# THE TIME BACK TO THE MOST RECENT COMMON ANCESTOR

Martin Möhle

University of Tübingen

# Outline

1. Exchangeable Population Models
2. Discrete Ancestral Process
3. Time back to the MRCA (Solution, Bounds, Algorithms)
4. Examples
5. Asymptotics, Coalescent
6. Conclusions, Remarks, Discussion

## An Exchangeable Population Model (Cannings)

Non-overlapping generations $r = \ldots, -1, 0, 1, \ldots$

Population size $N$, i.e.
$N$ genes (or individuals) in each generation

$\nu_i^{(r)}$ = number of offspring of gene $i$ in generation $r$

$$\sum_{i=1}^{N} \nu_i^{(r)} = N$$

No mutation, no selection (neutrality)

## Additional Assumptions for the Offspring

Exchangeability: $(\nu_{\pi 1}^{(r)}, \ldots, \nu_{\pi N}^{(r)}) \stackrel{d}{=} (\nu_1^{(r)}, \ldots, \nu_N^{(r)}) \ \forall\ \pi, r$

Homogeneity: $(\nu_1^{(r)}, \ldots, \nu_N^{(r)})_r$ i.i.d.

Avoid the trivial model ($\nu_1^{(r)} \equiv 1$)

For convenience: $(\nu_1, \ldots, \nu_N) := (\nu_1^{(0)}, \ldots, \nu_N^{(0)})$

**Classical Examples:**
**Wright-Fisher:** $(\nu_1, \ldots, \nu_N) \stackrel{d}{=} \mathrm{Multinomial}(N, \frac{1}{N}, \ldots, \frac{1}{N})$
**Moran:** $(\nu_1, \ldots, \nu_N)$ = random permutation of $(2, 0, \underbrace{1, \ldots, 1}_{N-2})$

## Backward Process

Sample $n$ $(\leq N)$ genes from generation $0$. For $r = 0, 1, \ldots$ define

$$\boxed{\mathcal{D}_r \ := \# \text{ of } \textbf{ancestors } r \text{ generations } \textbf{backward} \text{ in time}}$$

**Backward process** $\mathcal{D} := (\mathcal{D}_r)_{r=0,1,\ldots}$ is a time-homogeneous Markovian death process.

| | |
|---|---|
| state space: | $\{0, \ldots, n\}$ |
| initial state: | $n$ |
| absorbing states: | $1$ (and $0$) |
| transition matrix: | $P_N = (p_{ba})$ with entries |

$$\begin{aligned} p_{ba} \ = \ p_{ba}(N) \ &:= \ P(\mathcal{D}_r = a \,|\, \mathcal{D}_{r-1} = b) \\ &= \ \frac{\binom{N}{a}}{\binom{N}{b}} \sum_{\substack{b_1,\ldots,b_a \in \mathbb{N} \\ b_1+\cdots+b_a=b}} \mathrm{E}\left(\binom{\nu_1}{b_1} \cdots \binom{\nu_a}{b_a}\right) \\ &= \ \frac{b!}{a!} \sum_{\substack{b_1,\ldots,b_a \in \mathbb{N} \\ b_1+\cdots+b_a=b}} \frac{\Phi_a(b_1,\ldots,b_a)}{b_1! \cdots b_a!} \end{aligned}$$

with

$$\boxed{\Phi_a(b_1,\ldots,b_a) \ := \ \frac{(N)_a}{(N)_b} \mathrm{E}\Big((\nu_1)_{b_1} \cdots (\nu_a)_{b_a}\Big)}$$

(probability for a $(b_1, \ldots, b_a)$-merger)

Notation: $(x)_k := x(x-1) \cdots (x-k+1)$

## Ancestral Process

Sample $n$ ($\leq N$) genes from generation $0$. For $r = 0, 1, \ldots$ define a random relation $\mathcal{R}_r = {}_{(N)}\mathcal{R}_r^{(n)}$ on $\{1, \ldots, n\}$ via

$$\boxed{(i,j) \in \mathcal{R}_r \quad :\Longleftrightarrow \quad \begin{array}{l}\text{the chosen genes } i \text{ and } j \\ \text{have a common ancestor} \\ r \text{ generation backward in time}\end{array}}$$

**Ancestral process** $\mathcal{R} := (\mathcal{R}_r)_{r=0,1,\ldots}$ is a time-homogeneous Markov chain.

| | |
|---|---|
| state space: | $\mathcal{E}_n :=$ set of equivalence relations on $\{1, \ldots, n\}$ |
| initial state: | $\Delta := \{(i,i) \mid 1 \leq i \leq n\}$ |
| absorbing state: | $\Theta := \{(i,j) \mid 1 \leq i,j \leq n\}$ |
| typical path: | $\Delta = \mathcal{R}_0 \subseteq \mathcal{R}_1 \subseteq \cdots \subseteq \mathcal{R}_r \subseteq \Theta$ |
| transition matrix: | $P_N := (p_{\xi\eta})_{\xi,\eta \in \mathcal{E}_n}$ with entries |

$$\boxed{p_{\xi\eta} := P(\mathcal{R}_r = \eta \mid \mathcal{R}_{r-1} = \xi) = \Phi_a(b_1, \ldots, b_a)} \quad \text{for } \xi \subseteq \eta$$

$a := |\eta| :=$ number of equivalence classes of $\eta$
$b := |\xi| :=$ number of equivalence classes of $\xi$ $(= b_1 + \cdots + b_a)$
$b_1, \ldots, b_a =$ group sizes of merging classes of $\eta$

## Coalescence Probability

$$
\begin{aligned}
c_N \;&:=\; P(\text{Two individuals have same parent}) \\
&=\; p_{21} \;=\; \Phi_1(2) \;=\; \frac{\mathrm{E}(\nu_1(\nu_1-1))}{N-1} \;>\; 0
\end{aligned}
$$

## Eigenvalues of the Transition Matrix (Cannings)

$$
\lambda_i := p_{ii} = \Phi_i(1,\ldots,1) = \mathrm{E}(\nu_1\cdots\nu_i), \qquad i\in\{0,\ldots,N\}
$$

## Properties of the Eigenvalues

$1=\lambda_0=\lambda_1>\cdots>\lambda_k=\cdots=\lambda_N\geq 0$ for some $1\leq k\leq N$

$k=1 \iff P(\nu_1=1)=1$

$N_e \;=\; \dfrac{1}{c_N} \;=\; \dfrac{1}{1-\lambda_2}$ (Effective Population Size)

## Expected Time Back to the MRCA, Exact Solutions

$$
\begin{aligned}
T_n &:= \text{Number of generations back to the MRCA} \\
&= \inf\{r \in I\!N_0 \,|\, \mathcal{D}_r = 1, \mathcal{D}_0 = n\} \quad \text{(absorption time)}
\end{aligned}
$$

$$
\begin{aligned}
m_n &:= \text{Expected time back to the MRCA} \;=\; \mathrm{E}(T_n) \\
\mu_i &:= 1 - \lambda_i
\end{aligned}
$$

### Recursion

$$
m_1 = 0 \text{ and } \boxed{m_i \;=\; \frac{1}{\mu_i}\left(1 + \sum_{j=2}^{i-1} p_{ij}\, m_j\right)} \text{ for } i = 2, 3, \ldots, N
$$

### Algebraic Solution

$$
\boxed{m \;=\; (I - A)^{-1} a \;=\; \sum_{k=0}^{\infty} A^k a}
$$

with $m := (m_2, \ldots, m_N)'$, $a := (1, \ldots, 1)'$ and $A := (p_{ij})_{2 \le i,j \le N}$

### Explicit Solution

$$
\boxed{m_n = \sum_{k=1}^{n-1} \sum_{1=i_0<i_1<\cdots<i_{k-1}<i_k=n} \frac{p_{i_k,i_{k-1}}}{\mu_{i_k}} \cdots \frac{p_{i_1,i_0}}{\mu_{i_1}} \left(\frac{1}{\mu_{i_1}} + \cdots + \frac{1}{\mu_{i_k}}\right)}
$$

**Remark:** Recursion and explicit solution are numerically ‘unstable’ if some of the eigenvalues are very close to one (which is usually the case)

## Numerical Approximation

Iteration of the transformation $\boxed{Tm := Am + a}$
$m^{(0)}$ arbitrary, $m^{(1)} := Tm^{(0)}$, $m^{(2)} := Tm^{(1)}$, ...

Contraction property: $\|T^k\| < 1$ for sufficiently large $k$

Geometrical convergence: There exist $C > 0$ and $0 < q < 1$ such that $\boxed{\|m^{(n)} - m\| < Cq^n}$ for all $n$ (Banach fixed-point theorem)

### Remarks.

- Advantage: Iteration is self-correcting;

  approximates $m_2, \ldots, m_N$ simultaneously

- Disadvantage: Time-consuming (for example for $q$ close to one or large $N$)

- Typical behaviour for expectations of absorption times

## Upper Bounds for the Expected Time Back to the MRCA

For all $n \in \{1, \ldots, N\}$

### Bound 1

$$\boxed{m_n \ \leq \ \sum_{i=2}^{n} \frac{1}{\mu_i} \ =: \ B_1}$$

### Bound 2

$$\boxed{m_n \ \leq \ \sum_{i=2}^{n} \frac{1}{d_i} \ =: \ B_2}$$

with $d_i := i - \mathrm{E}^i(\mathcal{D}_1) = \sum_{j=1}^{i} (i-j) p_{ij}$

### Conjecture

$$\boxed{m_n \ \leq \ 2N_e \left(1 - \frac{1}{n}\right) \ =: \ B_3}$$

**Remarks:** $B_3 \leq B_2 \leq B_1$

$B_3$ is the asymptotic value known from coalescent theory

## Conjecture holds under an Additional Assumption

**Theorem.** Assume that

$$\boxed{\mathrm{E}^i\left(\frac{1}{\mathcal{D}_1}\right) \;\geq\; \frac{1}{i} + \frac{1}{c_N}} \qquad (A)$$

for all $i \in \{2, \ldots, N\}$. Then

$$\boxed{m_n \;\leq\; 2N_e\left(1 - \frac{1}{n}\right)} \qquad (B)$$

for all $n \in \{1, \ldots, N\}$. If the inequality in $(A)$ is strict $\forall\ i \in \{3, \ldots, N\}$, then the inequality in $(B)$ is strict $\forall\ n \in \{3, \ldots, N\}$.

**Remarks.**

1. Assumption $(A)$ is far from being intuitive, but can be motivated from the Moran model.
2. Helpful formula: $\boxed{\mathrm{E}^i\left(\frac{1}{\mathcal{D}_1}\right) \;=\; \frac{1}{(N)_i}\sum_{j=1}^{N}\frac{\mathrm{E}((C_j)_i)}{j}}$

$$\begin{aligned} C_j \;&:=\; \nu_1 + \cdots + \nu_j \\ &=\; \#\text{ of children of the individuals } 1, \ldots, j \end{aligned}$$

$$(x)_i \;:=\; x(x-1)\cdots(x-i+1)$$

## Examples

**Example 1: Moran model**
Effective population size: $N_e = \dfrac{N(N-1)}{2}$
Assumption (A), and hence bound (B) hold with equality

**Example 2: Wright-Fisher model**
Effective population size: $N_e = N$

$$\boxed{\mathrm{E}^i\left(\frac{1}{\mathcal{D}_1}\right) \;=\; \frac{1}{N}\sum_{j=1}^{N}\left(\frac{j}{N}\right)^{i-1} \;>\; \frac{1}{i}+\frac{1}{2N}}$$

$\Rightarrow$ Theorem is applicable $\Rightarrow$

$$\boxed{m_n \;<\; 2N\left(1-\frac{1}{n}\right)} \quad \forall\, n \in \{3,\ldots,N\}$$

In particular, $m_N \;<\; 2N-2$ for $N \geq 3$.
(expected time back to the MRCA of the total population)

**Example 3. Dirichlet compound multinomial model**
$(\nu_1, \dots, \nu_N)$ multinomially distributed with random probabilities $p_1, \dots, p_N$, symmetric Dirichet-distributed with parameter $\alpha > 0$.

Offspring distribution on $\Delta_N := \{(k_1, \dots, k_N) \mid k_1 + \dots + k_N = N\}$:

$$\boxed{P(\nu_1 = k_1, \dots, \nu_N = k_N) \;=\; \frac{N!}{k_1! \cdots k_N!} \frac{[\alpha]_{k_1} \cdots [\alpha]_{k_N}}{[N\alpha]_N}}$$

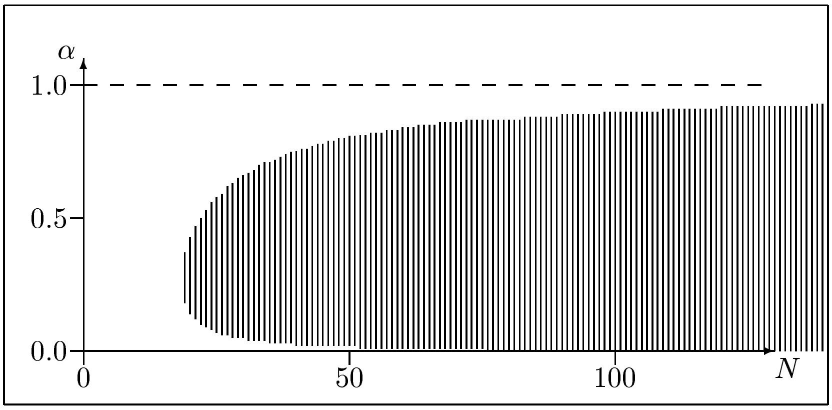
with $[x]_i := x(x+1) \cdots (x+i-1)$

**Special cases:**

- $\underline{\alpha \to \infty}$: Wright-Fisher model
- $\underline{\alpha = 1}$: $(\nu_1, \dots, \nu_N)$ uniformly distributed on the simplex $\Delta_N$
- $\underline{\alpha \to 0}$: $(\nu_1, \dots, \nu_N)$ = random permutation of $(N, 0, \dots, 0)$

**Results:**

- Numerical studies based on the iteration algorithm support the conjecture that $(B)$ is satisfied for all $N, n, \alpha$.
- $\mathrm{E}^i(1/\mathcal{D}_1) = \sum_{j=1}^N [j\alpha]_i/(j[N\alpha]_i)$ can be computed numerically and compared with $1/i + c_N/2$ quite easily:




Region of parameters $N$ and $\alpha$ where $(A)$ is not satisfied

## Asymptotics, Kingman coalescent

$T_n = T_{n,N}$ and $m_n = m_{n,N}$ depend on the total population size $N$ and on the sample size $n$

**Theorem:**
The following conditions are equivalent:

(i) $\boxed{\lim_{N\to\infty} \frac{p_{31}(N)}{p_{21}(N)} = 0}$

(triple mergers of ancestral lines are asymptotically negligible in comparison with binary mergers)

(ii) $\boxed{(\mathcal{R}_{[t/c_N]})_{t\geq 0} \Longrightarrow (R_t)_{t\geq 0}}$

(convergence to the Kingman coalescent in the Skorohod topology)

(iii) $\boxed{\lim_{N\to\infty} c_N \, \mathrm{E}(T_{n,N}) = \mathrm{E}(T_n) = 2\left(1 - \frac{1}{n}\right) \;\; \forall\; n \in I\!N}$

$T_{n,N} := \inf\{r \in I\!N_0 \,|\, \mathcal{D}_r = 1, \mathcal{D}_0 = n\}$

$T_n := \inf\{t > 0 \,|\, |R_t| = 1, |R_0| = n\}$

(convergence of the scaled expected number of generations back to the most recent common ancestor)

(iv) $\boxed{\lim_{N\to\infty} c_N \, \mathrm{E}(T_{3,N}) = \mathrm{E}(T_3) = \frac{4}{3}}$

(convergence as in (iii), but only for sample size $n = 3$)

## General Asymptotics
## Coalescent With Simultaneous Multiple Collisions

**Theorem:**
Assume that all the limits

$$g_{ij} \;:=\; \lim_{N\to\infty} \frac{p_{ij}(N)}{c_N}, \qquad i,j \in I\!N, j < i$$

exist. Then the following holds:

(i) $\boxed{(\mathcal{D}_{[t/c_N]})_{t\geq 0} \Longrightarrow D := (D_t)_{t\geq 0}}$

(convergence in the Skorohod topology to a time-continuous death process $D$ with with infinitesimal rates $g_{ij}$)

Remark: $D$ is the death process of a coalescent with simultaneous multiple collisions of ancestral lineages

(ii) $\boxed{\lim_{N\to\infty} c_N \, \mathrm{E}(T_{n,N}) = \mathrm{E}(T_n) \;\forall\; n \in I\!N}$

$T_{n,N} := \inf\{r \in I\!N_0 \,|\, \mathcal{D}_r = 1, \mathcal{D}_0 = n\}$

$T_n := \inf\{t > 0 \,|\, D_t = 1, D_0 = n\}$

(convergence of the scaled expected number of generations back to the most recent common ancestor)

## Conclusions

- For many haploid exchangeable models with total population size $N$ the expected number of generations $m_{n,N}$ back to the MRCA of a sample of size $n \in \{1, \ldots, N\}$ is bounded above by the asymptotic value derived from coalescent theory, i.e.

$$\boxed{m_{n,N} \;\leq\; 2N_e\left(1 - \frac{1}{n}\right)} \qquad \text{(B)}$$

- **Conjecture:** The bound (B) holds for the complete Cannings class of exchangeable haploid models. We have verified this conjecture under an additional assumption, which is satisfied for many, but not all models. General proof might be based on coupling methods?

- Asymptotic results can be extended to coalescent processes with simultaneous multiple collisions of ancestral lineages.