The Effect of Single Recombination Events on Coalescent Tree Height and Shape

Luca Ferretti[1,2], Filippo Disanto[1], Thomas Wiehe[1]*

1 Institut für Genetik, Universität zu Köln, Köln, Germany, 2 Center for Research in AgriGenomics, Barcelona, Spain

Abstract

The coalescent with recombination is a fundamental model to describe the genealogical history of DNA sequence samples from recombining organisms. Considering recombination as a process which acts along genomes and which creates sequence segments with shared ancestry, we study the influence of single recombination events upon tree characteristics of the coalescent. We focus on properties such as tree height and tree balance and quantify analytically the changes in these quantities incurred by recombination in terms of probability distributions. We find that changes in tree topology are often relatively mild under conditions of neutral evolution, while changes in tree height are on average quite large. Our results add to a quantitative understanding of the spatial coalescent and provide the neutral reference to which the impact by other evolutionary scenarios, for instance tree distortion by selective sweeps, can be compared.

Citation: Ferretti L, Disanto F, Wiehe T (2013) The Effect of Single Recombination Events on Coalescent Tree Height and Shape. PLoS ONE 8(4): e60123. doi:10.1371/journal.pone.0060123

Editor: Nadia Singh, North Carolina State University, United States of America

Received November 15, 2012; **Accepted** February 21, 2013; **Published** April 8, 2013

Funding: This work was supported by grants from the German Research Foundation (DFG-SFB680 and DFG-SPP1590) to TW, grant AGL2010-14822 (MICINN, Spain) to Miguel Perez-Enciso and by a Consolider Grant from the Spanish Ministry of Research, CSD2007- 00036 "Centre for Research in Agrigenomics." LF acknowledges support from CSIC (Spain) under the JAE-doc program, and by a visiting scholar grant from SFB680. The funders had no role in study design, data collection, and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: twiehe@uni-koeln.de

Introduction

Coalescent theory is a central part of modern population genetics [1–3]. It constitutes the basis of genealogical models, of statistical tests of the neutral evolution hypothesis [4] as well as of many simulation tools [5–7]. Besides application in population genetics, coalescent models and their various generalizations became an object of study in their own right in probability, graph theory and combinatorics [8–12].

The classical coalescent is a binary, rooted, unordered tree with a fixed number n of leafs. The latter is also called the *size* of the tree (Figure 1A). Such a tree can be interpreted as the genealogical history of a sample of DNA sequences, where mergers ("coalescents") of two lineages represent events of common ancestry. Thus, coalescent trees are naturally fitted with a time scale and for this reason they are sometimes called *labelled histories*. A biologically important generalization of the simple case is the coalescent with recombination. Recombination is a process by which two DNA sequences reciprocally exchange genetic material. In the coalescent framework this translates into lineage splits (Figure 1B). A split represents the un-coupling of the genealogical history of two sequence fragments. The ancestral recombination graph (ARG) [13] is a model to integrate such lineage splits into coalescent trees. Each sequence position x along the chromosome is associated with a coalescent tree T_x, which is the marginal tree of the ARG at position x. Depending on the rate of recombination, chromosomes are divided into smaller or larger sequence fragments f_i ("haplotype block") in such a way that all positions within a fragment are free of recombination and therefore have the same marginal tree T_f.

The spatial coalescent is the sequence $(T_{f_i})_i$ of coalescent trees along a sample of recombining chromosomes. Study of the spatial coalescent is of prominent interest in population genomics, since it contains information about the demographic and evolutionary history of a population. For instance, it has lately been used to infer demographic parameters in non-African human [14]. Unfortunately, the spatial coalescent is not a simple Markov process [15], complicating its probabilistic analysis and leaving many open problems to be addressed.

Here, we investigate the impact of single recombination events upon some measures of tree topology and shape. By *topology* we mean the branching pattern of a tree; by *shape* we mean its topology and branch lengths. In particular, we ask how recombination affects tree height and tree (im-)balance. The latter is measured by the difference in size of the left and right subtrees emerging from the root or any internal node. Depending on when and where a recombination event occurs, the effect on altering tree structure may be drastic, mild or completely silent. Informally, drastic events are those which lead to a large change of tree height or balance. These are events which typically involve splits by recombination of the branches emerging from the root of the tree. As such they may strongly affect the genealogical structure of haplotypes. Identifying and characterizing these events is very informative for population genetic inference. Mild events are typically those which occur along very recent branches, close to the leafs of the tree. They do not, or only mildly, affect haplotype structure and mutation frequency spectrum. Interestingly, there is a non-negligible portion of recombination events which do not alter tree topology, i.e. the branching pattern. We call these events *silent*. Sometimes, also the branch lengths remain unchanged; we call these events *hidden* (Figure 2).

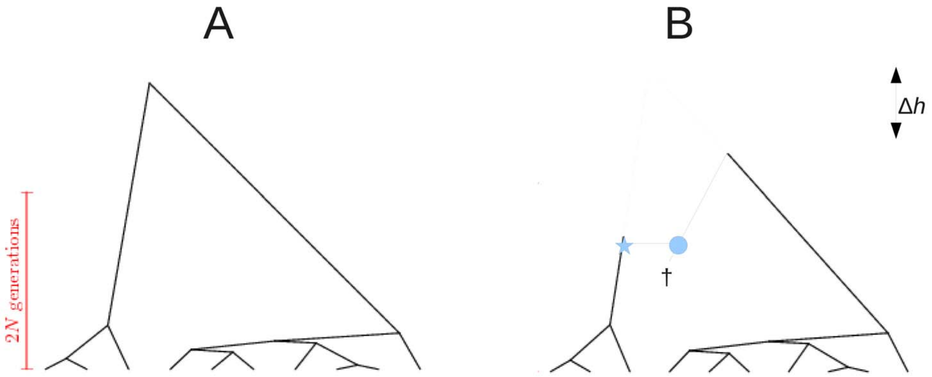

Figure 1. Example coalescent trees. A: Tree of size $n=10$ generated under the coalescent process. The y-axis represents a time scale, with leafs at the 'present', and the root in the 'past'. Starting from the present and going backwards in time, coalescent events are exponentially distributed with a parameter depending on population size ($2N$) and the number of lineages at any given point in time. B: Recombination is a prune (asterisk) and re-graft (circle) event: a lineage splits and merges onto another lineage which exists in the population at the time of recombination. This lineage does not need to extend to the present, and it may have become extinct from the entire population (cross). Recombination has changed the height of the coalescent tree with respect to the tree in panel A (Δh), but has not changed root imbalance: for both trees $\Omega=3$.
doi:10.1371/journal.pone.0060123.g001

Our goals are to formalize these concepts, to characterize in more detail the effect of single recombination events upon tree shape and to quantify the relative frequencies of drastic, mild and silent events. We explicitly calculate the probabilities of changes in height or root balance induced by a single recombination event. Our results are based on the assumption of a standard neutral model of constant population size. This means that for each coalescent event two lineages are chosen at random to merge. Further, the timing of events is exponentially distributed with a rate which, after re-scaling by population size N, depends only on the number of lineages at a given time.

In Results Section (a), we define a probability density for the trees in the spatial coalescent and we explain the difference between pointwise marginal trees T_x, evaluated at every basepair x of the DNA sequences, and the marginal trees T_f, evaluated at every fragment f. We derive a simple relation between the densities of T_x and T_f. In Section (b) we analyze the recombination events which lead to height-changes and derive their probabilities. In Section (c) we quantify the concept of root imbalance, called Ω, and derive the first-order transition probabilities under single recombination events. We focus on events which produce unbalanced trees and, at the same time, lead to an increase of tree height. This type of events is of particular interest for the analysis of biological data. Their effect on the mutation frequency spectrum and on haplotype structure is the basis of tests to reject the neutral evolution hypothesis (e.g., [16–18]). Therefore, for bench-marking it is highly interesting to know how often such events occur under purely neutral conditions, but it is not the goal of this paper to devise another neutrality test. Then, we generalize the results regarding the tree topology parameter Ω and derive the transition probability for arbitrary types of recombination events. Using this, we calculate the run-length distribution of Ω along recombining chromosomes. Finally, in Section 0.4, we calculate the average proportion of hidden recombination events and derive its limiting behavior for large sample sizes.

We remind the reader that the spatial coalescent is a non-Markovian process and not completely determined by transitions of any finite order. However, it is a homogeneous process. Therefore, first-order transition probabilities are well-defined and independent of the position in the sequence. Here, we compute first order probabilities for single recombination events from one tree to the next, averaging over all trees of the ARG which are not directly involved in the recombination event considered. Therefore, our results hold for the spatial coalescent as described by the ARG [13]. In fact, the ARG is the model which is underlying all our calculations.

Results

(a) Tree Distribution and Recombination

We consider a sample of n "chromosomes" from a diploid panmictic population of constant size N. Without recombination, the genealogical history for these chromosomes is described by the classical coalescent process [1,2]. The set of all possible coalescent trees of size n is a product $\mathbb{R}_+^{n-1}\otimes\mathcal{L}_n$, where $\mathbb{R}_+^{n-1}$ contains positive real waiting times of $n-1$ independent coalescent events and the discrete set $\mathcal{L}_n$ represents the set of all possible tree topologies. For our purposes here it is more convenient to consider labelled coalescent trees: this means that not only the internal nodes are ordered but also the leafs carry leaf labels. Hence [19] (see also http://oeis.org/A006472), the cardinality of $\mathcal{L}_n$ is

$$|\mathcal{L}_n|=\frac{n!(n-1)!}{2^{n-1}}. \tag{1}$$

Furthermore, all trees in $\mathcal{L}_n$ have the same probability $\frac{2^{n-1}}{n!(n-1)!}$, when they are generated under the standard coalescent process [20]. The waiting times t_k for a coalescent event, given k lineages, are exponentially distributed with mean $1/k(k-1)$. Time runs backward from the leafs to the root of the tree and is measured in units of the coalescent, i.e. time is scaled by four times the population size. Therefore, $\mathbb{R}_+^{n-1}\otimes\mathcal{L}_n$ can be regarded as being equipped with a probability mass function which factorizes into a probability density $p_k(t_k)$ for each waiting time ($2\leq k\leq n$) and the discrete probability for the topology $P^{(\mathrm{top})}$. For trees T in the above sense, we denote the resulting probability 'density' by

$$p^{(c)}(T)=\otimes_{k=2}^{n}p_k(t_k)\times P^{(\mathrm{top})}(T)$$

and we have

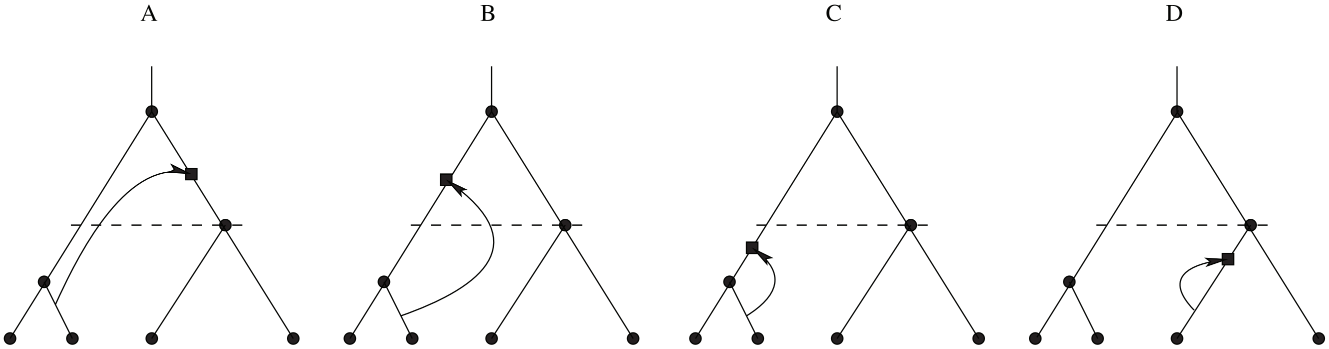

Figure 2. Non-silent, silent and hidden recombination events. A: Non-silent recombination changes tree topology. In the case shown, also Ω changes from 2 to 1. B: A recombination event which changes the order of internal nodes. Whether this event is classified as non-silent or silent, depends on the tree definition. It is non-silent for labelled histories (considered here; eq (1)), but it would be silent for unlabelled trees. C: A silent recombination event, which does not affect the branching pattern, but the lengths of the recombining branches. D: A hidden recombination event. It does neither affect branching pattern nor branch lengths.
doi:10.1371/journal.pone.0060123.g002

$$p^{(c)}(T) = \frac{2^{n-1}}{n!(n-1)!} \prod_{k=2}^{n} k(k-1)e^{-k(k-1)t_k(T)}, \qquad (2)$$

where $t_k(T)$ is the time interval during which the coalescent tree T has k lineages.

Modeling recombination as an ARG [13], there are two processes to be considered: coalescence and recombination. Given k independent lineages, in the coalescent process two lineages merge into a single one with rate $k(k-1)$. In the recombination process, a single lineage splits into two with rate $k\rho L$, where $\rho = 4Nr$ denotes the population recombination rate, r is the recombination rate per base and L is the finite length of the sequence. After a recombinational split the two ancestral lineages correspond to different sequence fragments, left and right of the point of recombination. This point is chosen uniformly along the sequence of length L. We assume that ρ is small, so multiple recombination events in the same position are negligible.

Given a tree $T(x)$ in position x, the length before the first recombination event downstream (or upstream) of x is geometrically distributed with parameter $\rho l(T)$, where $l(T)$ represents the total length of the tree. Since ρ is small, it can be safely approximated by an exponential distribution with the same parameter $\rho l(T)$.

Recombination events may change the shape of the tree. The local tree at position x in the genome may differ from the local tree at position y due to recombination. Moving along the genome, we consider two different sequences of trees: the sequence $\mathcal{S}_x = \{T(x_1), T(x_2), \ldots\}$ of local trees for all positions $x_1, x_2, \ldots$, and the sequence $\mathcal{S}_f = \{T(f_1), T(f_2), \ldots\}$ of local trees which are separated by a *single* recombination event (Figure 3). Note that a tree in $\mathcal{S}_f$ can span several base positions, as the typical length $1/\rho l(T_f)$ of the fragment f is greater than 1. Also, note that consecutive trees in $\mathcal{S}_f$ need not be different. This occurs when fragments are separated by hidden recombination events.

The standard coalescent without recombination is recovered when looking at the tree for a single position x in the sequence, ignoring all other trees. Neither the rate of coalescent events nor the choice of coalescing lineages in this tree are influenced by ancestral lineages at other positions. The local tree $T(x)$ at any position x is therefore a standard coalescent tree without recombination [21] and the marginal density of a tree in position x of the ARG is identical to $p^{(c)}(T)$; i.e., picking the tree in position x from a random sequence $\mathcal{S}_x$ is equivalent to generating one from the standard coalescent process without recombination.

On the other hand, picking a tree from a random sequence $\mathcal{S}_f$ results in a different distribution. The reason is that short trees recombine less, therefore they tend to span larger regions and to be under-represented in $\mathcal{S}_f$ compared to $\mathcal{S}_x$, as illustrated in Figures 4 and 5.

In fact, the two distributions differ by weights which are proportional to the length L_f of the fragments spanned by each tree. Since in the limit of large sequences the average length is $\mathrm{E}(L_f(T)) = 1/(\rho l(T))$, we have $p^{(c)}(T) \propto p^{(r)}(T)/l(T)$. Therefore, for large sequences, the tree density after a random recombination event is given by

$$p^{(r)}(T) = \frac{l(T)}{\mathrm{E}_c(l)} p^{(c)}(T), \qquad (3)$$

where $l(T)$ denotes the total length of the tree. For the standard neutral model, $\mathrm{E}_c(l) = a_n = \sum_{i=1}^{n-1} 1/i$. Note that the two distributions differ only in their weights of branch lengths, but not with respect to topology.

The argument leading to eq (3) can be made rigorous under the assumption of infinitely long chromosomes, using the fact that the coalescent with recombination is an ergodic process [22] (see Text S1, Supporting Information eqs (1)–(3)). As a check of eq (3), we show that $p^{(r)}(T)$ is invariant under a single recombination event. Let $\Pi_x(T'|T)$ be the transition density from tree T in a given position x to tree T' in position $x+1$, and $\Pi_r(T'|T)$ the transition density from tree T to tree T' obtained by a single recombination event. Since the marginal density $p^{(c)}(T)$ is the same for every position, we have

$$p^{(c)}(T') = \sum_T \Pi_x(T'|T) p^{(c)}(T) \qquad (4)$$

independent of the recombination rate. For small recombination rates and at first order in ρ, we have $\Pi_x(T'|T) = (1-\rho l(T))\delta_{T',T} + \rho l(T)\Pi_r(T'|T)$. Substituting this into (4) gives

$$l(T')p^{(c)}(T') = \sum_T \Pi_r(T'|T) l(T) p^{(c)}(T). \qquad (5)$$

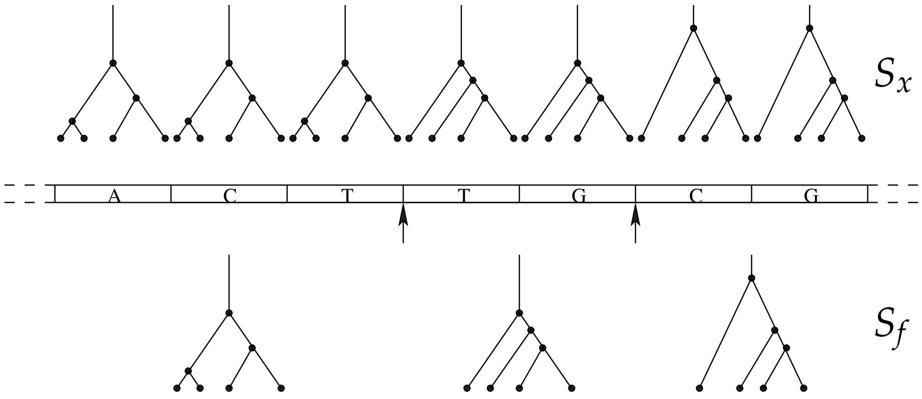

Figure 3. Distinction between sequences $\mathcal{S}_x$ and $\mathcal{S}_f$ along a recombining chromosome (sketched in the middle). Sequence $\mathcal{S}_x$ is the sequence of coalescent trees plotted for each nucleotide. Sequence $\mathcal{S}_f$ is the sequence of coalescent trees for each recombination fragment. Recombination breakpoints are indicated by arrows.
doi:10.1371/journal.pone.0060123.g003

That is, after normalization $p^{(r)}(T)\propto l(T)p^{(c)}(T)$ is an invariant distribution under $\Pi_r(T'|T)$. The normalization is $\sum_T l(T)p^{(c)}(T)=\mathrm{E}_c(l)$.

Furthermore, any marginal tree obtained from an ARG (conditioned on the number of recombinations in the sequence) by choosing randomly an ancestral lineage for every recombination event is distributed according to $p^{(r)}(T)$. This can be seen from symmetry: none of two trees separated by a single recombination event is distinguished, so they have the same distribution, which is the invariant distribution under a single recombination event, i.e. $p^{(r)}(T)$. This property has far-reaching consequences since it makes it possible to exploit the symmetries of the ARG.

Note that the two distributions, $p^{(r)}(T)$ and $p^{(c)}(T)$, become asymptotically identical when n becomes large. To see this, it suffices to consider the random variable $l/\mathrm{E}(l)$. Its mean is identical to 1. Since $\mathrm{Var}(l)=\sum_{i=1}^{n-1} i^{-2}\approx\pi^2/6$ for large n [2], one has

$$\mathrm{Var}(l/E(l))=\frac{\mathrm{Var}(l)}{\mathrm{E}^2(l)}\approx\frac{\pi^2/6}{a_n^2}\,. \qquad (6)$$

The right hand side of equation (6) converges to 0 with increasing n. Therefore the factor $l/\mathrm{E}(l)$ converges to 1 and $p^{(r)}(T)=(l/\mathrm{E}(l))p^{(c)}(T)\to p^{(c)}(T)$ (in the sense of local weak convergence). The relations between the empirical probability distributions $p[(T(x))_x]$ and $p[(T_f)_f]$ along the sequence and the probability densities $p^{(c)}(T)$ and $p^{(r)}(T)$ are summarized in the following diagram:

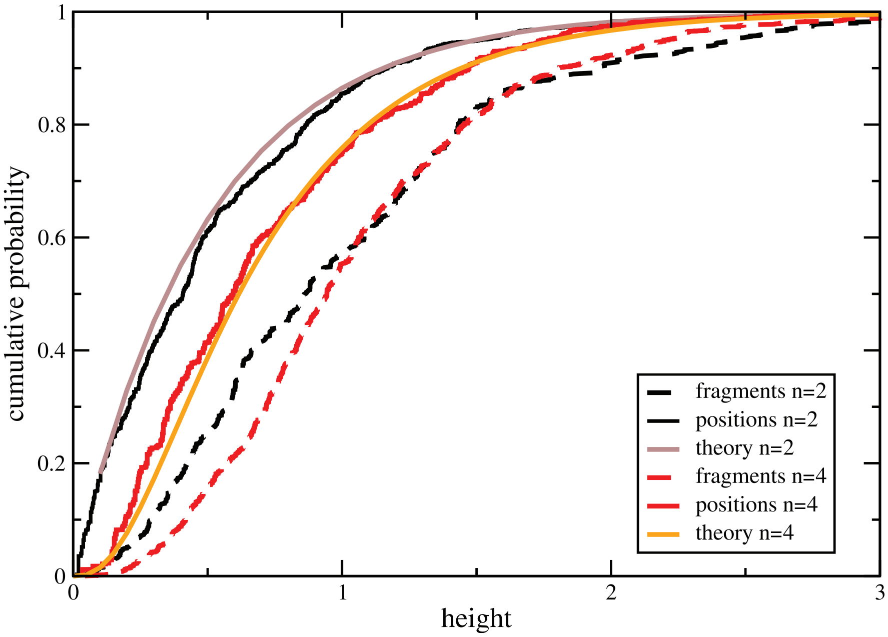

Figure 4. Cumulative distribution of tree height for $n=2$ (black) and $n=4$ (red) along a recombining chromosome of length 10^6 bp. Shown are the height distribution of trees in $\mathcal{S}_x$ (solid; "positions") and in $\mathcal{S}_f$ (dashed; "fragments"). For comparison, the theoretical distributions for $\mathcal{S}_x$ are plotted in light colors.
doi:10.1371/journal.pone.0060123.g004

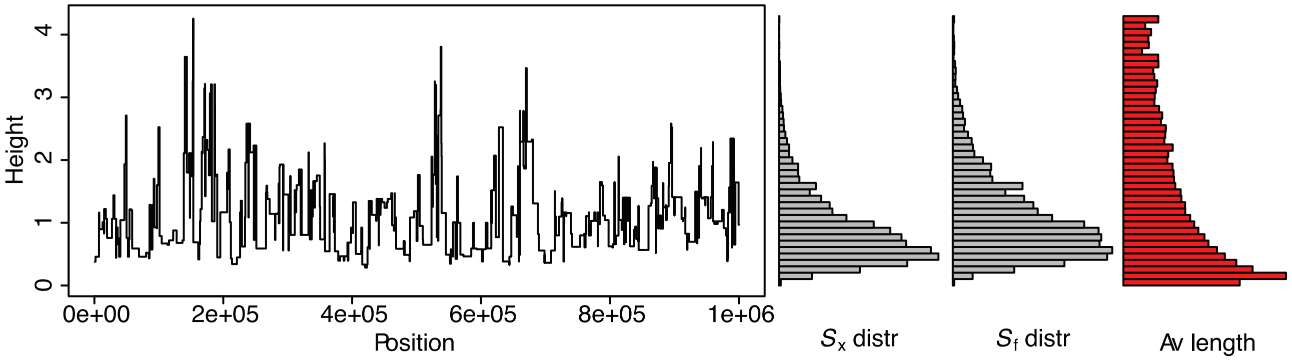

Figure 5. Height of neutral coalescent trees along the genome. One simulation run using *ms* [5] with $n=20$ and $\rho=4Nr=10^{-3}$. On the right, the distribution of the trees according to $\mathcal{S}_x$ and $\mathcal{S}_r$ and the average length before a recombination event, for a simulation of a sequence of length 10^6.
doi:10.1371/journal.pone.0060123.g005

$$\begin{tikzcd}
p\left[(T_x)_x\right] & p^{(c)}(T) \\
p\left[\left(T_f\right)_f\right] & p^{(r)}(T) \arrow[u, "n\to\infty"']
\end{tikzcd}$$

The distributions $p^{(c)}(T)$ and $p^{(r)}(T)$ need to be carefully distinguished when measuring the effect of a single recombination event. If one asks for the first recombination event downstream of a given position x in the genome, then the initial tree at position x is distributed with $p^{(c)}(T)$. If one asks instead for the effect of a randomly chosen recombination event, then the density $p^{(r)}(T)$ is the appropriate one.

(b) Height-changing Recombination Events

Probabilities of height changing events. Recombination can be interpreted as a random prune-and-regraft event on the tree [23]. First, a time point of pruning is selected uniformly anywhere on the tree; second, the node immediately above the selected branch is removed; third, the pruned branch is re-grafted onto the tree anywhere above the pruning point or onto the ancestral lineage of the root, forming a new node. For hidden recombination events, prune and re-graft occur on the same branch, without modifying topology or branch lengths of the tree.

We denote the root node by v_0 and the first internal node by v_1. There are four types of recombination events that change the height of the tree (Figure 6).

U ('up'): a prune-and-regraft event on the root branches generates a higher root without changing the topology;

D ('down'): a prune-and-regraft event on the root branches generates a lower root without changing the topology;

N ('new'): pruning a branch below the root branches and re-grafting onto the ancestral branch of the root creates a new root, while the old root becomes internal node v_1;

S ('substitute'): pruning a root branch and re-grafting onto a branch in the subtree of v_1 causes v_1 to become the root.

In fact, for the root to change height it must either be shifted (cases U and D) or be replaced (cases N and S). If the root is replaced, it can become an internal node v_1 (case N) or be lost (case S). Cases U and D leave the topology unchanged, while cases N and S do not.

We denote the probabilities of these events by $P_{\rm U}$, $P_{\rm D}$, $P_{\rm S}$, $P_{\rm N}$. We compute these quantities under both distributions, $p^{(c)}(T)$ and $p^{(r)}(T)$.

Given a coalescent tree of size n, let the *level* k be the time interval when exactly k independent lineages coexist, with $k=2,\ldots,n$. The waiting time at the kth level is $t_k(T)$, in the following called t_k for short. Tree height may be increased by recombination events of type U or N. The total probability for this, $P_{\rm UN}(T)$, is given by the sum of the probabilities of pruning at all possible levels, but never re-grafting lower than the root:

$$P_{\rm UN}(T)=\sum_{k=2}^{n}\int_0^{t_k}\frac{k\,d\tau}{l(T)}e^{-2k\tau}\prod_{j=2}^{k-1}e^{-2jt_j}, \tag{7}$$

where the product is defined to be 1 when $k=2$. This is a telescopic series that can be re-summed in a function of the total length of the tree

$$\begin{aligned}P_{\rm UN}(T)&=\sum_{k=2}^{n}\frac{k}{l(T)}\frac{1-e^{-2kt_k}}{2k}\prod_{j=2}^{k-1}e^{-2jt_j}\\&=\frac{1}{2l(T)}\sum_{k=2}^{n}\left[\prod_{j=2}^{k-1}e^{-2jt_j}-\prod_{j=2}^{k}e^{-2jt_j}\right]\end{aligned}$$

yielding the simple result

$$P_{\rm UN}(T)=\frac{1-e^{-2l(T)}}{2l(T)} \tag{8}$$

Interestingly, this probability depends only on the total length $l(T)$ of the tree and not on the topology. Very short trees grow with high probability, very long trees are unlikely to grow (Figure S1). The average probability of height-increase when passing from one recombination-delimited sequence fragment to the next is

$$\begin{aligned}P_{\rm UN}^{(r)}&=\sum_T P_{\rm UN}(T)p^{(r)}(T)=\sum_T\frac{1-e^{-2l(T)}}{2a_n}p^{(c)}(T)\\&=\frac{1}{2a_n}\left(1-\prod_{k=2}^{n}\int_0^{\infty}dt_k\;e^{-2kt_k}p_k(t_k)\right)=\\&=\frac{1}{2a_n}\left(1-\prod_{k=2}^{n}\frac{k-1}{k+1}\right)=\frac{1}{2a_n}\left(1-\frac{2}{n(n+1)}\right),\end{aligned} \tag{9}$$

which agrees very well with simulations (Figure 7). Note that $P_{\rm UN}^{(r)}$ approaches zero as slowly as $O(1/\log(n))$.

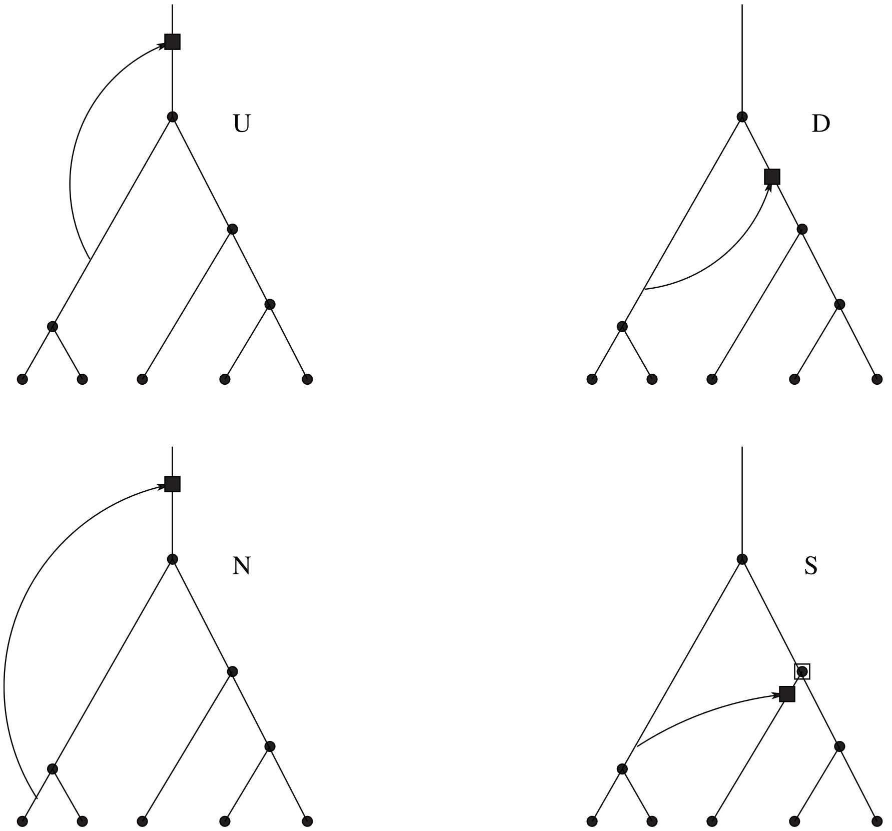

Figure 6. Types of height-changing recombination events. The square indicates the new node created by re-grafting. It forms the new root in cases U, D and N. In case S, an existing internal node becomes the new root (empty square overlaid on node v_1).
doi:10.1371/journal.pone.0060123.g006

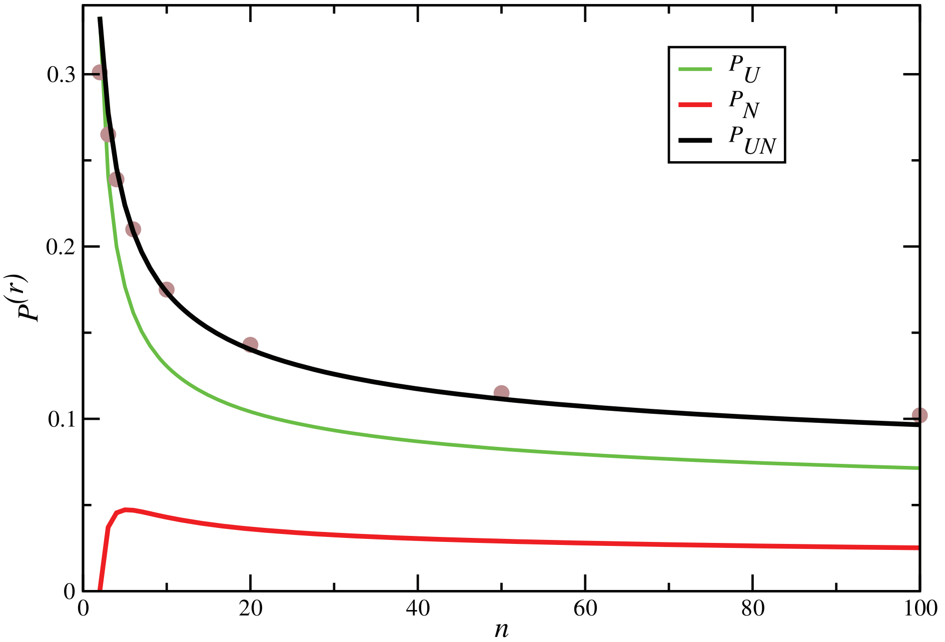

Figure 7. Increase of tree height. Probabilities $P_{\mathrm{UN}}^{(r)}$ (black), $P_{\mathrm{U}}^{(r)}$ (green) and $P_{\mathrm{N}}^{(r)}$ (red) of events that increase tree height as a function of sample size n. Dots represent the values of $P_{\mathrm{UN}}^{(r)}$ obtained by simulations using program ms [5] and selecting a random recombination event which is far from the sequence boundaries.
doi:10.1371/journal.pone.0060123.g007

This result can also be derived directly by counting ARGs, since $p^{(r)}(T)$ corresponds to the distribution of a random tree in an ARG. We will consider the case of a recombination event at a given level k and then average over all levels. To obtain the total number of ARGs $\mathcal{A}_{n,k}$ with a single recombination event at level k, choose a tree at random (among $|\mathcal{L}_n|$ possibilities), then choose the branch to be pruned (k possibilities) and the branch to which it is re-grafted at the same or a higher level ($\sum_{j=1}^{k} j$ possibilities). Therefore,

$$\mathcal{A}_{n,k}=\frac{k^2(k+1)}{2}|\mathcal{L}_n| \tag{10}$$

The number of ARGs where the new tree is higher than the old one is $k|\mathcal{L}_n|$, because there is just one possibility of re-grafting, namely on the ancestral lineage above the root of the old tree. The probability of pruning at level k in the old tree is $P_k=kt_k/l$. Therefore, one can average over $p^{(r)}(T)$ to obtain $P_{\mathrm{UN}}^{(r)}=\sum_{k=2}^{n} k\mathrm{E}_r(t_k/l)k|\mathcal{L}_n|/\mathcal{A}_{n,k}$, which is identical to equation (9).

Focusing now on pruning of the root branches, we obtain P_{U} analogously to equation (7). Let $N_k(v_j)$ be the number of direct descendants of node v_j at level k. $N_k(v_j)$ can take values 0,1,2. The average value of $N_k(v_j)$ satisfies the recursion

$$\bar{N}_{k+1}(v_j)=\bar{N}_k(v_j)\left(1-\frac{1}{k}\right)$$

$$\bar{N}_{j+2}(v_j)=2$$

that has the solution

$$\bar{N}_k(v_j)=\frac{2(j+1)}{k-1}.$$

In particular, the average number of direct descendants of the root at level k is $\bar{N}_k(v_0)=2/(k-1)$. The probability P_{U} is a modification of equation (7): multiplying by the fraction of events that are actually of type U, i.e. $N_k(v_0)/k$, one obtains

$$\begin{aligned}P_{\mathrm{U}}(T)&=\sum_{k=2}^{n}\int_0^{t_k}\frac{d\tau}{l(T)}N_k(v_0)e^{-2k\tau}\prod_{j=2}^{k-1}e^{-2jt_j}\\&=\frac{1}{l}\sum_{k=2}^{n}N_k(v_0)\frac{1-e^{-2kt_k}}{2k}\prod_{j=2}^{k-1}e^{-2jt_j}.\end{aligned} \tag{11}$$

In contrast to equation (7), equation (11) cannot be easily simplified since it depends also on the topology. After averaging over $p^{(r)}(T)$, we obtain

$$P_{\mathrm{U}}^{(r)}=\frac{1}{2a_n}\left(12b_n+\frac{10}{n}+\frac{2}{n+1}+\frac{8}{n^2}-19\right) \tag{12}$$

and

$$P_{\mathrm{N}}^{(r)}=\frac{1}{a_n}\left(10-6b_n-\frac{6}{n}-\frac{4}{n^2}\right), \tag{13}$$

where $b_n=\sum_{j=1}^{n-1}1/j^2$.

The probabilities $P_{\mathrm{D}}^{(r)}$ and $P_{\mathrm{S}}^{(r)}$ can be computed similarly to the above formulae, giving

$$\begin{aligned}P_{\mathrm{D}}(T)=&\frac{t_2}{l}-\frac{1-e^{-4t_2}}{4l}\\&+\frac{1}{l}\sum_{k=3}^{n}N_k(v_0)\frac{1-e^{-2kt_k}}{2k}\prod_{j=3}^{k-1}e^{-2jt_j}\frac{1-e^{-4t_2}}{2}\end{aligned} \tag{14}$$

and

$$P_{\mathrm{S}}(T)=$$

$$\frac{1}{l}\sum_{k=3}^{n}N_k(v_0)\left[\begin{array}{l}\sum_{j=3}^{k-1}\frac{1-e^{-2kt_k}}{2k}\prod_{d=j+1}^{k-1}e^{-2dt_d}\frac{j-1}{j}(1-e^{-2jt_j})\\+\frac{k-1}{k}\left(t_k-\frac{1-e^{-2kt_k}}{2k}\right)\end{array}\right] \tag{15}$$

(Text S1, Supporting Information eqs (4)–(9)). Alternatively, one may employ an argument based on symmetry properties of the ARG. Among two adjacent trees in the ARG, the left one is smaller or larger than the right one with equal probability. Therefore,

$$P_{\mathrm{DS}}^{(r)}=P_{\mathrm{UN}}^{(r)}. \tag{16}$$

The same is true when the root is only shifted. Thus,

$$P_{\mathrm{D}}^{(r)}=P_{\mathrm{U}}^{(r)}. \tag{17}$$

Hence, by subtraction,

$$P_{\mathrm{S}}^{(r)}=P_{\mathrm{N}}^{(r)}. \tag{18}$$

Note that the identities (17) and (18), being topological in nature, are also valid for models with variable population size. A related result about the probability that a random recombination event leaves tree height unchanged $(1-P_{\mathrm{UN}}^{(r)}-P_{\mathrm{DS}}^{(r)})$ has been obtained previously by Griffiths & Marjoram [24].

Equations (8), (11), (14), (15) are valid also when averaging over the distribution $p^{(c)}(T)$, instead of $p^{(r)}(T)$. However, exact results are available only for small sample sizes. For the case of arbitrary n we use the following Taylor approximation of the ratio moment

$$\mathrm{E}\left(\frac{X}{l}\right)\simeq\frac{\mathrm{E}(X)}{\mathrm{E}(l)}\left(1+\frac{\mathrm{Var}(l)}{\mathrm{E}(l)^2}+\frac{\mathrm{Cov}(X,l)}{\mathrm{E}(X)\mathrm{E}(l)}\right), \tag{19}$$

where $\mathrm{E}(X)/\mathrm{E}(l)$ represents the desired probability $P^{(c)}$. When the expansion is truncated at zeroth order (i.e., replacing the first moment of the ratio by the ratio of first moments), one obtains the results analogous to equations (12), (13), (17) and (18). More detailed calculations are given in Text S1, Supporting Information eqs (10)–(12). These yield, for instance, the probability of increasing tree height

$$P_{\mathrm{UN}}^{(c)}\simeq P_{\mathrm{UN}}^{(r)}\left(1+\frac{b_n}{a_n^2}+\frac{1}{a_n}\frac{3/2-1/n-1/(n+1)}{n(n+1)/2-1}\right). \tag{20}$$

Note that the scaling factor on the right hand side in equation (20) approaches 1 very slowly with increasing n. The case $P^{(c)}_{\mathrm{UN}}$ is actually an exception since an exact formula exists [15] for all values of ; in fact, $P_{\mathrm{UN}}(T)$ depends only on $l(T)$, therefore it is sufficient to average this quantity over the distribution of l obtained in [15]. For small samples there is a considerable difference between $P^{(c)}_{\mathrm{UN}}$ and $P^{(r)}_{\mathrm{UN}}$. For example, if $n=2$, we have $P^{(c)}_{\mathrm{UN}}=0.55$ while only $P^{(r)}_{\mathrm{UN}}=0.33$.

Amount of change in height. The variation in height Δh has a simple distribution. If the height increases, then the difference is given by the waiting time for coalescence of two lineages. It is

$$P_{\mathrm{U}}(\Delta h|T)=2e^{-2\Delta h}\eta(\Delta h)P_{\mathrm{U}}(T) \tag{21}$$

and

$$P_{\mathrm{N}}(\Delta h|T)=2e^{-2\Delta h}\eta(\Delta h)P_{\mathrm{N}}(T). \tag{22}$$

where $\eta(x)$ is the Heaviside function, $\eta(x)=1$ if $x\geq 0$ and 0 otherwise. If the height decreases because of an event of type D, its distribution is given by the waiting time for coalescence before time t_2, equivalent to the "bounded coalescent" for two lineages [25]

$$P_{\mathrm{D}}(\Delta h|T)=\frac{2e^{-2(t_2+\Delta h)}\eta(-\Delta h)\eta(t_2+\Delta h)}{1-e^{-2t_2}}P_{\mathrm{D}}(T). \tag{23}$$

For events of type S, the variation in height is simply the waiting time t_2 of the tree

$$P_{\mathrm{S}}(\Delta h|T)=\delta(\Delta h+t_2)P_{\mathrm{S}}(T), \tag{24}$$

where $\delta(x)$ is the Dirac delta distribution. Averaging these quantities over $p^{(r)}(T)$ and using the symmetries of the ARG, we obtain

$$P^{(r)}_{\mathrm{U}}(\Delta h)=P^{(r)}_{\mathrm{D}}(-\Delta h)=2e^{-2\Delta h}\eta(\Delta h)P^{(r)}_{\mathrm{U}} \tag{25}$$

and

$$P^{(r)}_{\mathrm{N}}(\Delta h)=P^{(r)}_{\mathrm{S}}(-\Delta h)=2e^{-2\Delta h}\eta(\Delta h)P^{(r)}_{\mathrm{N}}. \tag{26}$$

i.e., all these variations in height are exponentially distributed for an average tree.

Taking expectations, the average change in height after one of these events is

$$|\mathrm{E}(\Delta h)|=1/2,$$

irrespective of the type of event, i.e $\mathrm{E}(\Delta h|U)=\mathrm{E}(\Delta h|N)=-\mathrm{E}(\Delta h|D)=-\mathrm{E}(\Delta h|S)=1/2$. Comparing this to the average height of a tree, $\mathrm{E}(h)=1-1/n$, one notices that a single recombination event changes tree height by 50% on average.

(c) Root Imbalance and Recombination

Let L_{v_0} (R_{v_0}) be the number of left (right) descendants of the root. We have $L_{v_0}+R_{v_0}=n$. We call the random variable $\Omega=\min(L_{v_0},R_{v_0})$ *root imbalance*. Ω is a coarse-grained measure of tree topology. A recombination event may or may not change Ω and a change of Ω is neither sufficient nor necessary for a change in tree height. Since many recombination events induce rearrangements of the lower branches (close to the leafs) of the tree, they may affect Ω without affecting tree height. Still, large changes in Ω are often associated with height-changing recombination events of type N or S and thus are associated with drastic changes of tree topology.

In this section we calculate the transition probabilities $P(\omega|\omega_0)$ for Ω under a single recombination event, averaged over the initial tree. First, we focus on events of type UN, i.e. increasing height, and then we obtain the transition probabilities for all types of events separately.

Root imbalance and height-increasing events. Let the *size* of a branch be the number of leaves below the branch. A specific tree of size n can be fully described by the probability $P_{n,k}(i|T)$ that a randomly chosen branch at level k has size i. Averaging over trees of size n, the probability that a branch of level k has size i is

$$P_{n,k}(i)=\binom{n-i-1}{k-2}\Big/\binom{n-1}{k-1} \tag{27}$$

[26]. Let $\tilde{P}^{(r)}_{\mathrm{UN}}(i)$ be the probability that the height increases and the pruned branch has size i. It is obtained, similarly to $P^{(r)}_{\mathrm{UN}}$, by multiplying each term of the sum in equation (7) by $P_{n,k}(i|T)$. Thus, given a tree T,

$$\tilde{P}_{\mathrm{UN}}(i|T)=\sum_{k=2}^{n}\int_0^{t_k}\frac{d\tau}{l(T)}P_{n,k}(i|T)e^{-2k\tau}\prod_{j=2}^{k-1}e^{-2jt_j} \tag{28}$$

and, averaging over $p^{(r)}(T)$, one obtains

$$\tilde{P}^{(r)}_{\mathrm{UN}}(i)=\frac{2}{a_n}\sum_{k=2}^{n}\frac{\binom{n-i-1}{k-2}}{\binom{n-1}{k-1}}\frac{1}{k(k-1)(k+1)}. \tag{29}$$

More generally, the probability that the pruned branch has size i, given that recombination leads to an increase in height, is simply $\tilde{P}^{(r)}(i|UN)=\tilde{P}^{(r)}_{\mathrm{UN}}(i)/P^{(r)}_{\mathrm{UN}}$. The random variable Ω can take values between 1 and $n/2$ and is the folded version of the random variable i which ranges from 1 to $n-1$. Hence, the distribution of Ω, after an event that increases tree height, is

$$P^{(r)}_{\mathrm{UN}}(\omega)=\frac{\tilde{P}^{(r)}_{\mathrm{UN}}(\omega)+\tilde{P}^{(r)}_{\mathrm{UN}}(n-\omega)}{(1+\delta_{2\omega,n})}$$

and the distribution of Ω, conditioned on tree height increase, is

$$P^{(r)}(\omega|\mathrm{UN})=\frac{P^{(r)}_{\mathrm{UN}}(\omega)}{P^{(r)}_{\mathrm{UN}}}, \tag{30}$$

as illustrated in Figure S3.

Now we calculate the probability conditioned on the value ω_0 of Ω before recombination, i.e. the transition probability $P^{(r)}_{\mathrm{UN}}(\omega|\omega_0)$. The basic quantity for this computation is the probability $P_{n,k}(i|\omega_0)$ that a branch at level k has size i in a tree of total size n, given that the size of the root branches are ω_0 and $n-\omega_0$. To compute this, we need information about the actual size κ at level k of the subtree of size ω_0 of the root. We denote the distribution of κ by $P(\kappa|\omega_0,k,n)$ and the distribution of i given the sizes κ and ω_0 of its root subtree at levels k and n by $P(i|\kappa,\omega_0)$. Note that i does not depend on k nor on n, but only on the size of the root subtree to which it belongs (see Figure S4). Therefore we have

$$P_{n,k}(i|\omega_0)=\sum_{\kappa=i}^{\min(\omega_0,k-1)}\left[P(i|\kappa,\omega_0)\frac{\kappa}{k}+P(i|k-\kappa,n-\omega_0)\frac{k-\kappa}{k}\right]P(\kappa|\omega_0,k,n) \tag{31}$$

The probability $P(i|\kappa,\omega_0)$ is equal to

$$P(i|\kappa,\omega_0)=P_{\omega_0,\kappa}(i)+\delta_{i,\omega_0}\delta_{\kappa,1}=\frac{\binom{\omega_0-i-1}{\kappa-2}}{\binom{\omega_0-1}{\kappa-1}}+\delta_{i,\omega_0}\delta_{\kappa,1} \tag{32}$$

as can be shown by considering the corresponding subtree of the root as the whole tree and using equation (27). The probability $P(\kappa|\omega_0,k,n)$ depends only on the topology, therefore it can be obtained by counting the number of labelled coalescent trees (http://arxiv.org/abs/1112.1295v2) with a root branch of size ω_0 in the whole tree that reduces to size κ at level k, denoted by $\mathcal{L}_{n,\omega_0,\kappa,k}$, and dividing by the total number of trees with a root branch of size ω_0, denoted by $\mathcal{L}_{n,\omega_0}$. Using that $|\mathcal{L}_n|=n!(n-1)!/2^{n-1}$, that the coalescent process induces a uniform distribution on $\mathcal{L}_n$ and that the distribution of ω_0 is $2/(n-1)(1+\delta_{2\omega_0,n})$ [27], we have

$$|\mathcal{L}_{n,\omega_0}|=\frac{2|\mathcal{L}_n|}{(n-1)(1+\delta_{2\omega_0,n})}=\frac{n!(n-2)!}{2^{n-2}(1+\delta_{2\omega_0,n})} \tag{33}$$

The set of all trees in $\mathcal{L}_{n,\omega_0,\kappa,k}$ can be generated in the following way: (i) choose ω_0 leafs out of n; (ii) choose an relative order of the $n-2$ coalescent events among the two subsets with ω_0 and $n-\omega_0$ leafs such that among the first $n-k$ events $\omega_0-\kappa$ events belong to the first subset and $n-\omega_0-k+\kappa$ belong to the second; (iii) choose a topology for the root subtree of size ω_0; (iv) choose a topology for the complementary subtree of the root. This process generates exactly once all trees in $\mathcal{L}_{n,\omega_0,\kappa,k}$, except for the case $\omega_0=n/2$, where each tree is generated twice. Therefore, we have

$$|\mathcal{L}_{n,\omega_0,\kappa,k}|=\frac{1}{1+\delta_{2\omega_0,n}}\binom{n}{\omega_0}\binom{n-k}{\omega_0-\kappa}\binom{k-2}{\kappa-1}|\mathcal{L}_{\omega_0}||\mathcal{L}_{n-\omega_0}|. \tag{34}$$

Taking the ratio of tree counts, we obtain an hypergeometric distribution

$$P(\kappa|\omega_0,k,n)=\frac{|\mathcal{L}_{n,\omega_0,\kappa,k}|}{|\mathcal{L}_{n,\omega_0}|}=\mathrm{Hyp}_{\omega_0-1,k-2;n-2}(\kappa-1). \tag{35}$$

Finally, inserting the results (32) and (35) into (31), we obtain

$$P_{n,k}(i|\omega_0)=\frac{\delta_{i,\omega_0}\binom{n-\omega_0-1}{k-2}+\delta_{i,n-\omega_0}\binom{\omega_0-1}{k-2}}{\binom{kn-2}{k-2}}+\frac{\binom{n-i-2}{k-3}}{\binom{kn-2}{k-2}}\cdot\left[\begin{array}{l}\left(2B_{k-3,\omega_0-i-1;n-i-2}+M_{k-3,\omega_0-i-1;n-i-2}\right)\\+\left((k-1)B_{k-3,\omega_0-1;n-i-2}-M_{k-3,\omega_0-1;n-i-2}\right)\end{array}\right], \tag{36}$$

where $B_{x,y;z}$ and $M_{x,y;z}$ are the normalization and the mean (i.e., the zeroth and first moment) of the hypergeometric distribution with parameters x, y and z, if they satisfy $0\leq x,y\leq z$, and 0 otherwise. Note that $M_{x,y;z}=\frac{xy}{z}B_{x,y;z}$.

As before, we introduce $P_{n,k}(i|\omega_0)$ in equation (7) to obtain

$$\tilde{P}^{(r)}_{\mathrm{UN}}(i|\omega_0)=\frac{2}{a_n}\sum_{k=2}^{n}P_{n,k}(i|\omega_0)\frac{1}{k(k-1)(k+1)} \tag{37}$$

and, finally, the result

$$P^{(r)}_{\mathrm{UN}}(\omega|\omega_0)=\frac{\tilde{P}^{(r)}_{\mathrm{UN}}(\omega|\omega_0)+\tilde{P}^{(r)}_{\mathrm{UN}}(n-\omega|\omega_0)}{(1+\delta_{2\omega,n})} \tag{38}$$

$$P^{(r)}(\omega|\mathrm{UN},\omega_0)=\frac{P^{(r)}_{\mathrm{UN}}(\omega|\omega_0)}{\sum_{j=1}^{n-1}\tilde{P}^{(r)}_{\mathrm{UN}}(j|\omega_0)} \tag{39}$$

Figures 8 and S5 illustrate these probabilities. With a recombination event of type N, ω tends to change to smaller values. Thus, the tree becomes more unbalanced. However, by far the highest probability is attained for $\omega=\omega_0$, irrespective of ω_0 and mainly due to events of type U. This case is omitted from the figures for clarity.

Other recombination events that change root imbalance. Now we consider all possible recombination events that change Ω. Events of type U and D do not change Ω, so they can be ignored. Apart from the events of type N that we discussed above, other relevant recombination events are of type S and of type R ('root remains'), i.e. any event which leaves the root untouched. To compute the probability of a change in Ω for these types of events, we use the fact that random trees from an ARG have the distribution $p^{(r)}(T)$ and that the probability of each labelled ARG topology is the same. Due to this, we need only count the number of ARGs with a single recombination event at level k compatible with root imbalances ω_0 and ω, and denoted by $\mathcal{A}_{n,k,\omega_0,\omega,\mathrm{S}}$ and $\mathcal{A}_{n,k,\omega_0,\omega,\mathrm{R}}$. Then, we divide by the total number $\mathcal{A}_{n,k,\omega_0}$ of ARGs with a recombination at level k and root imbalance ω_0 for the original tree. Putting everything together, we obtain

$$P^{(r)}_{\mathrm{R}}(\omega|\omega_0)=\frac{1}{a_n}\sum_{k=3}^{n}\frac{1}{k^2(k+1)}\frac{1}{\binom{n-2}{k-2}}\cdot \tag{40}$$

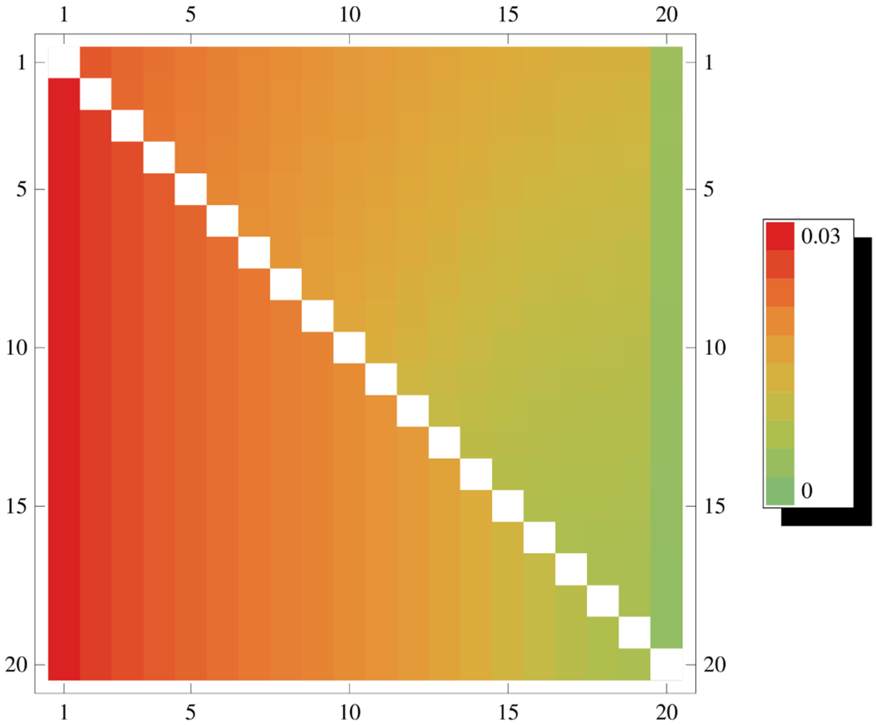

Figure 8. Transition probabilities of Ω. Distribution $P^{(r)}(\omega|\mathrm{UN},\omega_0)$ as a function of ω (horizontal axis) and ω_0 (vertical axis) for $n=40$. The diagonal terms ($\omega=\omega_0$) are not shown.
doi:10.1371/journal.pone.0060123.g008

$$\left[\binom{n-\omega_0+\omega-2}{k-3}H(\omega_0-\omega-1)(2(k-1)B_{k-3,\omega-1,n-\omega_0+\omega-2}+\right.$$

$$+(k-3)M_{k-3,\omega-1,n-\omega_0+\omega-2}-Q_{k-3,\omega-1,n-\omega_0+\omega-2})+$$

$$+\frac{1}{(1+\delta_{2\omega,n})}\binom{n-\omega+\omega_0-2}{k-3}H(\omega-\omega_0-1)(2(k-1)$$
$$B_{k-3,\omega_0-1,n-\omega+\omega_0-2}+$$

$$+(k-3)M_{k-3,\omega_0-1,n-\omega+\omega_0-2}-Q_{k-3,\omega_0-1,n-\omega+\omega_0-2})+$$

$$+\frac{1}{(1+\delta_{2\omega,n})}\binom{\omega+\omega_0-2}{k-3}H(n-\omega_0-\omega-1)(2(k-1)$$
$$B_{k-3,\omega_0-1,\omega+\omega_0-2}+$$

$$\left.+(k-3)M_{k-3,\omega_0-1,\omega+\omega_0-2}-Q_{k-3,\omega_0-1,\omega+\omega_0-2})\right],$$

where $Q_{x,y;z}$ is the second moment of the hypergeometric distribution with parameters x, y and z satisfying $0\le x,y\le z$, and 0 otherwise, and $H(n)$ is the Heaviside function, $H(n)=1$ if $n\ge 0$ and 0 otherwise. Note that the ARG symmetries imply the non-trivial relation

$$P_{\mathrm{R}}^{(r)}(\omega|\omega_0)=P_{\mathrm{R}}^{(r)}(\omega_0|\omega)\frac{1+\delta_{2\omega_0,n}}{1+\delta_{2\omega,n}}. \qquad (41)$$

The relative importance of $P_{\mathrm{R}}^{(r)}$ versus $P_{\mathrm{UN}}^{(r)}$ and $P_{\mathrm{DS}}^{(r)}$ is shown in Figure S6.

The contribution for events of type S can be obtained using the symmetry properties of the ARG. In fact, an ARG with a recombination event of type S changing ω_0 to ω is equivalent to an ARG with an event of type N changing ω to ω_0. Therefore,

$$P_{\mathrm{DS}}^{(r)}(\omega|\omega_0)=P_{\mathrm{UN}}^{(r)}(\omega_0|\omega)\frac{1+\delta_{2\omega_0,n}}{1+\delta_{2\omega,n}}. \qquad (42)$$

This result is essentially the transpose of the one shown in Figure 8, i.e. after an event of Type S, ω has an almost uniform distribution irrespective of ω_0.

Finally, the transition probability is

$$P^{(r)}(\omega|\omega_0)$$
$$=\begin{pmatrix} \omega\ne\omega_0: & P_{\mathrm{UN}}^{(r)}(\omega|\omega_0)+P_{\mathrm{DS}}^{(r)}(\omega|\omega_0)+P_{\mathrm{R}}^{(r)}(\omega|\omega_0) \\ \omega=\omega_0: & 1-\sum_{\omega\ne\omega_0}\left(P_{\mathrm{UN}}^{(r)}(\omega|\omega_0)+P_{\mathrm{DS}}^{(r)}(\omega|\omega_0)+P_{\mathrm{R}}^{(r)}(\omega|\omega_0)\right)\end{pmatrix} \qquad (43)$$

This distribution is shown in Figures S7 and S8 for $n=40$.

(d) Hidden and Silent Recombination Events

Counting ARGs we now determine the fraction of *hidden* recombination events, i.e. those which neither change tree topology nor branch lengths. Since these events are 'invisible' when analysing sequence polymorphisms or haplotype structure, their frequency can only be estimated by theoretical means.

Hidden recombination events are caused by pruning and re-grafting on the same branch (see Figure 2D). Let $\mathcal{A}_{n,k,H}$ denote the number of ARGs with a hidden event at level k. Since ARG topologies are equiprobable under $p^{(r)}(T)$, the probability that a recombination event is hidden is

$$P_H^{(r)}=\sum_{k=2}^{n} P_k \frac{\mathcal{A}_{n,k,H}}{\mathcal{A}_{n,k}}\,, \tag{44}$$

where $P_k=E(kt_k/l)=((k-1)a_n)^{-1}$ is the probability of pruning at level k. To calculate $\mathcal{A}_{n,k,H}$ we need to consider the following ingredients. A branch pruned under node v_j can be regrafted in $k-j-1$ topologically inequivalent ways on the same branch (but possibly on different levels). This number has to be multiplied by the number of branches under node v_j at level k (denoted by $N_k(v_j)$). Then, one has to sum over all possible nodes v_j and over all possible initial trees $T\in\mathcal{L}_n$. This yields

$$\begin{aligned}\mathcal{A}_{n,k,H}&=\sum_{T\in\mathcal{L}_n}\sum_{j=0}^{k-2} N_k(v_j)(k-j-1)\\ &=\sum_{j=0}^{k-2}\bar{N}_k(v_j)(k-j-1)|\mathcal{L}_n|\end{aligned} \tag{45}$$

Combining eqs (44) and (45) we obtain

$$\begin{aligned}P_H^{(r)}&=\sum_{k=2}^{n}\frac{1}{(k-1)a_n}\frac{\sum_{j=0}^{k-2}\bar{N}_k(v_j)(k-j-1)|\mathcal{L}_n|}{k^2(k+1)|\mathcal{L}_n|/2}\\ &=\frac{2}{3a_n}\left(1-\frac{1}{n}\right).\end{aligned} \tag{46}$$

This means that the fraction of hidden recombination events is of the order $O(1/\log(n))$. They are quite frequent for small to moderate n, but become increasingly rare with increasing n. Still, even when $n=1000$, about 9% of all recombination events are hidden.

Using the same technique of counting ARGs also the fraction of silent recombination events (i.e. events that do not change topology but that may change branch lengths) can be obtained. We start by counting events that are silent but not hidden. Given a tree, select a branch for pruning. Then, there are exactly two ways for re-grafting: either on the branch immediately above or on the branch immediately below the old parent node of the pruned branch (Figure 2B or C), but not on the pruned branch itself (the latter would be a hidden event). Performing similar calculations as before we obtain

$$\begin{aligned}P_{\text{silent}}^{(r)}-P_H^{(r)}&=\sum_{k=2}^{n}\frac{1}{(k-1)a_n}\frac{\mathcal{A}_{n,k,sil-H}-}{\mathcal{A}_{n,k}}\\ &=\sum_{k=2}^{n}\frac{1}{(k-1)a_n}\frac{\sum_{j=0}^{k-2}2\bar{N}_k(v_j)|\mathcal{L}_n|}{k^2(k+1)|\mathcal{L}_n|/2}\\ &=\frac{1}{a_n}\left(1-\frac{2}{n(n+1)}\right).\end{aligned} \tag{47}$$

Therefore,

$$P_{\text{silent}}^{(r)}=\frac{1}{3a_n}\left(5-\frac{8}{n}+\frac{6}{n+1}\right). \tag{48}$$

Note that the following holds:

$$P_{\text{silent}}^{(r)}=P_{UNDS}^{(r)}+P_H^{(r)}. \tag{49}$$

An intuitive explanation is the following: for any pruning point, there are two possible ways for re-grafting such that tree topology remains unchanged and there is exactly one way for re-grafting which leads to an increase of tree height. Therefore, $P_{\text{silent}}^{(r)}-P_H^{(r)}=2P_{UN}^{(r)}$. Then, eq (49) follows from symmetry of the ARG. Note that this argument is topological and does not depend on waiting times, i.e. branch lengths.

(e) Correlation Lengths

Since the spatial coalescent is a non-Markovian process, it is important to know over which chromosomal distances correlation and statistical dependence among trees persist. Correlation between trees, measured by any well-behaved tree statistic, decreases with distance. An interesting question is how quickly recombination reduces correlation. The answer depends on the particular statistic which is employed to measure correlation. Topology based statistics, such as Ω (measuring imbalance at the root) or Colless' index [28] (measuring imbalance at all internal nodes), behave differently from length based statistics, such as tree height (Figure 9).

We use our above results regarding events of type U, D, N, S and R to give a quantitative answer. The idea is to approximate the correlation length for a statistic by the inverse of the probability of recombination events that have a strong impact on this statistic.

Events of type U or D change height, but leave the topology unchanged. Events of type R preserve height but alter topology. Events of type N or S may change both, height and topology. They also lead to the fastest decay of correlation.

The average number of recombination events before an event of type N or S occurs is the inverse of this probability. This quantity is a rough estimate for the correlation length of tree shape. The numerical values of $P_{\text{NS}}^{(r)}=2P_{\text{N}}^{(r)}$ for $20\lesssim n\lesssim 100$ lie between $0.05-0.07$ (Figure S2). Based on this estimate, correlation between trees should decay strongly within 15 to 20 recombination events. This is in agreement with numerical simulations. More generally, the topological correlation length can be roughly estimated as

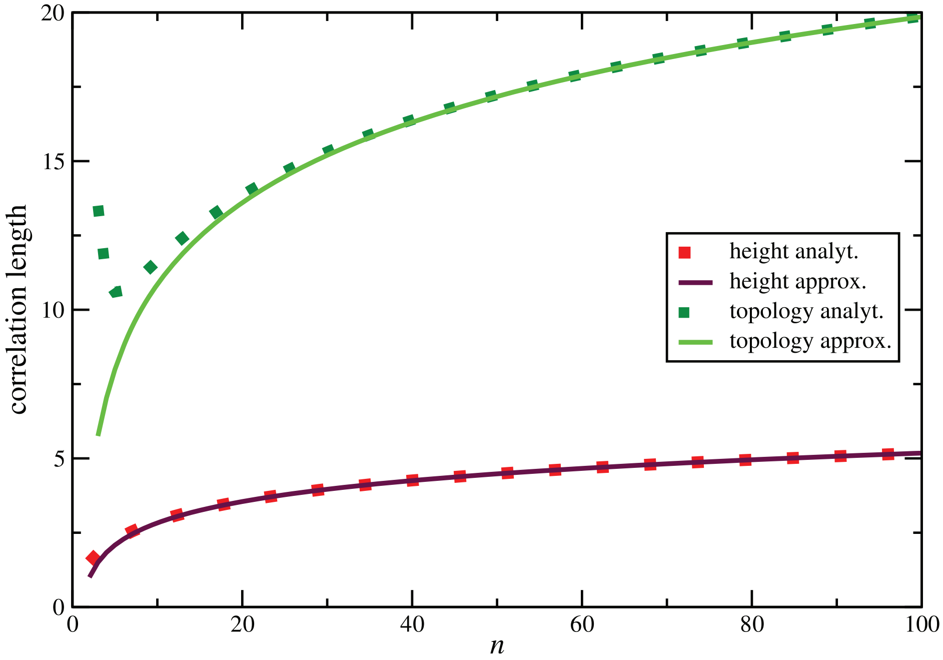

Figure 9. Correlation length $\Lambda_{top}^{(r)}$ (blue line) as a function of sample size *n*. The red line is the approximation $\log(n)/2(10-\pi^2)$.
doi:10.1371/journal.pone.0060123.g009

$$\Lambda_{top}^{(r)} \frac{1}{P_{\mathrm{NS}}^{(r)}} \frac{a_n}{2(10-\pi^2)} \approx 3.83\, a_n. \tag{50}$$

It Increases Logarithmically in n (Figure 9)

To translate this into physical length, we assume that the distance between two consecutive recombination events is exponentially distributed with mean $1/(\rho l(T))$. Averaging over $p^{(r)}(T)$ we obtain $1/(\rho a_n)$. Therefore, distance λ_{top} between two events of type N or S is approximately

$$\lambda_{top} = \frac{\Lambda_{top}^{(r)}}{\rho a_n} \sim \frac{1}{2(10-\pi^2)\rho} \sim \frac{3.83}{\rho}, \tag{51}$$

independent of n. For example, if the scaled recombination rate is $\rho \approx 10^{-3}$, the genomic distance between such events is about 4kb. Assuming that also the scaled mutation rate is $\theta \approx 10^{-3}$ per bp and assuming $n=100$, an interval between drastic recombination events of type N or S contains about $4a_{99} \approx 20$ polymorphic sites. This number should be sufficiently high to enable at least a rough tree re-construction from SNP data, and to estimate Ω. It will probably not be sufficient for the reconstruction of the fine topological structure of the lower branches.

To estimate the correlation length of Ω, also events of type R need to be taken into account. In fact, changes in Ω occur more often than events of type N or S. Using equation (43), we determined the run-length of Ω, i.e. the number of recombination events that occur before a change in Ω happens. Considering a random initial tree, an estimate for the run-length is given by

$$\Lambda_\Omega = \frac{1}{1-P^{(r)}(\omega|\omega)}. \tag{52}$$

The run-length is longer for more imbalanced trees, but always on the order of a few recombination events (between 2 and 6; Figure 10). This is also a reasonable estimate for the correlation length of the fine topological structure.

We now consider correlation in tree height. Height can change by events U,D,N and S. The average change in height is the same, $|\Delta h| = 1/2$, for all these events. Therefore, correlation length can be estimated as

$$\Lambda_h^{(r)} \sim 1/P_{\mathrm{UNDS}}^{(r)}.$$

Since
$P_{\mathrm{UNDS}}^{(r)} = 2P_{\mathrm{UN}}^{(r)}$ is between 0.25 and 0.3 for $20 \lesssim n \lesssim 100$ (Figure 7), drastic changes in height are expected on average every 3 to 4 recombination events. More generally, the correlation length also increases logarithmically in n and is

$$\Lambda_h^{(r)} \sim a_n. \tag{53}$$

For the physical correlation length we have.

$$\lambda_h = \frac{\Lambda_h^{(r)}}{\rho a_n} \sim \frac{1}{\rho}. \tag{54}$$

This is only about a quarter of the topological correlation length. Therefore, an exact reconstruction of tree height is difficult. For instance, for $n=100$ and $\theta = \rho = 10^{-3}$, one would have on average only 5 SNPs to estimate height or other tree parameters.

For the case $n=2$, Hudson [21] gives a formula for the correlation between the heights of two trees in dependence of the recombination rate ρ. The formula predicts that the correlation drops to about 0.5 with $\rho 1.4$, i.e. after approximately 1.4

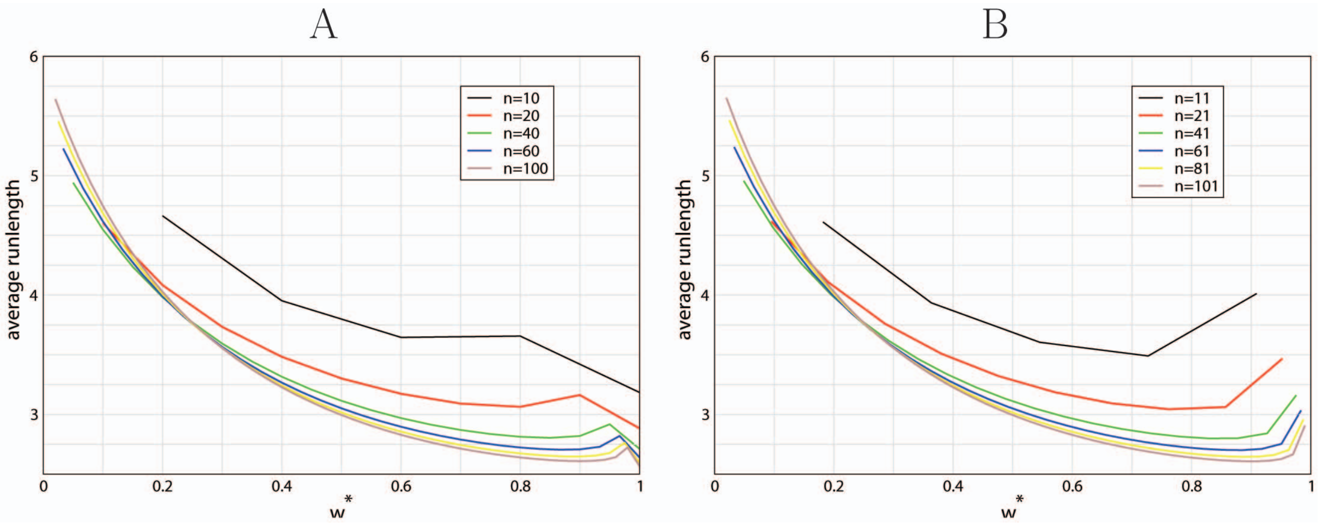

Figure 10. Run length $1/(1-P^{(r)}(\omega|\omega))$ **as a function of** $\omega^*=\frac{2\omega}{n}$ **for even sample sizes (A)** ($n=10,20,40,60,80,100$) **and for odd sample sizes (B)** ($n=11,21,41,61,81,101$)**.**
doi:10.1371/journal.pone.0060123.g010

recombination events. Our rough estimate for the correlation length in this case is $1/P^{(r)}_{\mathrm{UNDS}}=1.5$, and in good agreement with Hudson's result.

Finally, we briefly comment that linkage disequilibrium and haplotype block size depend strongly on the number and distribution of mutation and recombination events along coalescent trees, i.e. they depend strongly on tree topology and length. Since topology can in practice only be indirectly estimated from polymorphism patterns, not all changes in topology are actually visible for these statistics. The correlation lengths estimated from experimental data will tend to be larger than the theoretical estimates presented here. Assuming that haplotype blocks are mostly delimited by 'drastic' recombination events, involving a change of topology, we estimate the size of these haplotype fragments L_h, centered at some position x with a tree T. Assuming further that neither tree length $l(T)$ nor the probability of topology-changing drastic recombination events $P_{td}(T)$ change much after a 'non-drastic' recombination event, the probability distribution for the haplotype sizes is

$$P(L_h|T)=e^{-\rho l(T)P_{td}(T)L_h}\rho l(T)P_{td}(T). \qquad (55)$$

The average size is then

$$\mathrm{E}(L_h|T)=1/(\rho\, l(T)P_{td}(T)). \qquad (56)$$

The class of drastic recombination events that should be considered to determine $P_{td}(T)$ is probably larger than the class of type N and S events. However, $P_{td}(T)=P_{NS}(T)$ is a reasonable lower bound approximation.

Discussion

We have considered the effect of single recombination events on coalescent tree topology and explicitly determined the probability with which recombination triggers 'drastic' changes. We consider a change to be drastic if it leads to a change of tree height or of tree imbalance. These types of events are of practical interest because both have an effect on the pattern of polymorphic sites which are informative for genealogical reconstruction and evolutionary inferences. The primary effect of height change is upon the number of mutations, while a change in tree imbalance primarily affects the mutation site frequency spectrum.

Our results show important qualitative differences for the two types. The average change in height is quite drastic per se (50% of average tree height), while the average change in imbalance is quite mild, with large jumps occuring only very rarely. Our results hold for the standard neutral model, i.e. a model with constant population size and without substructure. As such, our results may serve as the analytical reference case for constructing formal tests of the neutral evolution hypothesis. For instance, the probabilities of height or topology change are markedly altered in the presence of selective sweeps, i.e. the fast fixation of a mutant allele due to positive selection. Recombination close to the sweep site, where tree height is severely reduced [29], tends to lead to both a drastic increase of tree height and highly imbalanced trees [16,18]. In contrast, variable population size leaves a different signature on the probabilities of drastic recombination events. Non-constancy of N is reflected in branch length variation, but it has no impact on the branching pattern, i.e. on topology. In fact, if panmixis continues to hold, the probability distribution of tree topologies does not depend on population size. Variation of N affects only branch lengths and waiting times. Since all our results, averaged over $p^{(r)}(T)$, depend implicitly on the first moments of the waiting times through the quantity $P_k=k\mathrm{E}(t_k/l)$, they can in principle be adapted to models with variable population size using the theory developed earlier [26,30]. A detailed treatment is left to further investigation. Here we just note that the relations (17), (18) and (49) are valid for all models of variable population size.

Population substructure is another important case of deviation from the standard neutral model. Restricted gene flow between subpopulations strongly affects the transition probabilities of root imbalance, but less the distribution of height change. A more detailed discussion of the impact of these evolutionary scenarios upon a test statistic of the neutral evolution hypothesis is given in [18].

We have derived a number of further results which shed more light on the details and consequences of recombination. We analysed the correlation length between trees on a recombining chromosome and showed that topological correlation is generally longer-ranging than correlation in tree height. Still, for both types very few recombination events – on the order of ten – are sufficient to unlink the genealogical histories of two genomic fragments, given standard neutral conditions. The calculations also make clear that correlation length (number of recombinations) scales logarithmically in *n*. This is important to take into account for deep sequencing association studies.

It is perhaps surprising to see that a considerable fraction of recombination events remains hidden. Even for large sample sizes, about 10% of the recombination events are not visible. An even larger fraction is silent, i.e. does not cause topological changes of the underlying genealogy.

Analyzing root imbalance in more detail, we found that the distribution of Ω-run lengths is biased towards unbalanced trees: under the standard neutral model, unbalanced trees tend to span larger genomic regions than balanced trees. Interestingly, the Ω-run length, when normalized, is asymptotically independent of *n*. Our results provide a basis to tackle problems of correlation between tree statistics in coalescent models. They extend known results, such as the one by Hudson [21] concerning tree height correlation, to the more general case of arbitrary sample size *n*.

Some of the quantities studied here involve counting problems of ancestral recombination graphs with a single recombination event. These problems are related to counting problems of phylogenetic networks [31]. Unlike counting problems of trees, which can often be tackled by generating function techniques ([20], arxiv.org/abs/1112.1295v2, arxiv.org/abs/1202.5668v3), only few results are available for tree-like structures with independent cycles so far [32]. Our results represent a step towards a combinatorial treatment of these problems.

Supporting Information

Figure S1 Probability of increasing height after a recombination event as a function of the total tree length l.
(PDF)

Figure S2 Probability of recombination events $P^{(r)}_{\mathrm{NS}}$ which change tree height and topology as a function of the sample size n.
(PDF)

Figure S3 Distribution $P^{(r)}(\omega|\mathrm{UN})$ of Ω after an event that increases tree height, for $n=20$.
(PDF)

Figure S4 Illustration of the sizes x and y of the subtrees at the levels k and j corresponding to pruning and regrafting, respectively.
(PDF)

Figure S5 Probability distribution $P^{(r)}(\omega|\mathrm{UN},\omega_0)$ for $\omega_0=1,5,10,15$ (in blue, pink, yellow, green) and $n=40$. For clarity, only the probabilities for $\omega\neq\omega_0$ are shown.
(PDF)

Figure S6 Ratio $P^{(r)}_{\mathrm{UNDS}}(\omega|\omega_0)/P^{(r)}_{\mathrm{R}}(\omega|\omega_0)$ as a function of ω (x-axis) and ω_0 (y-axis) for $n=40$. For clarity, only the probabilities for $\omega\neq\omega_0$ are shown.
(PDF)

Figure S7 Distribution $P^{(r)}(\omega|\omega_0)$ of Ω for $\omega_0=1,5,10,15$ (in blue, pink, yellow, green) and $n=40$.
(PDF)

Figure S8 Distribution $P^{(r)}(\omega|\omega_0)$ as a function of ω (x-axis) and ω_0 (y-axis) for $n=40$. For clarity, only the probabilities for $\omega\neq\omega_0$ are shown.
(PDF)

Text S1 Supporting information.
(PDF)

Acknowledgments

We would like to thank Jeff Thorne and two anonymous reviewers for very constructive comments, and A. Klassmann and S. Ramos-Onsins for numerous discussions.

Author Contributions

Analyzed the data: LF TW. Wrote the paper: LF TW. Designed the study: TW LF. Performed calculations and derived theorems: LF FD. Performed computer simulations: LF TW.

References

1. Kingman JFC (1982) The coalescent. Stochastic Processes and their Applications 13: 235–248.
2. Hudson RR (1990) Gene genealogies and the coalescent process. In: Oxford Surveys in Evolutionary Biology, Oxford University Press, volume 7. 1–44.
3. Wakeley J (2009) Coalescent theory – an introduction. Greenwood Village, Colorado: Roberts&Company.
4. Kimura M (1987) Molecular evolutionary clock and the neutral theory. J Mol Evol 26: 24–33.
5. Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337–338.
6. Kim Y, Wiehe T (2009) Simulation of DNA sequence evolution under models of recent directional selection. Brief Bioinform 10: 84–96.
7. Ewing G, Hermisson J (2010) MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics 26: 2064–2065.
8. Griffiths RC (1984) Asymptotic line-of-descent distributions. J Math Biol 21: 67–75.
9. Sagitov S (1999) The general coalescent with asynchronous mergers of ancestral lines. J Appl Probab 36: 1116–1125.
10. Greven A, Pfaffelhuber P, Winter A (2009) Convergence in distribution of random metric measure spaces (Λ-coalescent measure trees). Probab Theory Relat Fields 145: 285–322.
11. Bhaskar A, Kamm JA, Song YS (2012) Approximate sampling formulae for general finite-alleles models of mutation. Adv Appl Probab 44: 408–428.
12. Angel O, Berestycki N, Limic V (2012) Global divergence of spatial coalescents. Probab Theory Relat Fields 152: 625–679.
13. Griffiths RC, Marjoram P (1996) Ancestral inference from samples of DNA sequences with recombination. J Comput Biol 3: 479–502.
14. Li H, Durbin R (2011) Inference of human population history from individual whole-genome sequences. Nature 475: 493–496.
15. Wiuf C, Hein J (1999) Recombination as a point process along sequences. Theor Popul Biol 55: 248–259.
16. Fay J, Wu C (2000) Hitchhiking under positive Darwinian selection. Genetics 155: 1405–1413.
17. Li H (2011) A new test for detecting recent positive selection that is free from the confounding impacts of demography. Mol Biol Evol 28: 365–375.
18. Li H, Wiehe T (2012) Coalescent tree imbalance as an indicator of selective sweeps. (in review).
19. Murtagh F (1984) Counting dendrograms: A survey. Discrete Applied Mathematics 7: 191–199.
20. Disanto F, Wiehe T (2013) Exact enumeration of cherries and pitchforks in ranked trees under the coalescent model. Mathematical Biosciences 242: 195–200.
21. Hudson R (1983) Properties of a neutral allele model with intragenic recombination. Theor Popul Biol 23: 183–201.
22. Wiuf C (2006) Consistency of estimators of population scaled parameters using composite likelihood. J Math Biol 53: 821–841.
23. Paul J, Steinrücken M, Song Y (2011) An accurate sequentially Markov conditional sampling distribution for the coalescent with recombination. Genetics 187: 1115–1128.

24. Griffiths RC, Marjoram P (1997) An ancestral recombination graph. In: Progress in population genetics and human evolution (Minneapolis, MN, 1994), New York: Springer, volume 87 of IMA Vol. Math. Appl. 257–270.
25. Rasmussen M, Kellis M (2012) Unified modeling of gene duplication, loss, and coalescence using a locus tree. Genome Res 22: 755–765.
26. Zivkovic D, Wiehe T (2008) Second-order moments of segregating sites under variable population size. Genetics 180: 341–357.
27. Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. Genetics 105: 437–460.
28. Colless DH (1982) Review: [untitled]. Systematic Zoology 31: 100–104.
29. Kaplan N, Hudson R, Langley C (1989) The ''hitchhiking effect'' revisited. Genetics 123: 887–899.
30. Griffiths RC, Tavaré S (2003) The genealogy of a neutral mutation. In: Green P, Hjort N, Richardson S, editors, Highly Structured Stochastic Systems, Oxford Statistical Science Series, Oxford University Press, volume 27. 393–412.
31. Huson DH, Rupp R, Scornavacca C (2011) Phylogenetic Networks: Concepts, Algorithms and Applications. Cambridge University Press.
32. Semple C, Steel M (2006) Unicyclic networks: compatibility and enumeration. IEEE/ACM Trans Comput Biol Bioinform 3: 84–91.