

## Supplementary material

### 1 Restriction rules for ARGs

In the following a formal description of the restriction rules (introduced in Section 2.1.3) imposed to an ARG by a classification is given:

- A pair of sequences can only coalesce if
  - both sequences belong to the same subtype ( $\times$ ) or CRF ( $\times$ ), or
  - the sequences ( $\star$ )
    - \* are the only sequences of their subtype left or
    - \* belong to more than one subtype and are the parent of a coalescent event.

Here, a sequence generated by a coalescent event is defined to belong to the same subtype(s) or CRF, resp., as its children, i.e.,

- the parent of two subtype A sequences belongs to subtype A,
- the parent of two CRF1 sequences belongs to CRF1,
- the parent of one subtype A and one subtype B sequence belongs to subtype A and subtype B.

A sequence generated by a recombination event belongs to the subtype(s) its segments belong to.

- The sequences of a CRF must all coalesce before they undergo a recombination event. Only the last sequence left ( $\times$ ) is allowed to recombine. (Multiple) breakpoints have to be chosen such that the parental subtypes get separated and recombination events have to take place until all parental subtypes are separated.

### 2 MCMC details and move types

The Markov chain Monte Carlo algorithm for ARGs fulfilling the restrictions imposed by a given classification is described, including proposal mechanism used.

Let  $G$  and  $H$  be ARGs. Then the change from  $G$  to  $H$  is accepted if

$$r := \frac{P(D|H)P(H|\Theta, r, R)Q(H, G)}{P(D|G)P(G|\Theta, r, R)Q(G, H)} > u$$

where  $u$  is sampled from a uniform distribution on  $[0, 1]$ .  $Q(G, H)$  denotes the proposal probability specifying the probability to generate  $H$  in the next step given  $G$  is the current ARG.

Note that, if

$$Q(H, G) = CP(G|\Theta, r, R), \quad Q(G, H) = CP(H|\Theta, r, R) \quad (1)$$

with  $C > 0$ ,

$$r = \frac{P(D|H)}{P(D|G)} \quad (2)$$

Hence, if a proposal ARG is sampled with respect to a conditional coalescent distribution,  $r$  only depends on the probability of the data with respect to the genealogy.

In total, we apply five different types of proposal mechanisms (moves), chosen such that the whole space of legal (i.e. fulfilling the classification-given restrictions) ARGs can be entirely be traversed and the MCMC algorithm converges fast into areas of ARGs with high likelihood. The last three of them

fulfill (2). Except the first move (which is a global rescaling operation), all perform local rearrangements, i.e., among all subgraphs fulfilling specific topological and typological properties one subgraph is chosen randomly and is rearranged.

In the description of the moves, we will use the following notation:

- Given an ARG  $G$ , its nodes are denoted by  $N = N_G$ . Let  $\text{Tip}(G)$  be the tip nodes of  $G$  and  $\text{Int}(G) = N_G \setminus \text{Tip}(G)$  the internal nodes of  $G$ . Denoting the subtypes of the classification by  $S = \{S_1, \dots, S_{m_p}\}$  and its CRFs by  $C = \{C_1, \dots, C_{m_r}\}$ , we define  $\text{Type} : N \rightarrow S \cup C \cup \{\text{Imp}\}$

$$n \rightarrow \begin{cases} S_i, & \text{if } n \text{ belongs only to subtype } S_i \\ C_i, & \text{if } n \text{ belongs to CRF } C_i \\ \text{Imp}, & \text{else} \end{cases}$$

where Imp is a symbol standing for “impure”.

- The child(ren) and parent(s), resp., of a node  $n \in N$  is denoted by  $C(n) \in \mathcal{P}(N)$  and  $P(n) \in \mathcal{P}(N)$ , resp. If  $n$  has only one child or parent, resp.,  $C(n)$  or  $P(n)$ , resp., are also interpreted as elements of  $N$ . If  $n$  has two children or parents, resp., they are denoted by  $C_1(n)$  and  $C_2(n)$  or  $P_1(n)$  and  $P_2(n)$ , resp. In case  $n$  has only one child, it has to have a spouse, which is denoted by  $S(n)$ . Furthermore, we define

$$P^d(n) := \begin{cases} P(n), & \text{if } d = 1 \text{ and } \#P(n) = 1 \\ P(P^{d-1}(n)), & \text{if } d > 1 \text{ and } \#P(P^{d-1}(n)) = 1 \\ \text{undefined}, & \text{else} \end{cases}$$

for  $d \in \mathbb{N}$ . Moreover, the age (i.e. time of generation) of  $n \in N$  is denoted by  $T(n)$ .

- The container of  $n \in N$  is defined by

$$B(n) := \begin{cases} \{n\}, & \text{if } n \in \text{Tip}(G) \text{ or } \#C(n) = 2 \\ \{n, S(n)\}, & \text{if } S(n) \text{ is defined} \end{cases}$$

(“B” stands for “Box”). We denote the set of all containers of  $G$  by  $B = B_G$ , i.e.,

$$B := \{B(n) : n \in N\}.$$

In detail, the five moves are:

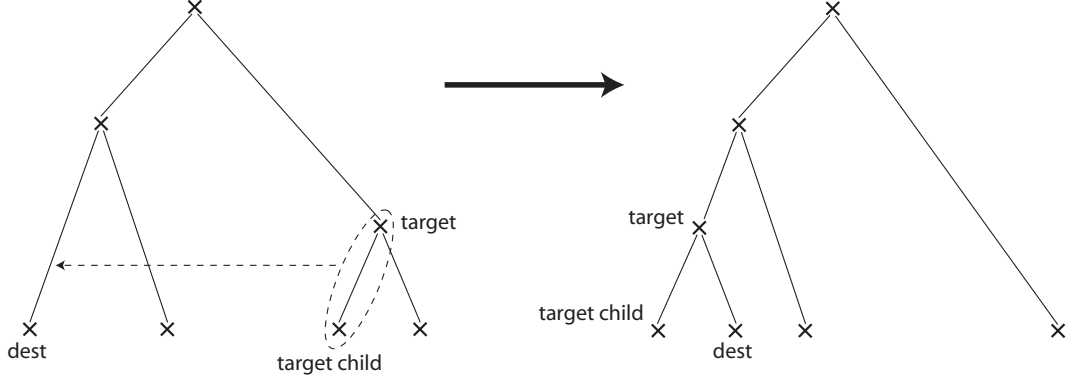
1. Scaling move: For all non-tip nodes  $n$ ,  $T(n)$  is multiplied by  $c \sim U([1 - \delta, \frac{1}{1-\delta}])$  with  $0 < \delta \ll 1$ .
2. Branch-Swapping move: This move is similar to the Wilson-Balding move described in (1). Among the set

$$\{n \in N : \#C(n) = 2\}$$

one (target) node  $n_t$  is chosen randomly. Then, one (destination) node  $n_d$  is chosen randomly among the set

$$\{n \in N : \#P(n) = 1, \text{Type}(n) = \text{Type}(n_t), \\ T(n) < T(n_t) < T(P(n))\},$$

Finally,  $n_t$  is moved with one of its children such that  $n_t$  becomes the parent of  $n_d$  and the child of the former parent of  $n_d$ , and the other child of  $n_t$  becomes the child of the former parent of  $n_t$  (see Figure 1).



**Figure 1.** The branch-swapping move

3. Node age move: Let  $b_1, \dots, b_n$  be the non-tip containers ordered by age, i.e.,  $T(b_i) \leq T(b_j)$  for  $i < j$ . For this move, a non-root container  $b_i \in B$  is chosen randomly among  $b_1, \dots, b_{n-1}$ . Then one of the three following moves are carried out with equal probability: The age of the container is drawn from a conditional coalescent distribution (with given population parameters) conditioned on that
  - (a) the order of the containers does not change
  - (b)  $T(b_1) \leq \dots \leq T(b_{i+1}) \leq T(b_i) \leq T(b_{i+2}) \leq \dots \leq T(b_n)$  (this move is forbidden if  $b_{i+1}$  contains the parent(s) of  $b_i$ )
  - (c)  $T(b_1) \leq \dots \leq T(b_{i-2}) \leq T(b_i) \leq T(b_{i-1}) \leq \dots \leq T(b_n)$  (this move is forbidden if  $i = 1$  or  $b_{i+1}$  contains a child of  $b_i$ )

The move under b) is also called an “up move”, the one under c) a “down move”.

4. Coalescent move: A (target) node  $n_t$  is chosen at random from

$$\{n \in N : \#C(n) = 2, \#P(n) = 1\}.$$

The so-called neighborhood of rearrangement consists of the target node, its children, parent, and parent’s other child. This move makes changes of two kinds: it may reassign the three children among target and parent, and it modifies the branch lengths within the neighborhood. The new branch lengths must remain within the constraints imposed by the times of the three children and of the parent’s ancestor (if existing); these times define the boundaries of the neighborhood. Conceptually, the portion of the genealogy involving these nodes is erased and must now be redrawn. This move is based on the rearrangement move introduced by (2) (Large parts of this description were taken from (2)). Technical details about this move for ARGs without recombination events are described in (2), our extension to ARGs with recombination events is not shown due to the length of our deduction.

5. Recombination move: This is the most complicated move and is introduced in order to reorder nodes involved in recombination events. Among

$$\{n \in N : \#P(n) = 2\}$$

a (target) node  $n_t$  is chosen randomly. Let the sets  $\{R_i\}_{i \in \mathbb{N}_0}$  and  $R$  be defined by

$$\begin{aligned} R_0 &:= \{n_t\}, \\ R_i &:= \{n \in N : \exists n_0 \in R_{i-1} : \#P(n_0) = 2, n \in P(n_0)\}, \\ &\quad i \in \mathbb{N}, \\ R &:= \bigcup_{i \in \mathbb{N}} R_i \end{aligned}$$

and

$$H := \{n \in N : \exists n_0 \in R : \#P(n_0) = 1, P(n_0) = n\}$$

(cf. Figure 2a). All nodes belonging to  $R$  and  $H$  (except  $n_t$ ) are removed from the ARG and

$$\begin{aligned} \forall h \in H \forall n_0 \in C(h), n_0 \notin R : P(n_0) &\leftarrow P^d(n_0), \\ d &= \min\{i \geq 2 : P^i(n_0) \notin H\} \end{aligned}$$

(cf. Figure 2b). Realize that  $d = 2$  if no unknown subtype occurs and at least two subtypes have to be present in order to make this move work. Now, denote by

$$(\{S_1, \dots, S_n\} \rightarrow \{S_{i_1}, \dots, S_{i_{n_1}}\}, \{S_{j_1}, \dots, S_{j_{n_2}}\})$$

a recombination event which separates the subtypes  $\{S_1, \dots, S_n\}$  into the subtypes  $\{S_{i_1}, \dots, S_{i_{n_1}}\}$  and  $\{S_{j_1}, \dots, S_{j_{n_2}}\}$  (called R-event) and by

$$s \rightarrow T$$

the event of a node belonging to subtype  $s$  being connected to the rest of the ARG by a coalescent event (called C-event). Let  $M$  be the set of finite sequences of R- and C-events such that, if carried out chronologically on  $n_t$ , lead to a legal ARG. As next step of the move,  $m \in M$  is chosen randomly and  $n_t$  is reconnected according to  $m$ , where the age of the newly generated nodes is chosen randomly from a simply sampleable distribution (not the conditional coalescent distribution). E. g.,

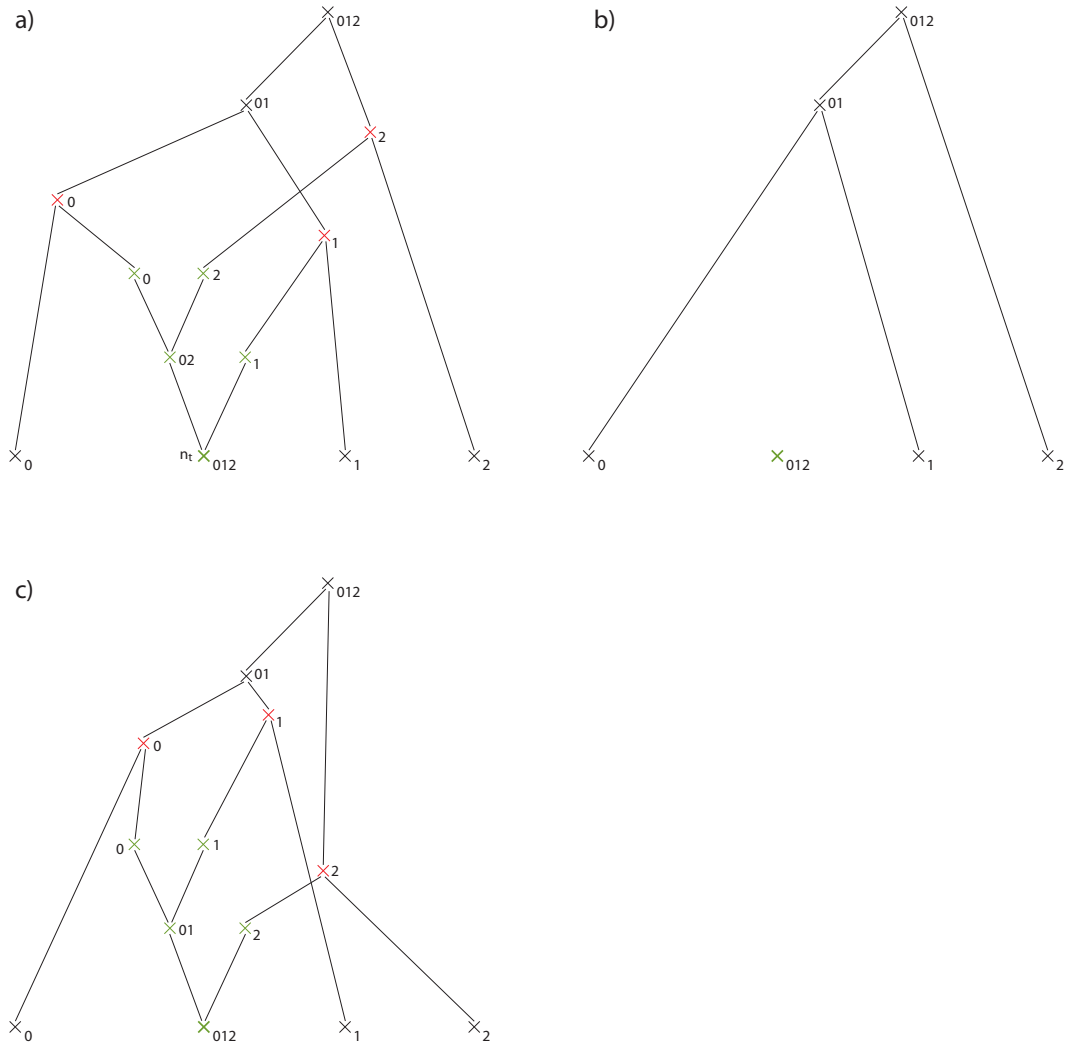
$$\begin{aligned} \{0, 1, 2\} &\rightarrow \{0, 1\}, \{2\}, \quad 2 \rightarrow T, \quad \{0, 1\} \rightarrow \{0\}, \{1\}, \\ 0 &\rightarrow T, \quad 1 \rightarrow T \end{aligned}$$

would lead to an ARG like shown in Figure 2c. Then a fixed number of extended “node age moves” is applied to the newly generated nodes, where “extended” means that, additionally to the move described under 3., a movement of nodes of  $H$  beyond their parent and children is allowed under suitable circumstances (cf. Figure 3). In more detail, we relax the conditions (b) and (c) under 3. by allowing “up moves” also if  $b_{i+1}$  is non-root and the parent of  $b_i$  and “down moves” if  $\#C(b_{i-1}) = 2$ . Such moves are carried such that the ARGs yielded by a “recombination move” are samples with respect to a conditional coalescent distribution.

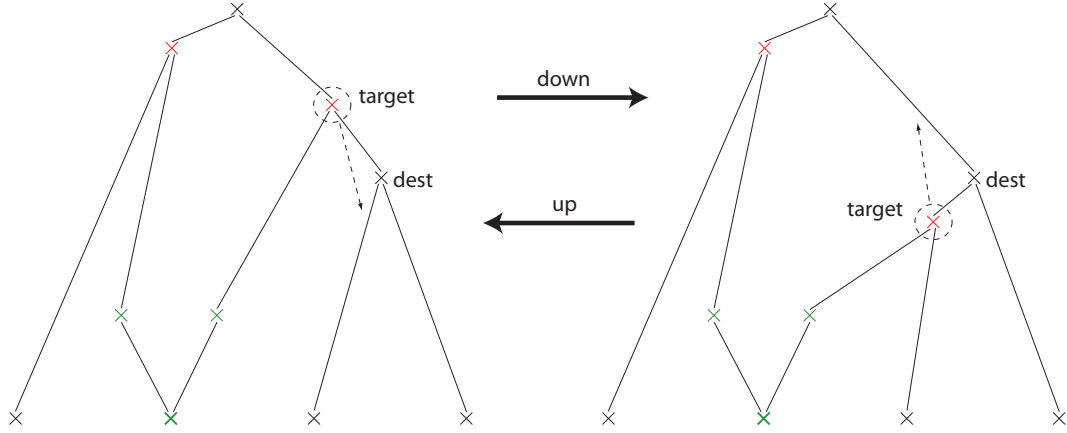
After having carried out these moves, we have to reconnect the nodes accordingly to the ARG like follows:

- “down move”: Sample  $j \sim U(1, 2)$ , and set  $P(n_d) \leftarrow P(n_t)$ ,  $P(n_t) \leftarrow n_d$ ,  $P(C_j(n_d)) \leftarrow n_t$
- “up move”: Let  $n_c := \{n \in C(n_t) : \#C(n) \neq 1\}$  and set  $P(n_t) \leftarrow P(n_d)$ ,  $P(n_d) \leftarrow n_t$ ,  $P(n_c) \leftarrow n_d$ .

The ARG obtained by this procedure is the result of the “recombination move”. Notice that this move would only not violate (1) if  $P(T_{m_1}) = P(T_{m_2})$  for  $m_1, m_2 \in M$ , where  $T_m$  is the set of ARGs



**Figure 2.** The “recombination move”. a) Nodes belonging to  $R$  are colored in green, the ones belonging to  $H$  in red. Numbers denote to which subtypes the nodes belong to. b) All nodes belonging to  $R$  and  $H$  except the target node have been removed and the non- $R$  children of  $H$  have been connected to the next valid ancestor. c) The removed part of the ARG has been regenerated.



**Figure 3.** The modified “up move” and “down move”. Left to right shows a “down move”, right to left an “up move”.

which could be generated according to  $m$ . But since all  $m$  involve the same number of coalescent and recombination events and we do not sample ARGs, but seek a maximum, this seems to be an acceptable compromise between exactness on the one hand and complexity and speed on the other hand. In case we intend to sample ARGs in the future, we will have to allow for moves between  $T_{m_1}$  and  $T_{m_2}$  for different  $m_1$  and  $m_2$  instead of only carrying out extended “node age moves”.

### 3 Separating and noise distance

Let

$$N_i^p = \{n \in N : \text{Type}(n) = S_i\}$$

for  $i \in \{1, \dots, m_p\}$  and

$$n_i^f = \operatorname{argmax}_{n \in N_i^p} T(n).$$

Moreover, for  $n_1, n_2 \in N$ ,  $n_1 \neq n_2$ , let  $n_{mrca}(n_1, n_2)$  be the most recent common ancestor node of  $n_1$  and  $n_2$ . Then the separating distance  $d_{sep}$  is defined by

$$d_{sep} = \sum_{i=1}^{m_p} \sum_{j \neq i} 2 \cdot T(n_{mrca}(n_i^f, n_j^f)) - T(n_i^f) - T(n_j^f)$$

and the noise distance  $d_{noise}$  by

$$d_{noise} = \sum_{i=1}^{m_p} T(n_i^f).$$

## References

1. Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W (2002) Estimating Mutation Parameters, Population History and Genealogy Simultaneously From Temporally Spaced Sequence Data. *Genetics* 161: 1307-1320.
2. Kuhner MK, Yamato J, Felsenstein J (1995) Estimating Effective Population Size and Mutation Rate From Sequence Data Using Metropolis-Hastings Sampling. *Genetics* 140: 1421-1430.