:    TRIALS $\leftarrow$ TRIALS $+ 1$

12: **until** TRIALS $\geq$ MAXTRIALS or SUCCESS $\leftarrow$ true

13: **return:** U

- PROBLEM 1: but what is approximately sufficient?
- PROBLEM 2: the "epsilon-dilemma" —ABCDE Fixes 1 & 2

Let $v \in \mathcal{V}_n^m \equiv \{0,1\}^{n \times m}$ be a BIM, then the SFS
$x \equiv (x_1, \ldots, x_{n-1}) \in \mathcal{X}_n^m \equiv \{x \in \mathbb{Z}_+^{n-1} : \sum_{i=1}^{n-1} x_i \leq m\}$

$$x_i = N_i(v^T \cdot (1,1,\ldots,1)), \qquad N_i(y_1, y_2, \ldots, y_s) = \sum_{j=1}^{s} \mathbf{1}_{\{i\}}(y_j), \qquad i = 1, \ldots, n-1.$$
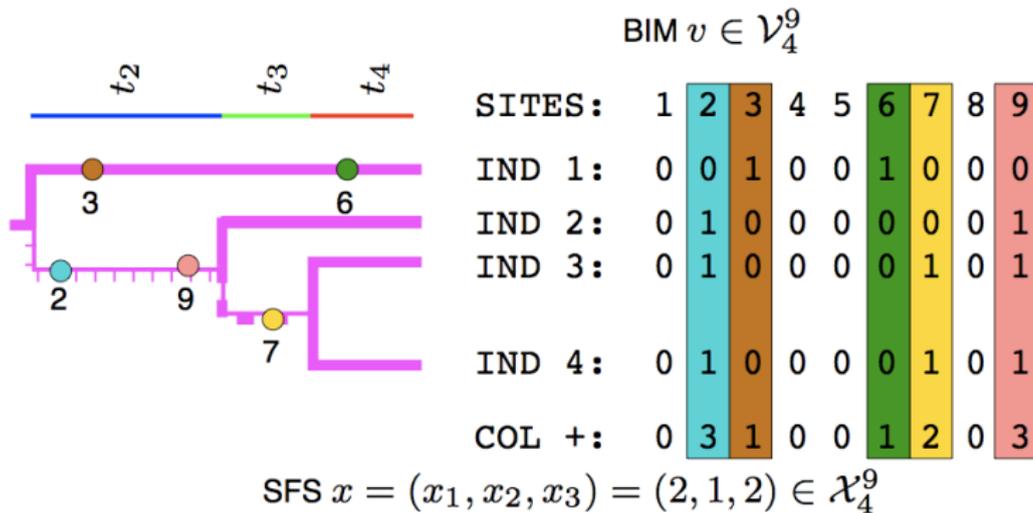
BIM $v \in \mathcal{V}_4^9$

SITES:	1	2	3	4	5	6	7	8	9
IND 1:	0	0	1	0	0	1	0	0	0
IND 2:	0	1	0	0	0	0	0	0	1
IND 3:	0	1	0	0	0	0	1	0	1
IND 4:	0	1	0	0	0	0	1	0	1
COL +:	0	3	1	0	0	1	2	0	3

SFS $x = (x_1, x_2, x_3) = (2, 1, 2) \in \mathcal{X}_4^9$

$$((\bullet,_3\bullet),_1(\bullet,_2\bullet))$$

$$\mathbf{f}^\wedge = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 2 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 4 & 0 & 0 & 0 \end{pmatrix}$$

$$P(f^\wedge) = 1/3$$

$$(((\bullet,_3\bullet),_2\bullet),_1\bullet)$$

$$\mathbf{f}^\curlywedge = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 2 & 1 & 0 & 0 \\ 4 & 0 & 0 & 0 \end{pmatrix}$$

$$P(f^\curlywedge) = 2/3$$

$$\mathscr{X}_4{}^{10^5}$$

$$\mathscr{F}_4 = \{f^\wedge, f^\curlywedge\}$$
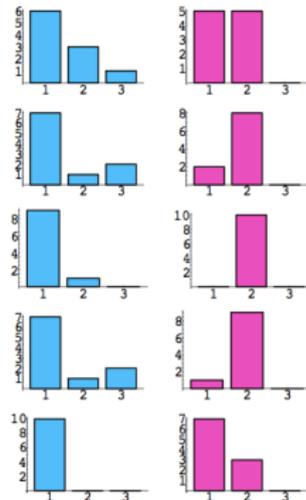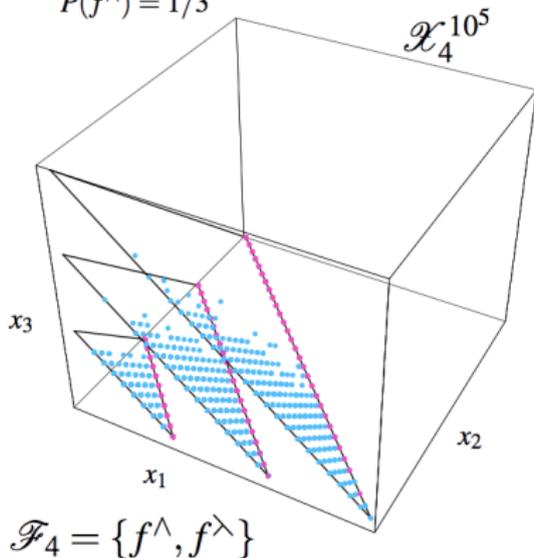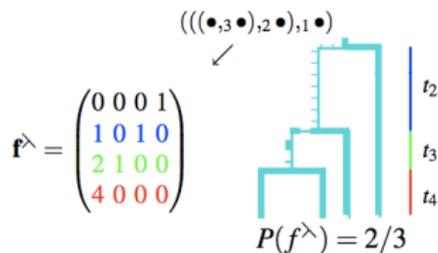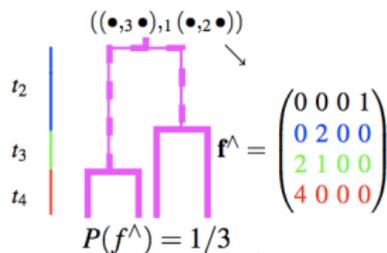
Transition-Diagram
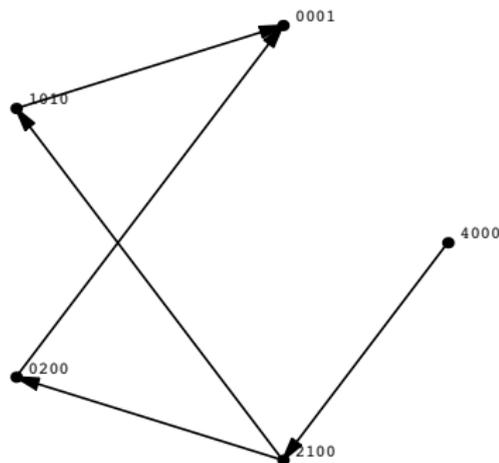
Hasse-Diagram
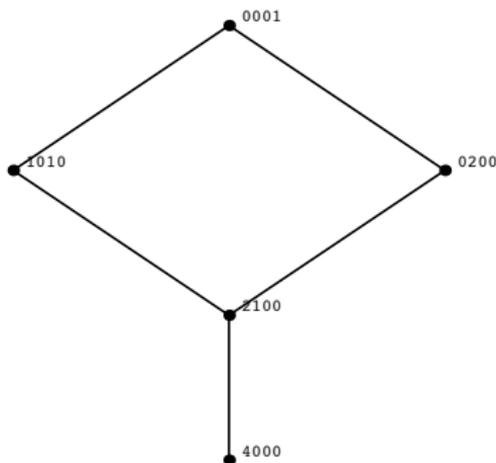
Ex 1:
$[\{1\}, \{2\}, \{3\}, \{4\}], [\{1, 2\}, \{3\}, \{4\}], [\{1, 2, 3\}, \{4\}], [\{1, 2, 3, 4\}] \rightarrow$
$[(4, 0, 0, 0), (2, 1, 0, 0), (1, 0, 1, 0), (0, 0, 0, 1)]$
Ex 2:
$[\{1\}, \{2\}, \{3\}, \{4\}], [\{1, 2\}, \{3\}, \{4\}], [\{1, 2\}, \{3, 4\}], [\{1, 2, 3, 4\}] \rightarrow$
$[(4, 0, 0, 0), (2, 1, 0, 0), (0, 2, 0, 0), (0, 0, 0, 1)]$
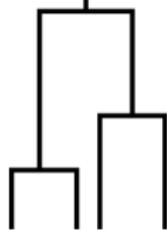
$((\bullet, _3\bullet), _1(\bullet, _2\bullet))$

$\mathbf{f}^\wedge = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 2 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 4 & 0 & 0 & 0 \end{pmatrix}$

0001

0200

1010

2100

4000

$\mathbf{f}^\lambda = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 2 & 1 & 0 & 0 \\ 4 & 0 & 0 & 0 \end{pmatrix}$

$(((\bullet, _3\bullet), _2\bullet), _1\bullet)$
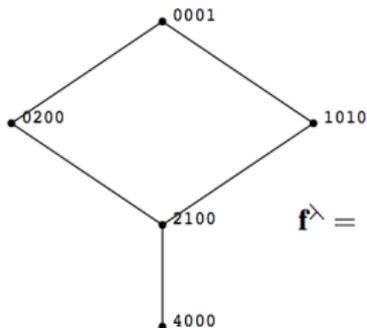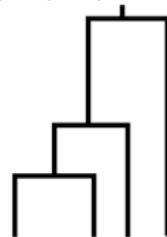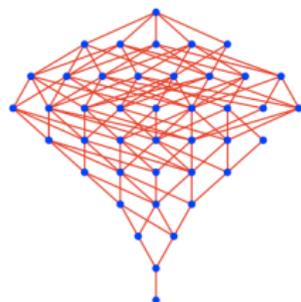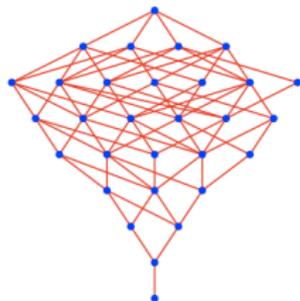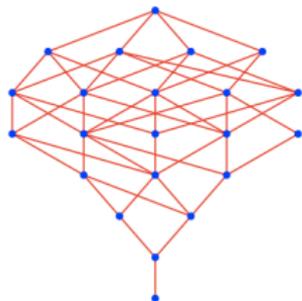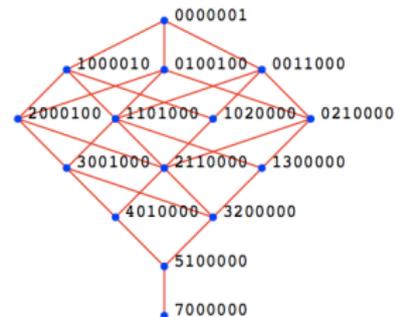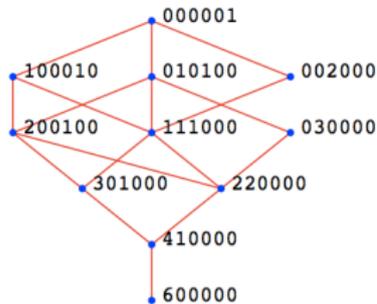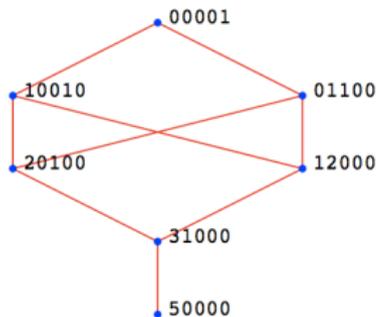
Kingman's Unlabeled *n*-Coalescent

Consider, the integer partitions of n with i blocks:

$$\mathbb{F}_n^i \equiv \{f_i \equiv (f_{i,1}, f_{i,2}, \ldots, f_{i,n}) \in \mathbb{Z}_+^n : \sum_{j=1}^n j f_{i,j} = n, \sum_{j=1}^n f_{i,j} = i\}.$$

where $f_{i,j}$ is the number of lineages subtending j leaves at the i-th epoch.

Theorem (Kingman's Unlabeled *n*-coalescent)

It is the continuous time Markov chain on $\mathbb{F}_n \equiv \cup_{i=1}^n \mathbb{F}_n^i$, the set of integer partitions of n, whose infinitesimal generator $\mathbf{q}(f_h|f_g)$ for any two states $f_g, f_h \in \mathbb{F}_n$ is:

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

Initial state: $f_n = (n, 0, 0, \ldots, 0)$ and absorbing state: $f_1 = (0, 0, \ldots, 1)$.

Any realization of the chain is an f-sequence: $f = (f_n, f_{n-1}, \ldots, f_1) \in \mathcal{F}_n$.

1: **input:**

    1. scaled mutation rate θ

    2. sample size n

2: **output:** a SFS sample x from the n-coalescent

3: generate an f-sequence under the unlabeled n-coalescent

4: draw $t \sim T = (T_2, T_3, \ldots, T_n)$, where T_i's are independently distributed as $\mathrm{Exponential}\left(\binom{i}{2}\right)$

5: $l \leftarrow t^{\mathbf{T}} \cdot \mathbf{f}$ and $l_\bullet = \sum_{i=1}^{n-1}$

6: draw x from Poisson-Multinomial distribution
$e^{-\theta l_\bullet} (\theta l_\bullet)^{\sum_{i=1}^{n-1} x_i} \prod_{i=1}^{n-1} \bar{l}_i^{x_i} / \prod_{i=1}^{n-1} x_i!$

7: **return:** x

Likelihood of a Site Frequency Spectrum

Theorem (Likelihood of SFS)

Let c, f and t be the c-sequence, f-sequence, and epoch times of tree a, then

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

where l is lineage lengths subtending $1, 2, \ldots, n-1$ *leaves. Then:*

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

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

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

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

$$X^\circledast(x) = x^\circledast \equiv (x_1^\circledast, \ldots, x_{n-1}^\circledast) \equiv (\mathbf{1}_{\mathbb{N}}(x_1), \ldots, \mathbf{1}_{\mathbb{N}}(x_{n-1})) \in \{0, 1\}^{n-1}$$
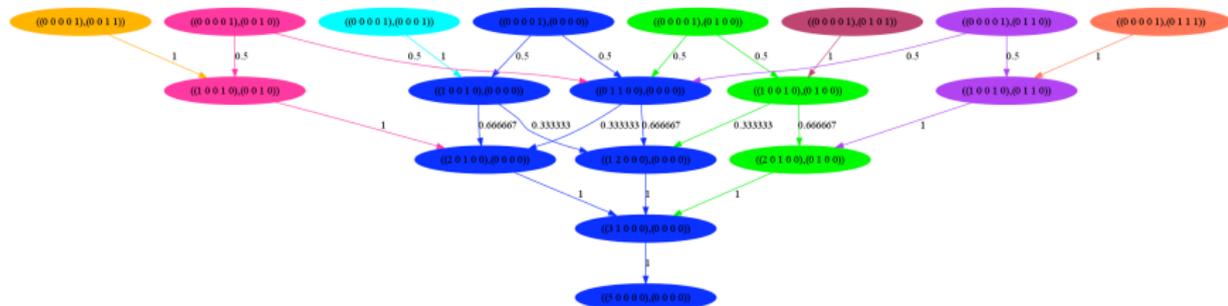
Transition diagram of $\{F^{\downarrow x^{\circledast}}(k)\}_{k \in [5]_+}$ over states in $\mathbb{F}_n^{x^{\circledast}}$. The simplified diagram replaces the states that do not affect the transitions, namely, $x_1^{\circledast}$ and $x_2^{\circledast}$, with $* \in \{0, 1\}$.

Transition diagram of $\{F^{\downarrow x^\circledast}(k)\}_{k \in [5]_+}$
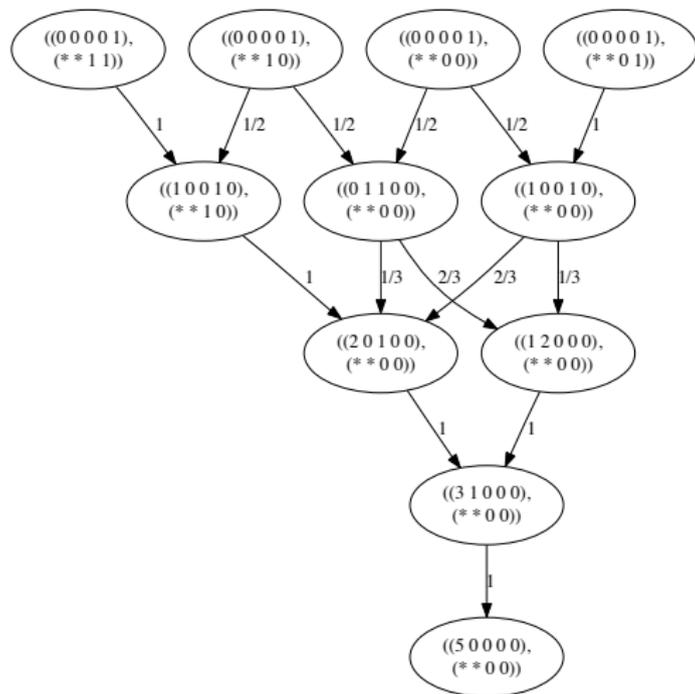
Transition diagram of $\{F^{\downarrow x^\circledast}(k)\}_{k \in [5]_+}$ over states in $\mathbb{F}_n^{x^\circledast}$. The simplified diagram replaces the states that do not affect the transitions, namely, $x_1^\circledast$ and $x_2^\circledast$, with $* \in \{0, 1\}$.

An Importance Sampler over $F_n^c(x^\circledast)$

Theorem (A Proposal over $F_n^c(x^\circledast)$)

For a given $x \in \mathcal{X}_n^m$, consider the following discrete time Markov chain
$\{F^{|x^\circledast}(k)\}_{k \in [n]_+}$ *on the augmented state space* $\mathbb{F}_n \times \{0,1\}^{n-1} \ni (f_h, z_h)$:

$$P^*((f_h, z_h)|(f_g, z_g)) = \begin{cases} P(f_h|f_g)/\Sigma(f_g, z_g) & : \text{if} \quad (f_h, z_h) \prec_{f,z} (f_g, z_g), \\ 0 & : \text{otherwise} \end{cases}$$

where,

$$\Sigma(f_g, z_g) = \sum_{(j,k) \in H(f_g, z_g)} P(f_g - e_{j+k} + e_j + e_k | f_g),$$

$$H(f_g, z_g) = \{(j,k) : f_{g,j+k} > 0, \, 1 \leq j \leq \max\{\min\{\hat{g}, j+k-1\}, \lceil \frac{j+k}{2} \rceil\} \leq k \leq j+k-1\},$$

$$\hat{g} = \max\{i : z_{g,i} = 1\},$$

$$(f_h, z_h) \prec_{f,z} (f_g, z_g) \Leftrightarrow f_h = f_g + e_j + e_k - e_{j+k}, \, z_h = z_g - \mathbf{1}_{\{1\}}(z_{g,j}) \, e_j - \mathbf{1}_{\{1\}}(z_{g,k}) \, e_k$$

where, the initial state is $(f_1, X^\circledast(x)) = ((0, 0, \dots, 1), x^\circledast)$ and the final absorbing state is $(f_n, (0, 0, \dots, 0)) = ((n, 0, \dots, 0), (0, 0, \dots, 0))$.

Table 1 10^4 loci were simulated under each hypothesised model $H_0, H_1, \ldots H_8$ and tested for the extremeness of the observed Tajima's D statistic with and without conditioning on the observed $x^{\circledR}$ in an attempt to reject the null hypothesis H_0 at significance level $\alpha = 5\%$.

Model: parameters	Proportion of loci rejected by null distribution of test statistics			
$H_i : (m\phi_1, \phi_2, \rho)$	$P_{H_0} (D \geq d)$	$P_{H_0} (D \geq d\|x^{\circledR})$	$P_{H_0} (D \leq d)$	$P_{H_0} (D \leq d\|x^{\circledR})$
$H_0 : (100, 0, 0)$	0.0495	0.0501	0.0499	0.0501
$H_1 : (100, 0, 10)$	0.0074	0.8640	0.0061	0.0017
$H_2 : (100, 0, 100)$	0.0000	0.9999	0.0000	0.0000
$H_3 : (100, 10, 0)$	0.0000	0.0019	0.0326	0.1759
$H_4 : (100, 10, 10)$	0.0001	0.2023	0.0135	0.0797
$H_5 : (100, 10, 100)$	0.0000	0.5559	0.0006	0.0180
$H_6 : (100, 100, 0)$	0.0000	0.0000	0.1696	0.6882
$H_7 : (100, 100, 10)$	0.0000	0.0002	0.1580	0.6668
$H_8 : (100, 100, 100)$	0.0000	0.0020	0.1321	0.6617

(A)   (B)
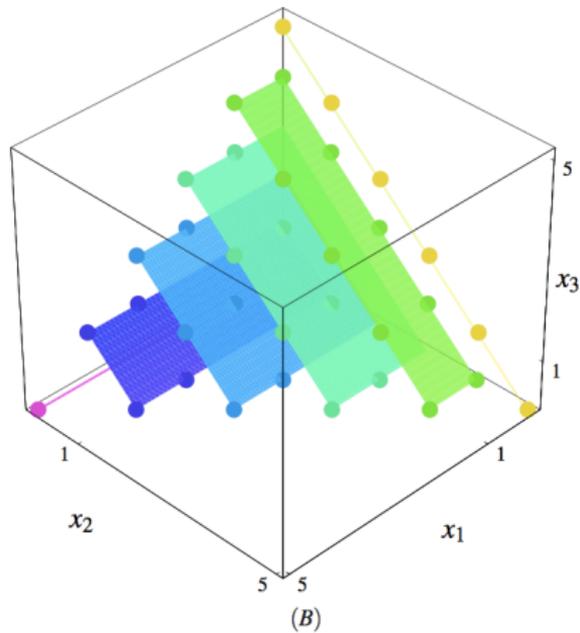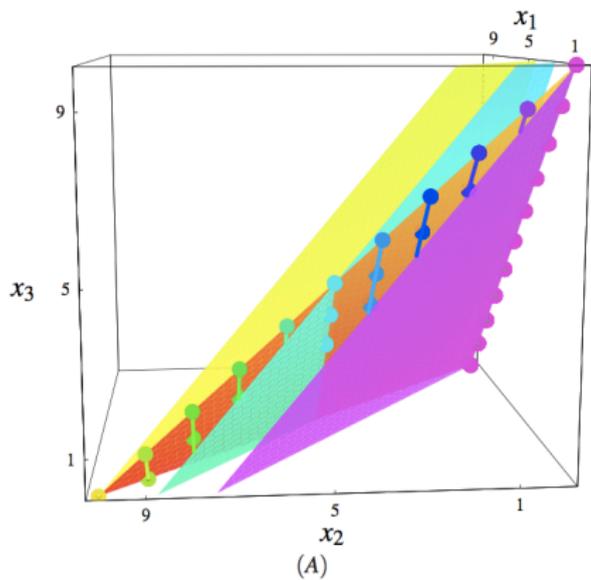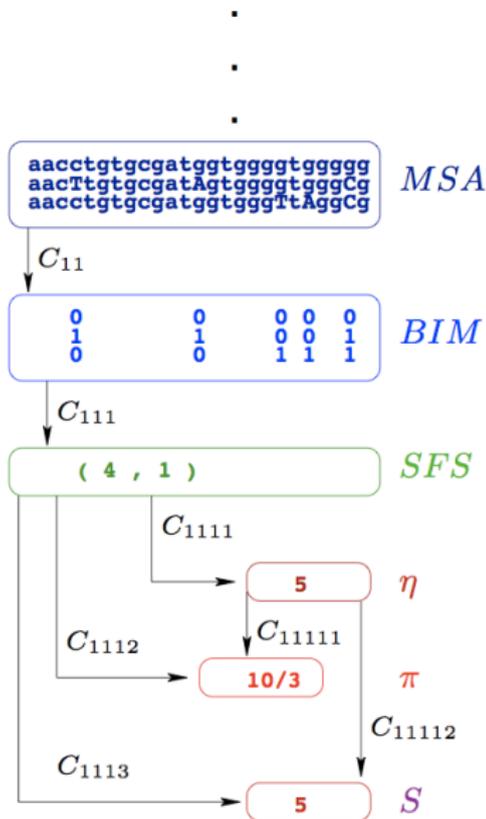
- Limits on Inference from Finest Empirical Resolutions

- Inference from Coarser Site Frequency Spectrum is Possible via a Collapsed Kingman's n-coalescent Markov chain

- Algebraic Geometry is useful to infer from classical summaries of SFS.

- MSEs are smaller – the exponential growth model

- Helps speed-up intensive SIS methods (Particle filtering on Experiment Graph)

- Topological unfolding of SFS and D $\Rightarrow$ Tree-less Genome Scans are essentially meaningless

- A Decision-theoretic formalism – partially-ordered coalescent experiments graph

- Possible to generalize

- Saves electricity and slows down global warming!

The End

- Many thanks to:
  - Research Fellowship of The Royal Commission for the Exhibition of 1851
  - Mike Steel for pointers to Definition of Lumped Markov chain.
  - Allan Wilson Centre Summur Studentship to Jenny Harlow
  - Simon Tavare and Michael Nussbaum for discussions on Approximate Sufficiency

For the Full Story See:

R. Sainudiin, K. Thornton, J. Harlow, J. Booth, M. Stillman, R. Yoshida, R. Griffiths, G. McVean and P. Donnelly *Experiments with the Site Frequency Spectrum*, Bulletin of Mathematical Biology, Algebraic Biology Special Edition, pp. 1-44, 2010. http://www.springerlink.com/content/0748966716753484/.

R. Sainudiin, K. Thornton, J. Harlow and B. Bycroft, *LCE: a C++ Class Library for Lumped Coalescent Experiments*, GPL licensed, available from http://www.math.canterbury.ac.nz/~r.sainudiin/codes/lce, 2010.