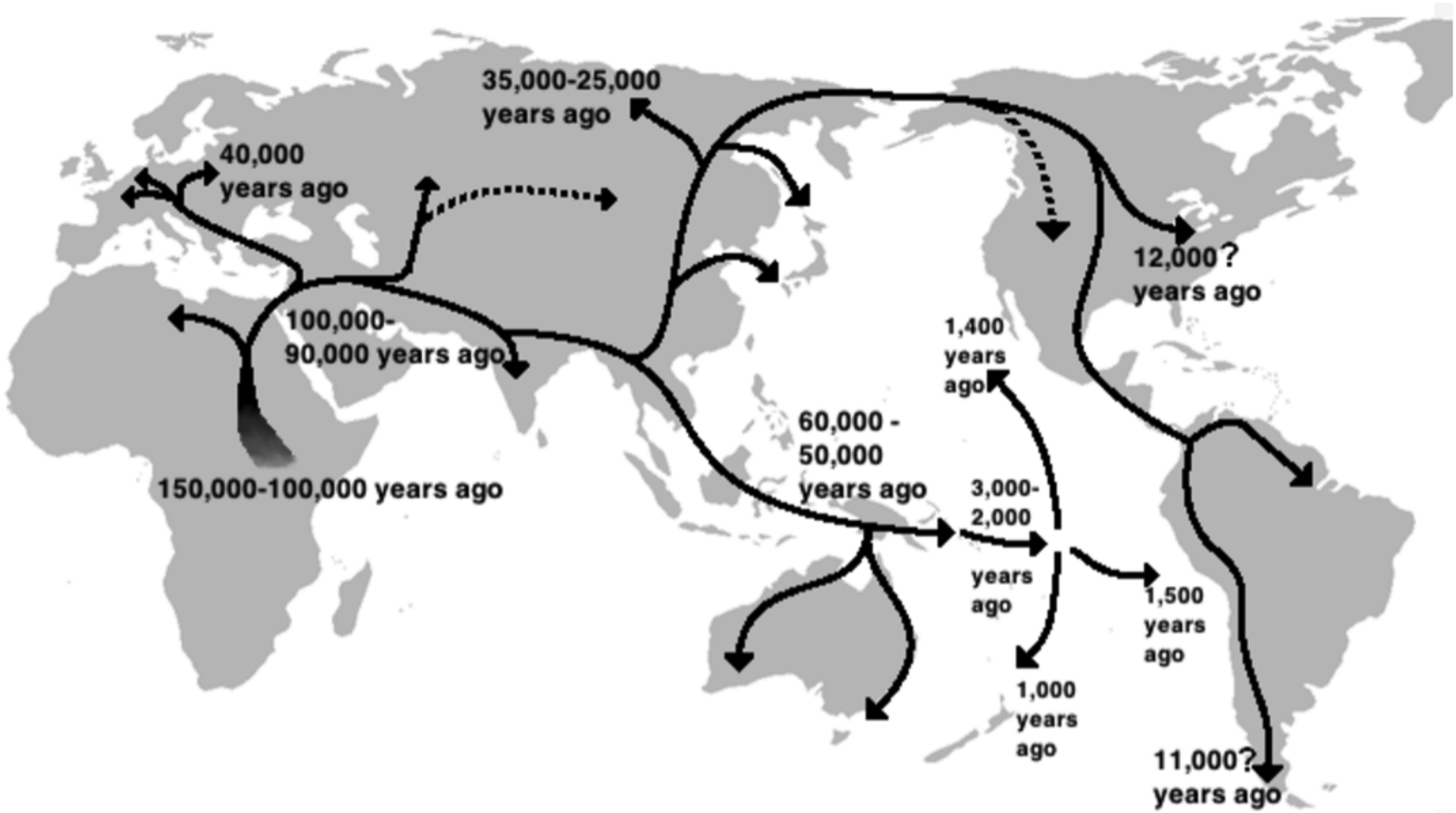
Population History Estimations: A Coalescent Theory Approach

Xuehua Lan
Supervisor: Nathan Ross

The University of Melbourne
Deaprtment of Mathematics and Statistics

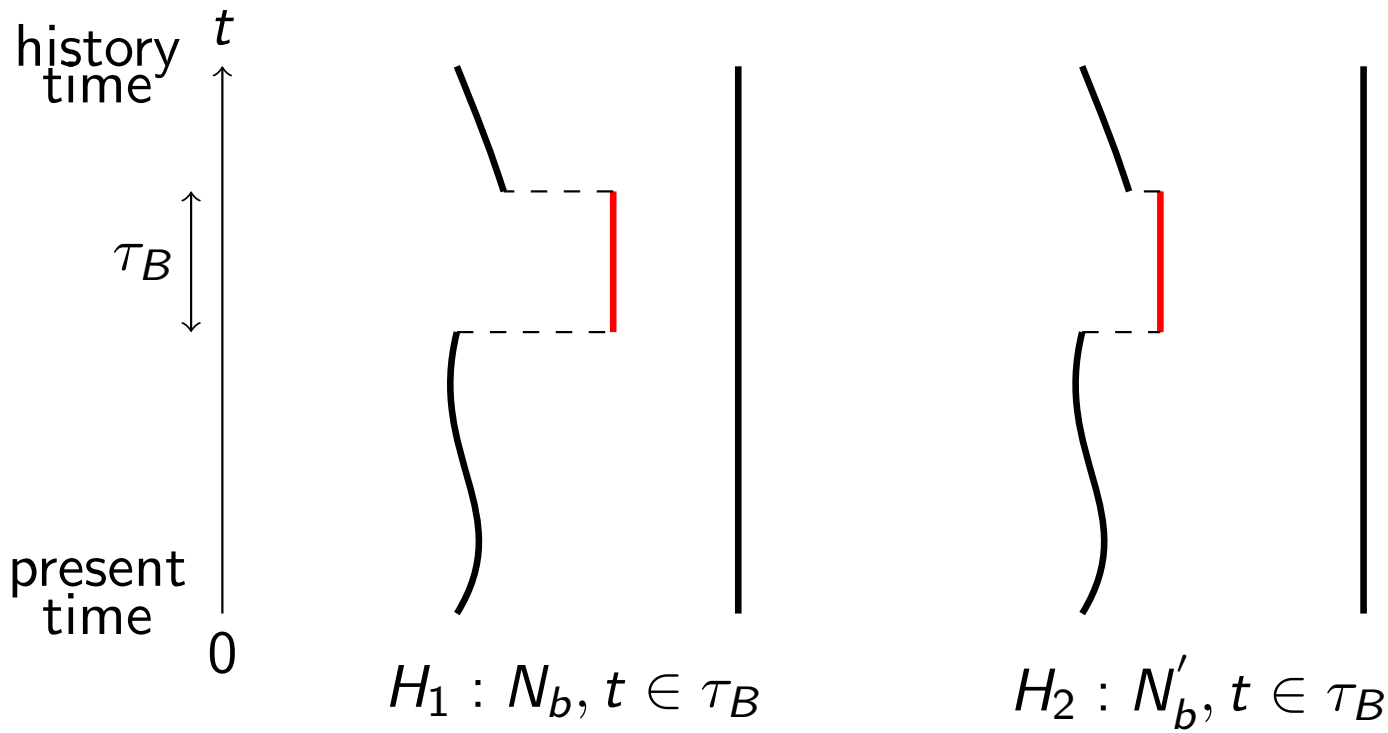
May, 2017

Human Migration out of Africa

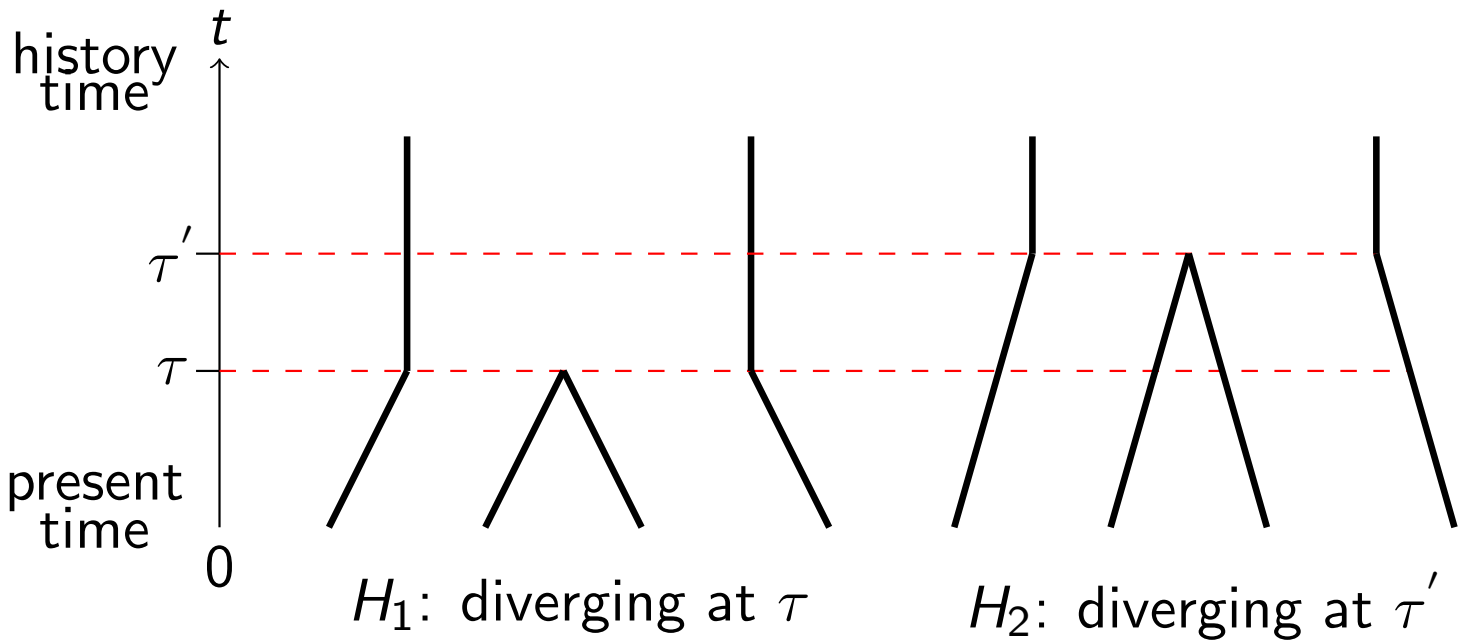


Source *https* : *//wikimedia.org/wiki/File* : *Human_migration_out_of_Africa*

Interest: is there a bottleneck?



Interest: when is the diverging time?



Introduction

Interests

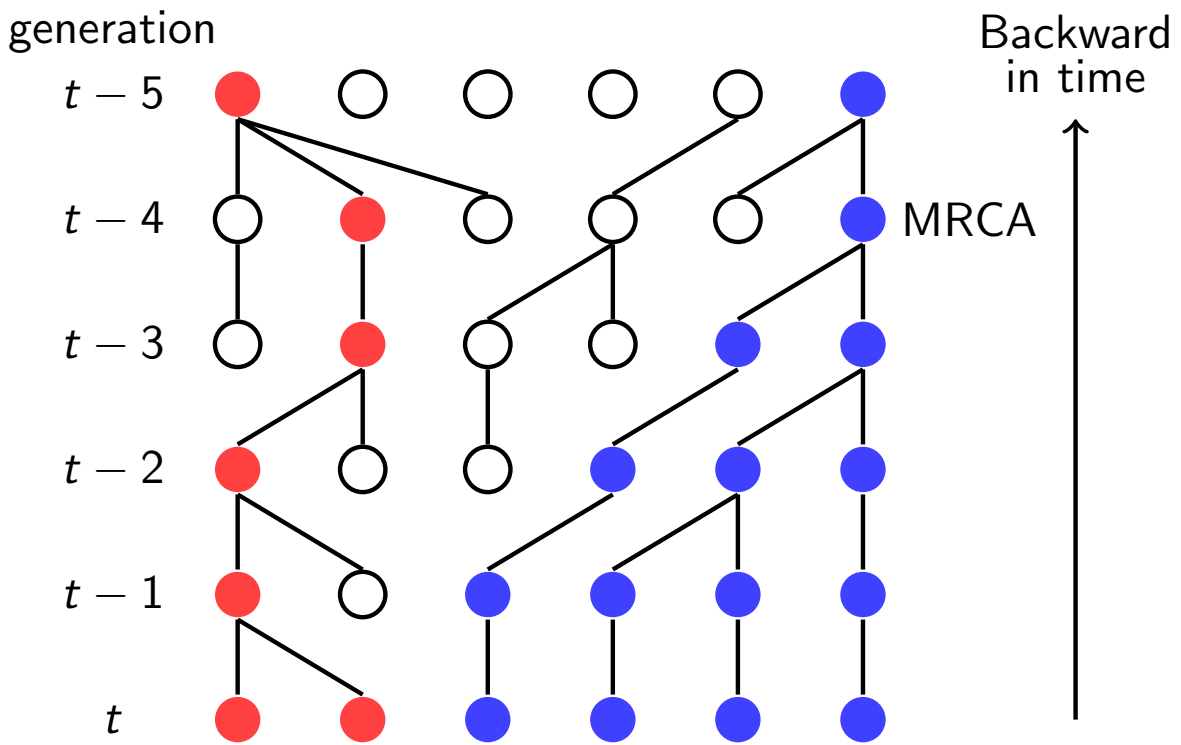
theoretical lower bounds w.r.t. amount of data used.

Outlines

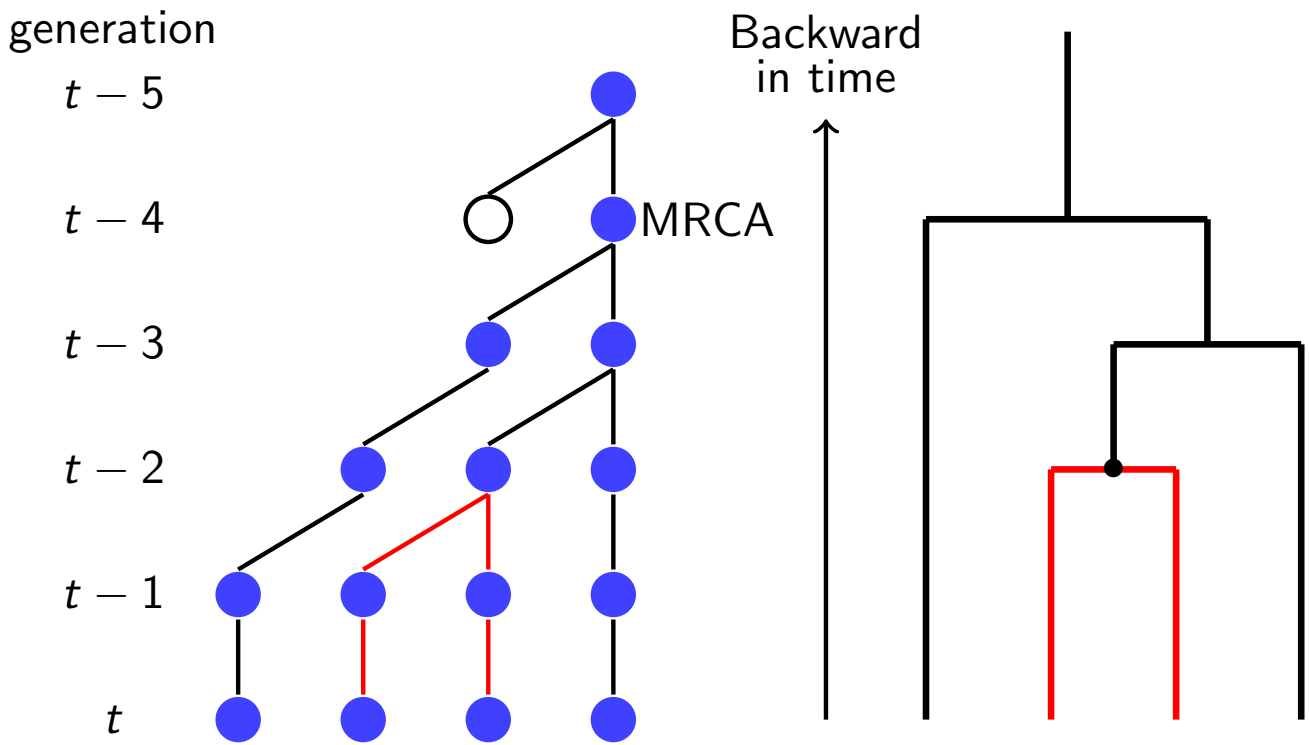
- Models
- Statistics
- Data used
- Lower Bounds
- Results

Wright-Fisher Model and Coalescent

- Wright-Fisher: population has constant size N
  uniformly sampling for parents
- Coalescent: find the most recent common ancestor (MRCA)



Gene Genealogies and Coalescent Tree



Kingman Coalescent from Wright-Fisher Model

- n sample chosen from population N

discrete-time coalescent

- 0 coal: $p_{n,n} = \frac{N-1}{N} \ldots \frac{N-n+1}{N} = 1 - \frac{n(n-1)}{2N} + \mathcal{O}(N^{-2})$
- 1 coal: $p_{n,n-1} = \frac{N-1}{N} \ldots \frac{N-n+2}{N} \cdot \frac{n(n-1)}{2N} = \frac{n(n-1)}{2N} + \mathcal{O}(N^{-2})$

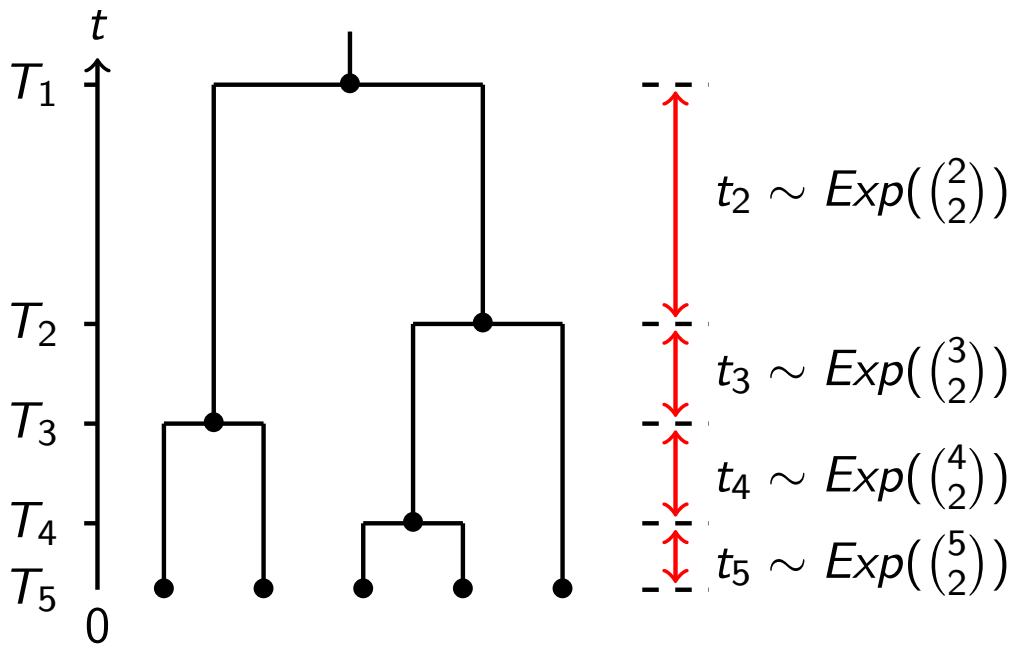
continuous-time coalescent

- time measured in units of N generations, $N \to \infty$
- t_n: time for n lineages coalescent into $n-1$ lineages

$$P(t_n > t) = (1 - \frac{n(n-1)}{2N})^{N \cdot t} \to \exp(-t\frac{n(n-1)}{2}) \qquad (1)$$

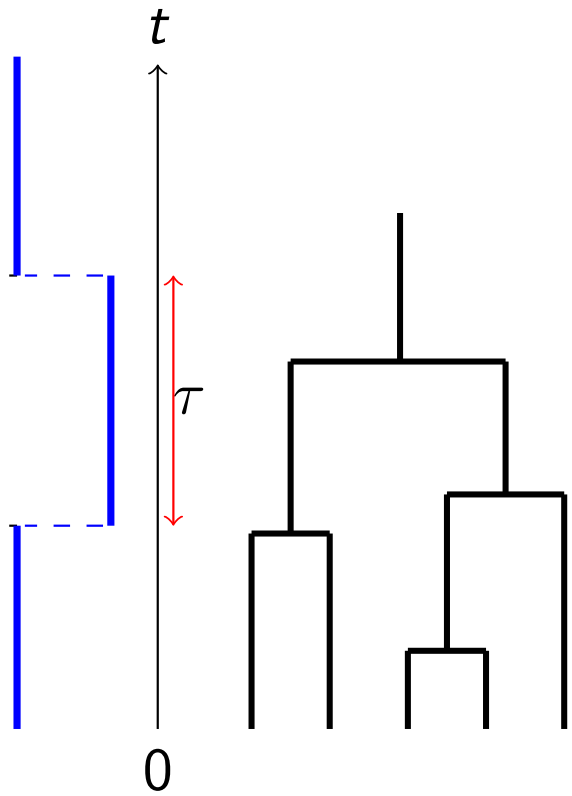
Kingman Coalescent with Constant Populations

- $n = 5$ sample from population N
- time measured in units of N generations

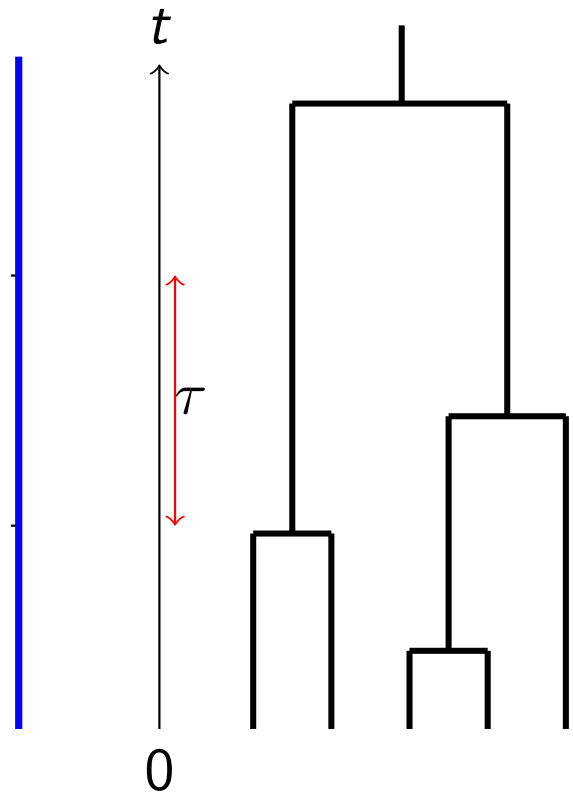


- coalescent time: $T_n := 0 < T_{n-1} < \ldots < T_2 < T_1$
- coalescent waiting time: $t_n := T_{n-1} - T_n, \ldots, t_2 := T_1 - T_2$

Coalescent Tree and Variable Population



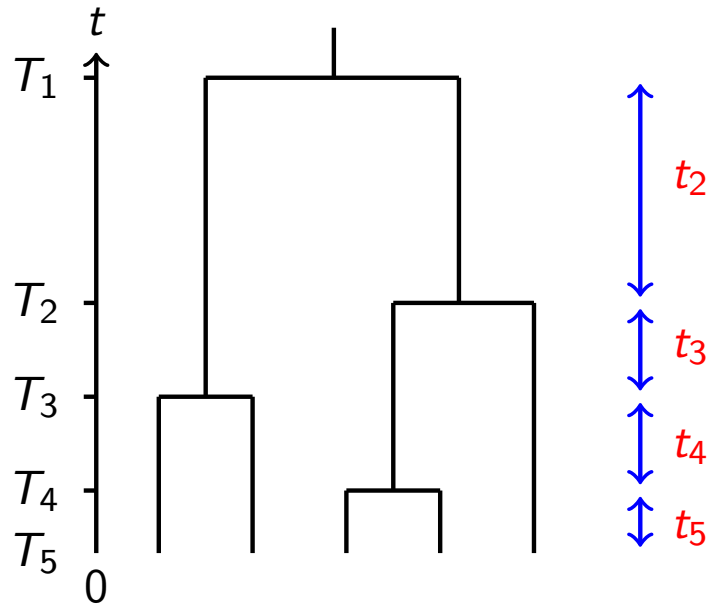
(*a*): Bottleneck in τ



(*b*): Constant

Coalescent Statistics

- Population size function $\eta(t), t \in [0, \infty)$
- Given colescent time $s_n = 0 < s_{n-1} < \ldots < s_{j+1}$



- Expected waiting time for $j+1$ lineages coalescent into j

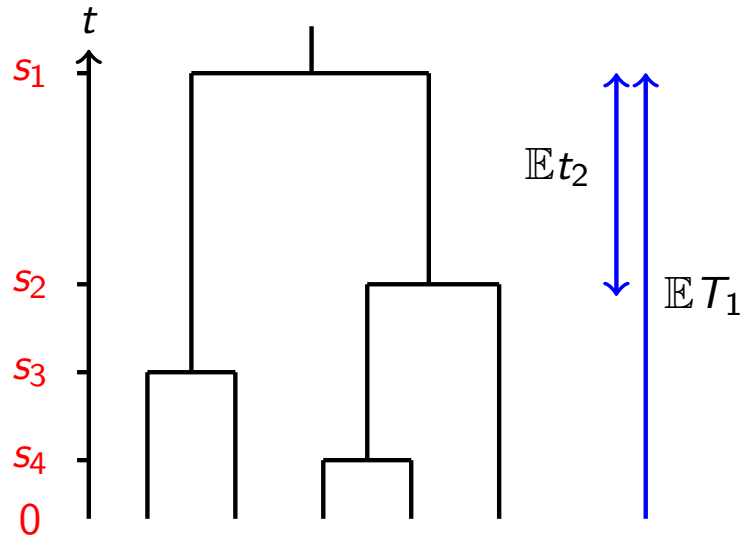
$$\mathbb{E}t_{j+1} = \int_{s_{j+1}}^{\infty} \exp(-\int_{s_{j+1}}^{t} \frac{\binom{j+1}{2}}{\eta(u)}\, du)\, dt;$$

- Expected Total Length (branches) of the Coalescent Tree

$$\mathbb{E}\, T_{tot} = \sum_{k=2}^{n} k \cdot \mathbb{E}t_k.$$

Distributions of Coalescent Time

- coalescent time
  $\mathbf{s} = (s_1, \ldots, s_{n-1})$
- Population size
  $\eta(t), t \in [0, \infty)$



- Density of coalescent time

$$p(s_{n-1}, \ldots, s_1) = \prod_{k=2}^{n} \frac{\binom{k}{2}}{\eta(s_{k-1})} \cdot \exp\left(-\binom{k}{2} \int_{s_k}^{s_{k-1}} \frac{1}{\eta(t)}\, dt\right)$$

- approach by 2 sample with 1 coalescent time data s_1

$$p(s_1) = \frac{1}{\eta(s_1)} \cdot \exp\left(-\int_0^{s_1} \frac{1}{\eta(t)}\, dt\right) \qquad (2)$$

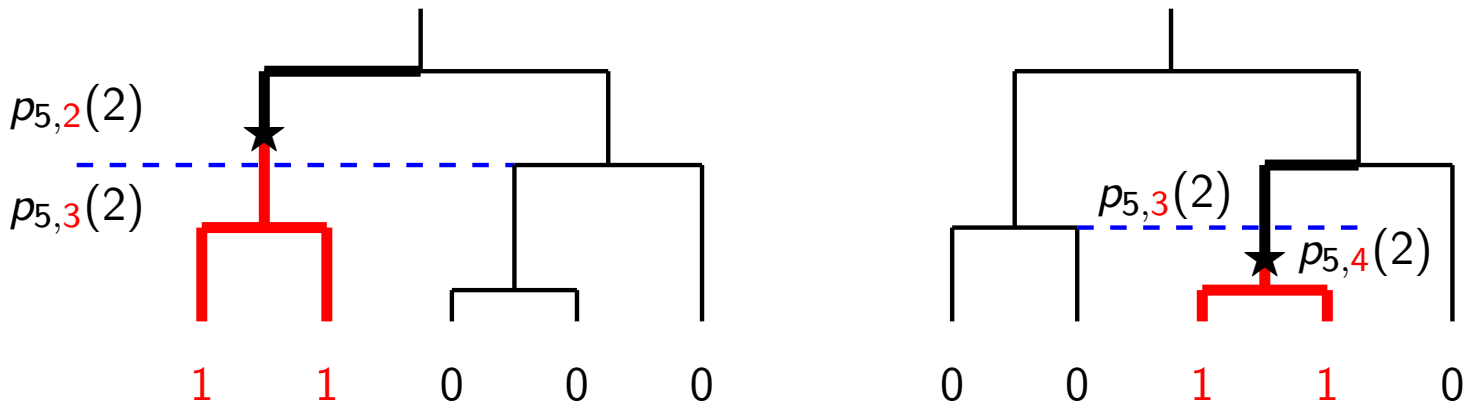
Sample Frequency Spectrum

- 6 DNA sequences (individuals);
- 7 SNPs or segregating sites;
- '1' for mutant; '0' for normal;
- SFS: $\chi^6 = \frac{1}{7}(3, 1, 2, 0, 1)$.

6 sample	1	2	3	4	5	6	7
1	0	0	1	1	0	0	1
2	0	0	0	1	0	1	0
3	0	1	0	0	1	0	1
4	1	1	0	1	0	0	1
5	0	0	0	1	0	0	0
6	0	1	1	1	0	0	0
$\sum$mutant	1	3	2	5	1	1	3

Distributions of Sample Frequency Spectrum

- Given n, $p_{n,k}(b) = \binom{n-b-1}{k-2} / \binom{n-1}{k-1}$ prob. mutation happen in k lineages and eventually derived b mutant genes at present



- prob. of a SNP has b mutant alleles given n sample:

$$q_{n,b} = \frac{\sum_{k=2}^{n} k \cdot p_{n,k}(b) \cdot \mathbb{E}(t_k)}{\sum_{k=2}^{n} k \cdot \mathbb{E}(t_k)} \tag{3}$$

Distributions of Sample Frequency Spectrum

- dist. of expected SFS: $q_{n,b} = \frac{\sum_{k=2}^{n} k \cdot p_{n,k}(b) \cdot \mathbb{E}(t_k)}{\sum_{k=2}^{n} k \cdot \mathbb{E}(t_k)}$.
- Is $q_{n,b}$ unique represent $\eta(t)$?
  Myers et al. (2008) unbounded frequency of oscillatory;
  Bhaskar and Song (2014) n is of sign change complexity.
- Is $\mathbb{E}(t_k)$ calculable?
  Polanski et al. (2003) explicit expression of $\mathbb{E}t_k$;
  Polanski and Kimmel (2003) dist. $q_{n,b}$.

Minimax Bounds between Two Hypotheses

- m indep. data $\hat{\theta}^{n,m} := \hat{\theta}^{n,m}(X_1, \ldots, X_m)$;
- η_1, η_2 measured in a metric space $(\mathcal{D}, d)$;
- distributions P_1 and P_2 induced by η_1, η_2 resp.;
- $P_1^m = P_1 \times \ldots \times P_1$ and $P_2^m = P_2 \times \ldots \times P_2$;
- Total variation distance

$$d_{TV}(P_1, P_2) = \sup_A |P_1(A) - P_2(A)|;$$

- Modified Le Cam theorem;

$$\inf_{\hat{\theta}} \sup_{\eta_1, \eta_2} \mathbb{E} d(\hat{\theta}, \theta(\eta)) \geq \frac{d(\eta_1, \eta_2)}{2} \cdot (1 - d_{TV}(P_1^m, P_2^m)) \qquad (4)$$

Distribution Divergences: Hellinger distance

- Hellinger distance

$$d_H^2(P_1, P_2) = \frac{1}{2}\int_D (\sqrt{f_{P_1}} - \sqrt{f_{P_2}})^2 \, d\mu$$
$$= 1 - \int_D \sqrt{f_{P_1}}\sqrt{f_{P_2}} \, d\mu$$

- For $P_1^m = P_1 \times \ldots \times P_1$ and $P_2^m = P_2 \times \ldots \times P_2$

$$d_{TV}^2(P_1^m, P_2^m) \leq 2m \cdot d_H^2(P_1, P_2) \tag{5}$$

Distribution Divergences: Kullback-Leibler distance

- Kullback-Leibler distance

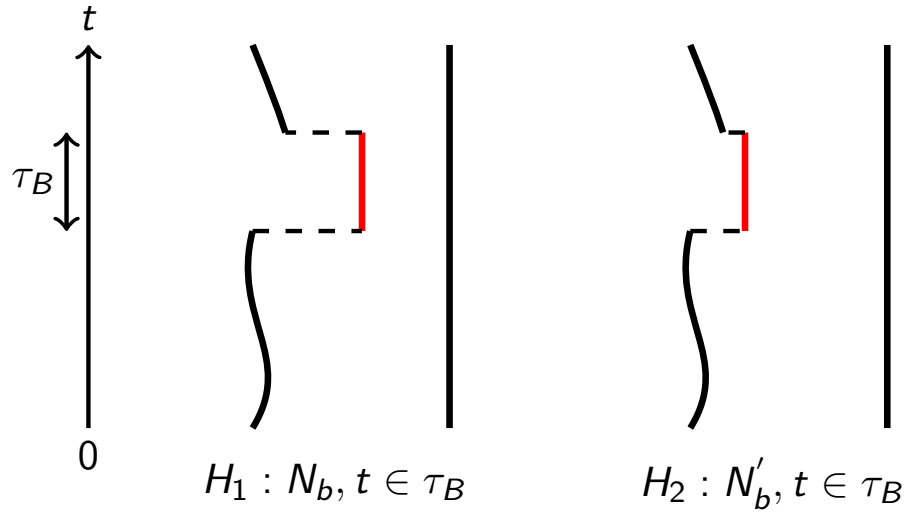
$$d_{KL}(P_1||P_2) = \int_D f_{P_1} \log \frac{f_{P_1}}{f_{P_2}} \, d\mu$$
$$= \sum_{x \in D} p_1(x) \log \frac{p_1(x)}{p_2(x)}$$

- For $P_1^m = P_1 \times \ldots \times P_1$ and $P_2^m = P_2 \times \ldots \times P_2$

$$d_{TV}^2(P_1^m, P_2^m) \leq \frac{m}{2} \cdot d_{KL}(P_1||P_2) \tag{6}$$

Results: bounds on bottleneck

- $d(\eta_1, \eta_2) = \tau_B \cdot (N_b' - N_b)$
- maximize $\varepsilon := N_b' - N_b$



- L coalescent time data $s_1, ..., s_L$ sampled indep. from η

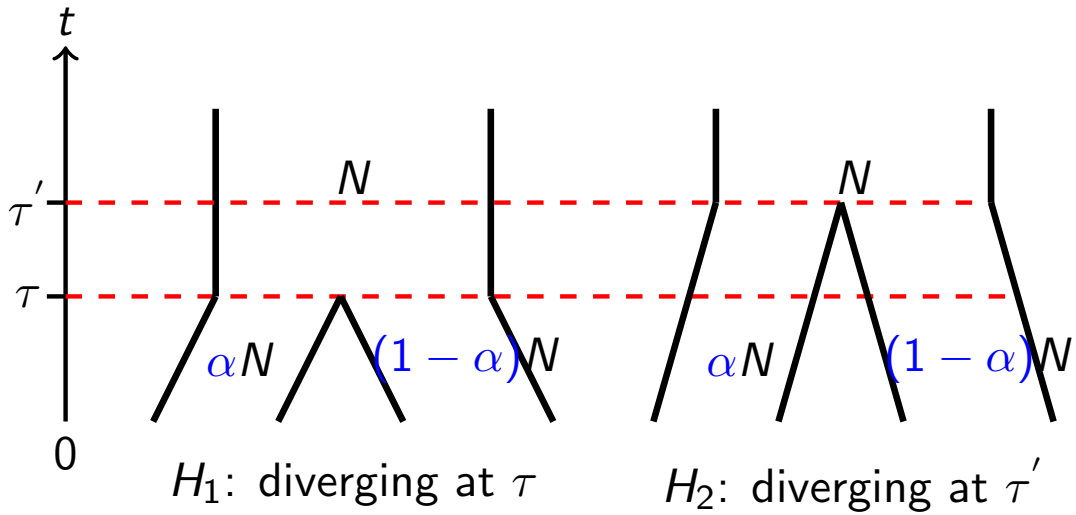
$$\inf_{\hat{\mathcal{E}}} \sup_{\eta_1, \eta_2} \mathbb{E} d(\hat{\mathcal{E}}, \mathcal{E}(\eta)) \geq \frac{\tau_B N_b}{4\sqrt{2L}} \cdot \min\{\frac{2}{\tau_B}, \frac{N_b N_b'}{N_b + N_b'}\}$$

- S segregating sites data $X_1, ..., X_S$ sampled indep. from η

$$\inf_{\hat{\chi}} \sup_{\eta_1, \eta_2} \mathbb{E} d(\hat{\chi}, \chi(\eta)) \geq \frac{4}{27} \frac{\tau_B N_b}{S}$$

Results: Hellinger distance on diverging populations

- Def: diverging population $\zeta[\tau]$ split at time τ
- sample size n_1, n_2 from each sub-populations



- $n_1 = n_2 = 1$.
- $s, s^{'}$ are coalescent time data induced from $\zeta[\tau]$ and $\zeta[\tau^{'}]$

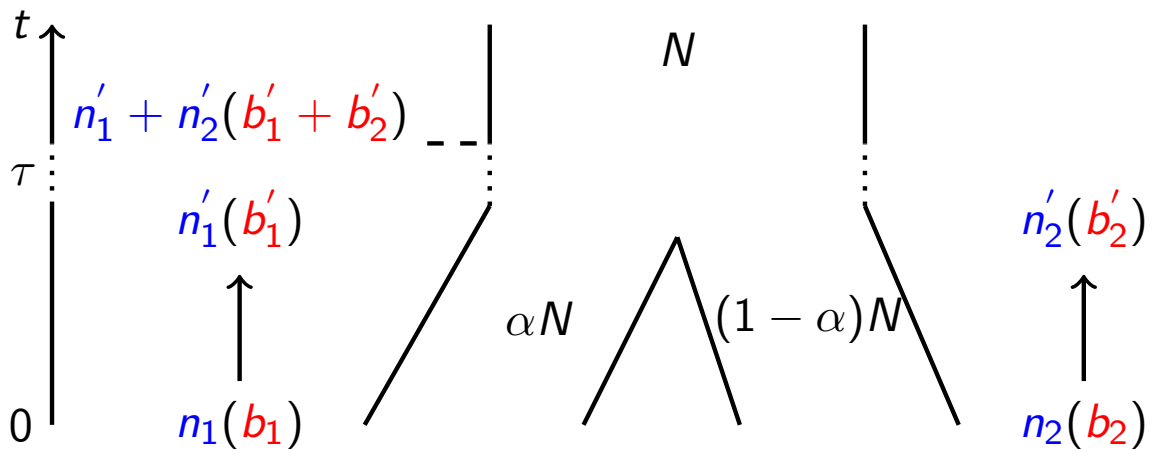
$$d_H^2(s, s) \leq \frac{\tau^{'} - \tau}{2N}$$

Results: Stochastic Flows of Mutant Genes

- Given sample size of (n_1, n_2) with mutant genes (b_1, b_2) resp., and $(n_1^{'}, n_2^{'})$ lineages left at τ, prob. of $(b_1^{'}, b_2^{'})$ mutant genes

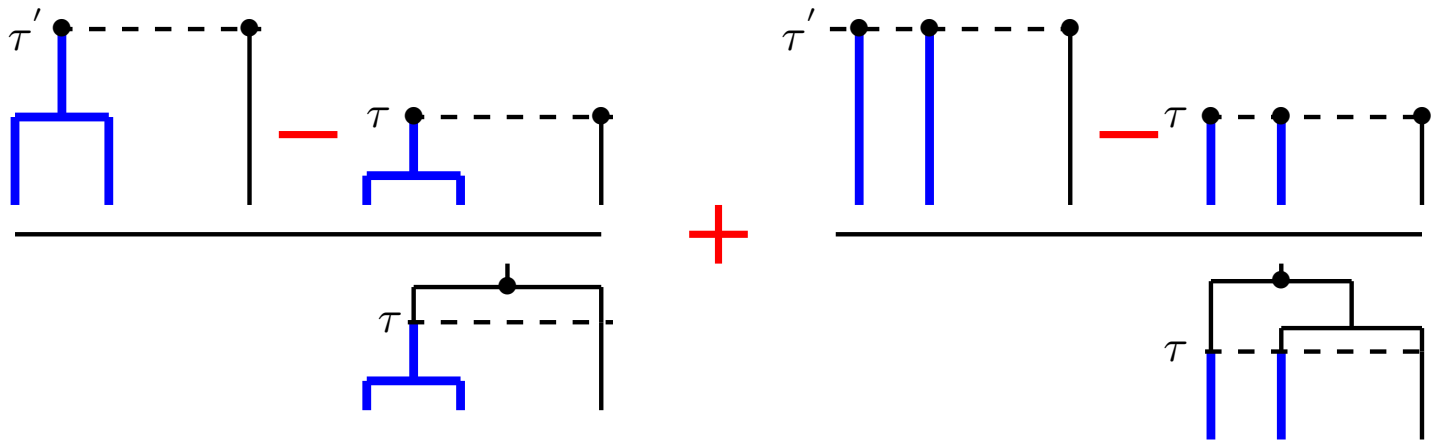
$$\frac{\binom{b_1}{b_1^{'}}\binom{n_1-b_1}{n_1^{'}-b_1^{'}}}{\binom{n_1}{n_1^{'}}} \cdot \frac{\binom{b_2}{b_2^{'}}\binom{n_2-b_2}{n_2^{'}-b_2^{'}}}{\binom{n_2}{n_2^{'}}}$$

where $b_1^{'} \leq \min\{b_1, n_1^{'}\}$, $b_2^{'} \leq \min\{b_2, n_2^{'}\}$.



Results: Kullback-Leibler distance on diverging population

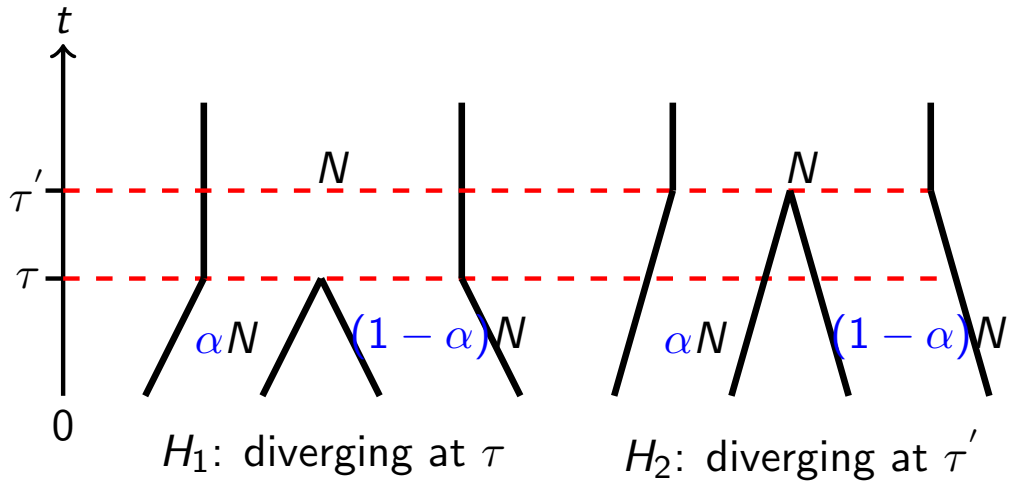
- $d_{KL}(\chi^{(\tau)}_{(n_1,n_2)} || \chi^{(\tau')}_{(n_1,n_2)}) \leq \sum_{n_1'=n_2'=1}^{n_1+n_2} \frac{C(\tau')-C(\tau)}{\bar{T}^{(\tau)}_{(n_1',n_2')}}$
- Example: $n_1 = 2$ and $n_2 = 1$, then $n_1' = 1$ or 2, $n_2' = 1$



- with sample size (n_1, n_2), there are $n_1 \cdot n_2$ terms

Results: bounds on diverging population

- $d(\zeta, \zeta^{'}) = \alpha N \cdot (\tau^{'} - \tau)$
- maximize $\upsilon := \tau^{'} - \tau$



- Given L coalescent time $(\hat{\mathcal{E}})^{n_1+n_2,L} = (\hat{\mathcal{E}})^{n_1+n_2,L}(s_1, ..., s_L)$,

$$\inf_{\hat{\mathcal{E}}} \sup_{\zeta[\tau],\zeta[\tau']} \mathbb{E}d(\hat{\mathcal{E}}, \mathcal{E}(\zeta)) \geq \frac{4\alpha \cdot N^2}{27(1-\alpha)} \cdot \frac{1}{L}$$

- Given S segregating sites $\chi^{(\zeta[t])}_{(n_1,n_2)} := (\chi^{(\eta_1)}_{n_1}, \chi^{(\eta_2)}_{n_2})$, $(2,1)$ samp.

$$\inf_{\hat{\chi}} \sup_{\zeta[\tau],\zeta[\tau']} \mathbb{E}d(\hat{\chi}, \chi(\zeta)) \geq \alpha N \cdot (\frac{3}{3\tau + 3N} + \frac{5}{5\tau + 4N})^{-1} \cdot \frac{1}{S}$$

Thank you!