OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Coalescent Tree Imbalance and a Simple Test for Selective Sweeps Based on Microsatellite Variation

Haipeng Li[1,2¤], Thomas Wiehe[2]*

1 Department of Computational Genomics, CAS Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China, **2** Institut für Genetik, Universität zu Köln, Köln, Germany

Abstract

Selective sweeps are at the core of adaptive evolution. We study how the shape of coalescent trees is affected by recent selective sweeps. To do so we define a coarse-grained measure of tree topology. This measure has appealing analytical properties, its distribution is derived from a uniform, and it is easy to estimate from experimental data. We show how it can be cast into a test for recent selective sweeps using microsatellite markers and present an application to an experimental data set from *Plasmodium falciparum*.

Citation: Li H, Wiehe T (2013) Coalescent Tree Imbalance and a Simple Test for Selective Sweeps Based on Microsatellite Variation. PLoS Comput Biol 9(5): e1003060. doi:10.1371/journal.pcbi.1003060

Editor: Wen-Hsiung Li, University of Chicago, United States of America

Received September 5, 2012; **Accepted** March 28, 2013; **Published** May 16, 2013

Funding: HL was supported by the National Key Basic Research Program of China (2012CB316505), the NSFC grants (31172073 and 91131010) and the Bairen Program, and through a grant to TW by the German Research Foundation (DFG-SFB680, www.dfg.de). 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

¤ Current address: Department of Computational Genomics, CAS Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China.

Introduction

The coalescent process is an established tool to describe the evolutionary history of a sample of genes drawn from a natural population [1–3]. For a neutrally evolving population of constant size N the coalescent has well understood analytical properties concerning tree shape and mutation frequency spectrum which provide a firm basis for a variety of statistical tests of the neutral evolution hypothesis [4–8]. Adding recombination as an evolutionary mechanism, the coalescent is usually studied in the framework of the ancestral recombination graph (ARG) [9]. The combined action of selection and recombination has been analyzed first in detail by Hudson and Kaplan [10] and, in terms of genetic hitchhiking, by Kaplan *et al.* [11]. More recently, it was shown that the (non-Markovian) ARG can well be approximated by a simpler, more tractable model, the so-called Sequential Markov Coalescent [12–14], which is of particular interest for the efficient simulation of genealogies across large genomic regions. How single recombination events reflect on tree shape under neutrality has recently been analyzed by Ferretti et al. [15]. Here, we concentrate on tree shape in the vicinity of a selected locus.

Selection changes the rate by which coalescent events occur and hence can lead to distortions of tree shape. It is well known [6,16] that selective sweeps can produce highly unbalanced trees when selection acts in concert with limited recombination, i.e. at some chromosomal distance from the site under selection. Conversely, observing unbalanced trees should provide information about recent selection in a particular genomic region. In fact, this property is also the basis of Li's MDFM test [16]. A practical concern is how such distorted gene genealogies may reliably be estimated or re-constructed using polymorphism data. When working with SNPs a large genomic fragment with many polymorphic sites has to be analyzed to obtain a clear phylogenetic signal. Since for many organisms recombination and mutation rates are on the same order of magnitude [17, Table 4.1], one harvests about as many recombination as polymorphic sites when sampling genomic sequences, thus complicating tree shape estimation. To alleviate this problem one may turn to multi-allelic markers, such as microsatellites, complementing or replacing bi-allelic SNPs.

In this paper we introduce the statistic Ω of tree balance and, first, derive theoretical properties of this and derived statistics. Second, we show how a selective sweep affects these statistics. Third, we investigate the possibility and reliability of estimating Ω from experimental data. Fourth, we define an easily applicable microsatellite based test statistic for selective sweeps. It requires clustering of microsatellite alleles into two disjoint sets and examining whether these sets are sufficiently different in size and/or whether they have a sufficiently large distance from each other. Finally, we demonstrate a practical application.

Terminology

Consider the coalescent tree for a sample of size n. It is a binary tree without left-right orientation, with ordered internal nodes and branch lengths representing a measure of time. All leaves are aligned on the bottom line, representing the present. We use the term *tree topology* when talking about the branching pattern and *tree shape* when talking about topology and branch lengths. We remark that topology and shape can be conceptually distinguished, but in practice estimating topology relies on polymorphism patterns. Since these depend on branch lengths, i.e. on shape, topology can usually not be estimated independently. We call the *size* of a tree the number of leaves and the *length* of a tree the combined length of

Author Summary

It is one of the major interests in population genetics to contrast the properties and consequences of neutral and non-neutral modes of evolution. As is well-known, positive Darwinian selection and genetic hitchhiking drastically change the profile of genetic diversity compared to neutral expectations. The present-day observable genetic diversity in a sample of DNA sequences depends on events in their evolutionary history, and in particular on the shape of the underlying genealogical tree. In this paper we study how the shape of coalescent trees is affected by the presence of positively selected mutations. We define a measure of tree topology and study its properties under scenarios of neutrality and positive selection. We show that this measure can reliably be estimated from experimental data, and define an easy-to-compute statistical test of the neutral evolution hypothesis. We apply this test to data from a population of the malaria parasite *Plasmodium falciparum* and confirm the signature of recent positive selection in the vicinity of a drug resistance locus.

all branches. The *height* is the time interval between present and root, indicated by t_0 in Figure 1. Let the label of the root be v_0. The n leaves can be grouped into two disjoint sets, L_0 and R_0, the 'left-' and 'right-descendants' of the root. Let L_0 be the smaller of the two sets and $|L_0|=\Omega_0$. Hence, $|R_0|=n-\Omega_0\geq n/2$. Let v_1 be the 'right' child of v_0, i.e. the root of the subtree with leaf set R_0. The descendants of v_1 can again be grouped into two disjoint subsets, L_1 and R_1, the left- and right-descendants of v_1. Again, without loss of generality, let $|L_1|\leq|R_1|$ and denote $|L_1|=\Omega_1$. Hence, $|R_1|=n-\Omega_0-\Omega_1$. Proceed in this way to define subsets L_2, R_2, and so on. For any tree there are h such pairs (L_i,R_i) where $\log_2(n)\leq h\leq n-1$, with h depending on the topology of the tree. The set $R_0,\ldots,R_h$ constitutes a – not necessarily unique – top-down sequence of maximal subtrees.

Results

Tree topology of the neutral coalescent

Consider a coalescent tree of size n under the neutral model with constant population size, where n is assumed to be large. Root imbalance is measured by the random variable Ω_0. The distribution of Ω_0 is 'almost'-uniform [18,19] on $\{1,2,\ldots,\lfloor n/2\rfloor\}$. More precisely,

$$p(n,\omega_0) := \mathrm{Prob}(\Omega_0=\omega_0)=\frac{2-\delta_{\omega_0,n/2}}{n-1}, \qquad (1)$$

where $\delta_{.,.}$ denotes here the Kronecker symbol. The expectation is

$$E(\Omega_0)=\sum_{\omega_0=1}^{n/2}\omega_0\,p(n,\omega_0)=\frac{n^2}{4(n-1)}\approx\frac{n}{4}.$$

The variance is

$$V(\Omega_0)=\sum_{\omega_0=1}^{n/2}\omega_0^2\,p(n,\omega_0)-E^2(\Omega_0)=\frac{(n-2)n(4+(n-2)n)}{48(n-1)^2}\approx\frac{n^2}{48}$$

and the standard deviation

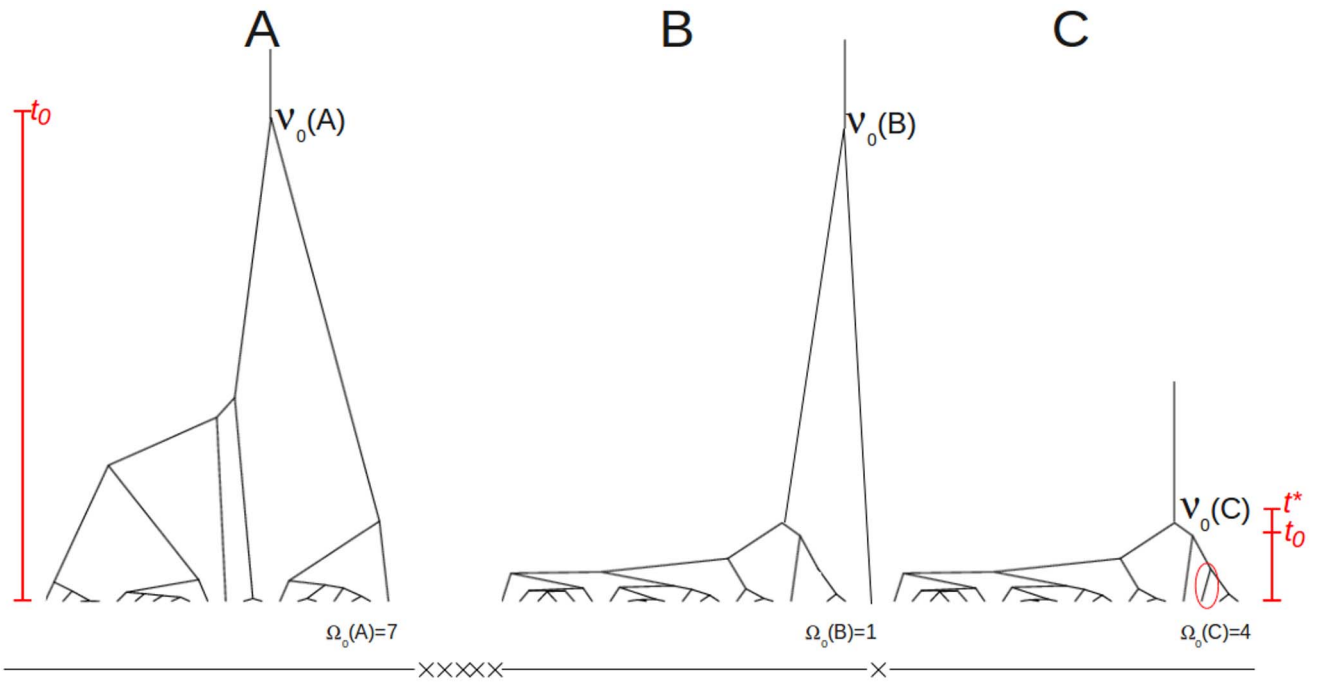

Figure 1. Coalescent trees under recombination and selection. A: Sketch of a neutral coalescent tree with tree size $n=20$. **B** and **C**: A selective sweep in locus C leads to a tree of low height (t_0 small). The selective sweep was initiated by a beneficial mutation at time t^*. At some distance from C, a single lineage (circled branch in C) has "recombined away" leading to the unbalanced tree shown at locus B. Note that tree height between trees B and C changes drastically and that $\Omega_0=4$ at locus C and $\Omega_0=1$ at locus B. Multiple recombination events (indicated by the crosses at the bottom line) between loci A and B lead to essentially uncorrelated trees at A and B.
doi:10.1371/journal.pcbi.1003060.g001

$$\sigma(\Omega_0) \approx n/(2\sqrt{12}),$$

provided n is sufficiently large.

The compound random variables Ω_i, $i>0$, have support which depends on Ω_j, $j<i$. More precisely, the distribution of Ω_i, given Ω_j, $j<i$, is almost-uniform on $\{1,2,\ldots,\lfloor n_i/2 \rfloor\}$ with

$$\text{Prob}(\Omega_i=\omega_i)=p(n_i,\omega_i), \tag{2}$$

where $n_i=n-\omega_0-\ldots-\omega_{i-1}$ $(i>0)$ is a random variable which is bounded below by $n/2^i$ and above by $n-i$. The moments are somewhat more complicated. For instance,

$$\begin{aligned} E(\Omega_1) &= \sum_{\omega_0=1}^{n/2} p(n,\omega_0) \sum_{\omega_1=1}^{(n-\omega_0)/2} \omega_1\, p(n-\omega_0,\omega_1) \\ &= \sum_{\omega_0=1}^{n/2} p(n,\omega_0) \frac{(n-\omega_0)^2}{4(n-\omega_0-1)} \approx \sum_{\omega_0=1}^{n/2} p(n,\omega_0) \frac{(n-\omega_0)}{4} \\ &= \frac{n(3n-4)}{16(n-1)} \approx \frac{3n}{16}. \end{aligned}$$

Continuing this way, evaluating sums iteratively and using the above approximation, one derives

$$E(\Omega_i) \approx \frac{3^i n}{4^{i+1}}. \tag{3}$$

Similary, one can obtain the second moments and combine these to

$$V(\Omega_i) \approx \frac{1}{3}\left(1-\frac{3^i n}{4^{i+1}}\right)^2. \tag{4}$$

Define now the normalized random variables $\Omega_i^*=2\Omega_i/n_i$. Since n is a constant, we have for $\Omega_0^*=2\Omega_0/n$

$$E(\Omega_0^*) \approx 1/2$$

and

$$\sigma(\Omega_0^*) \approx \sqrt{1/12}.$$

To calculate the moments of Ω_i^*, $i>0$, we replace n_i by $E(n_i)$. Simulations suggest that this is acceptable, as long as n_i is not too small. Figure 2 shows this fact for $i \le 3$. Here we focus on Ω_i^* for $i \le k < h$, where k is small and n is large ($k=2$, $n \ge 100$, say). Since,

$$E(n_i)=E(n-\Omega_0-\ldots-\Omega_{i-1}) \approx n\left(\frac{3}{4}\right)^i,$$

we obtain

$$E(\Omega_i^*) \approx \frac{E(2\Omega_i)}{E(n_i)} = \frac{1}{2}. \tag{5}$$

Similarly,

$$V(\Omega_i^*) = \frac{1}{12} + \frac{1}{n^2}\left(\frac{4}{3}\right)^{2i+1} - \frac{2}{3n}\left(\frac{4}{3}\right)^i \approx 1/12 \tag{6}$$

and

$$\sigma(\Omega_i^*) \approx \sqrt{1/12}.$$

It is very convenient to work with the normalized random variables Ω_i^* instead of Ω_i. Their support is bounded by 0 and 1 for all i and they are well approximated by independent continuous uniforms on the unit interval. This considerably facilitates the handling of sums and products of Ω_i^*. For instance, the joint distribution $F^{(k+1)}$ of $\Omega_0^*,\Omega_1^*,\ldots,\Omega_k^*$ is then approximated by the continuous uniform product with distribution function

$$F^{(k+1)}(k,u_0,\ldots,u_k) \approx F(k,u=u_0\cdot\ldots\cdot u_k) = \frac{\Gamma(k+1,-\log(u))}{k!}, \tag{7}$$

expectation

$$E(\otimes_{i=0}^{k}\Omega_i^*) \approx (1/2)^{k+1}$$

and variance

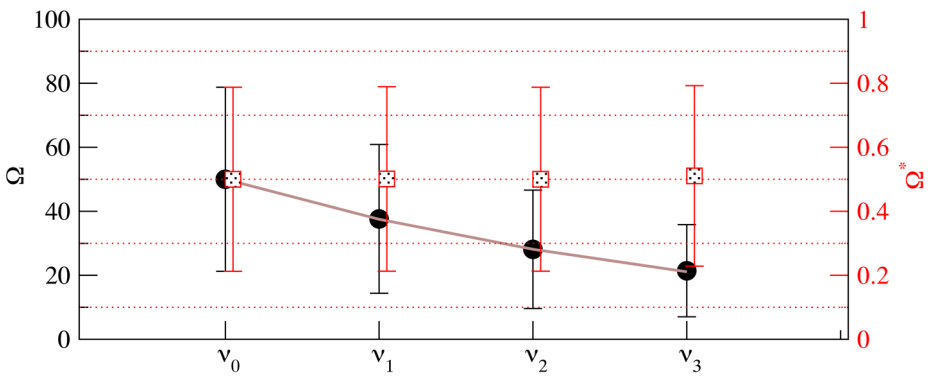

Figure 2. Mean and standard deviation of Ω and Ω^* for coalescent trees of size $n=200$. Shown are the values for 10^4 independent realizations. x-axis: values of Ω (black circles) and Ω^* (red squares) are determined for the subtrees originating at node v_i, $i=0,\ldots,3$. The solid gray line shows the theoretical expectation according to eq (3).
doi:10.1371/journal.pcbi.1003060.g002

$$V(\otimes_{i=0}^{k}\Omega_i^*) \approx (1/3)^{k+1} - (1/2)^{2(k+1)}.$$

The coefficient of variation, $\sqrt{V}/E$, is

$$c_v(\otimes_{i=0}^{k}\Omega_i^*) \approx \sqrt{\left(\frac{4}{3}\right)^{k+1} - 1}.$$

As is well known, the normalized sum of continuous uniforms converges in distribution to a normal random variable rather quickly. In fact, we have for the standardized sum

$$S_k = \sum_{i=0}^{k}\left(\sqrt{\frac{12}{k+1}}\left(\Omega_i^* - E(\Omega_i^*)\right)\right) \sim N(0,1). \qquad (8)$$

In practice, already $k=2$ yields a distribution which is reasonably close to a normal (see Suppl. Figure S1).

Linked trees. Consider now a sample of recombining chromosomes. Coalescent trees along a recombining chromosome are not independent. In particular, tree height and tree topology of closely linked trees are highly correlated. However, under conditions of the standard neutral model, correlation breaks down on short distances (Figure 3) [15]. Roughly 10 recombination events in the sample history reduce correlation by about 50%. Under neutrality and when N is constant, a sample of size n has experienced on average $4Nra_{n-1}$ recombination events [20] (Suppl. Figure S2), where a_n is the n-th harmonic number and represents the length of the tree. Assuming a recombination rate of 1 cM/Mb, population size $N=10^4$ and sample size $n=200$, this amounts to roughly 10 recombination events per 4 kb. If $N=10^5$, in an interval of only about 400 bp correlation is reduced to 50% (Figure 3). Thus, if correlation half-life is determined by roughly 10 events in the sample, we estimate the correlation length L_{half} as

$$L_{\text{half}} \approx \frac{5}{2Nc\,a_{n-1}}, \qquad (9)$$

where c is the recombination rate per bp per unit time. Hence, trees may be regarded as essentially uncorrelated when considering physical distances of some 10 kb and sufficiently large populations and samples.

Eq (9) may be violated if population size N is not constant. As a biologically relevant example we consider a population bottleneck, during which the population is reduced to size N_b. A bottleneck is characterized by three parameters, time of onset, duration (both in units of $4N$) and depth ($d=N_b/N$). A bottleneck induces time dependent changes of the coalescent rate [21] and a reduction of effective population size. Particularly drastic effects on the genealogy are observed when the duration is similar to or larger than the depth [22]. Given biologically reasonable parameters, this inflation may even be larger under a bottleneck than under a selective sweep (Figure 3).

Tree topology in the vicinity of a selective sweep

A positively selected allele sweeping through a population leads to a drastic reduction of tree height due to its short fixation time t^* (see Figure 1C). The fixation time depends on the selection coefficient s and population size N. In units of $4N$, $t^* \approx (1/\alpha)\log(\alpha)$, where $\alpha=2Ns$ [23]. This is much smaller than the neutral average fixation time $t^* \approx 1$. The reduced fixation time leads to a severe reduction of genetic variability. Furthermore, external branches of the tree are elongated relative to internal branches, yielding a star-like phylogeny of an approximate length of nt^*. Replacing the neutral tree length a_{n-1} in eq (9) by this

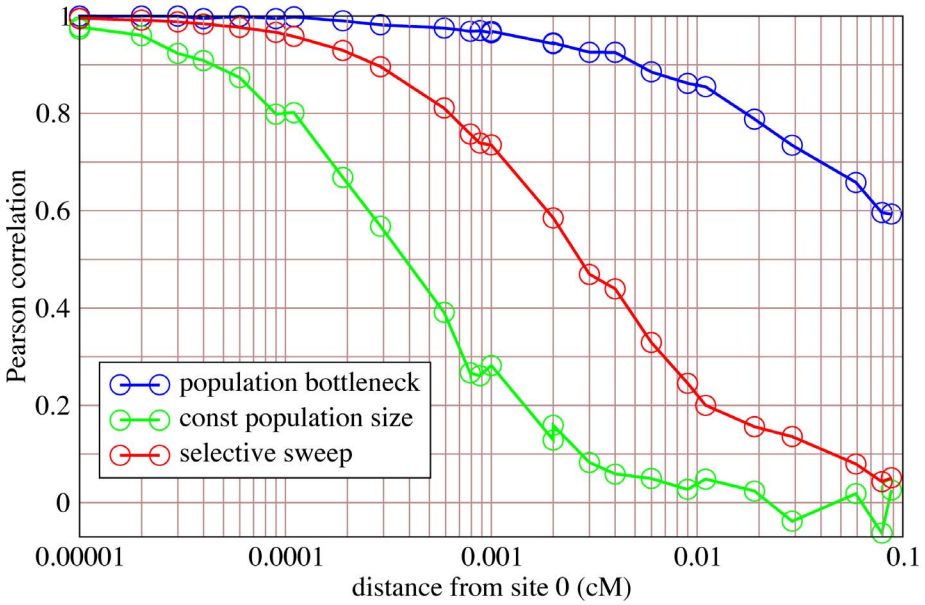

Figure 3. Correlation across distance. Correlation based on simulations (1000 replicates) of the statistic $S_3=2\sum_{k=0}^{2}(\Omega_k^*-0.5)$ of the true tree. Pearson's correlation coefficient is measured between $S_3(0)$ and $S_3(x)$ for pairs of trees at position 0 and position x. Three scenarios are compared: standard neutral model with constant population size (green), population bottleneck (blue) and selective sweep (red). Sample size $n=200$, $\theta=20$, $N=10^5$ and a recombinaton rate of $1cM/Mb$ is assumed. The bottleneck parameters are: duration = depth = 10^{-3}, $\tau=10^{-2}$. The selective sweep has a strength of $\alpha=2Ns=2000$. The selected site is at position $x=0$. Under standard neutrality, 50% correlation is reached at position $x=4\,10^{-4}$ cM, corresponding to about 400 bp.
doi:10.1371/journal.pcbi.1003060.g003

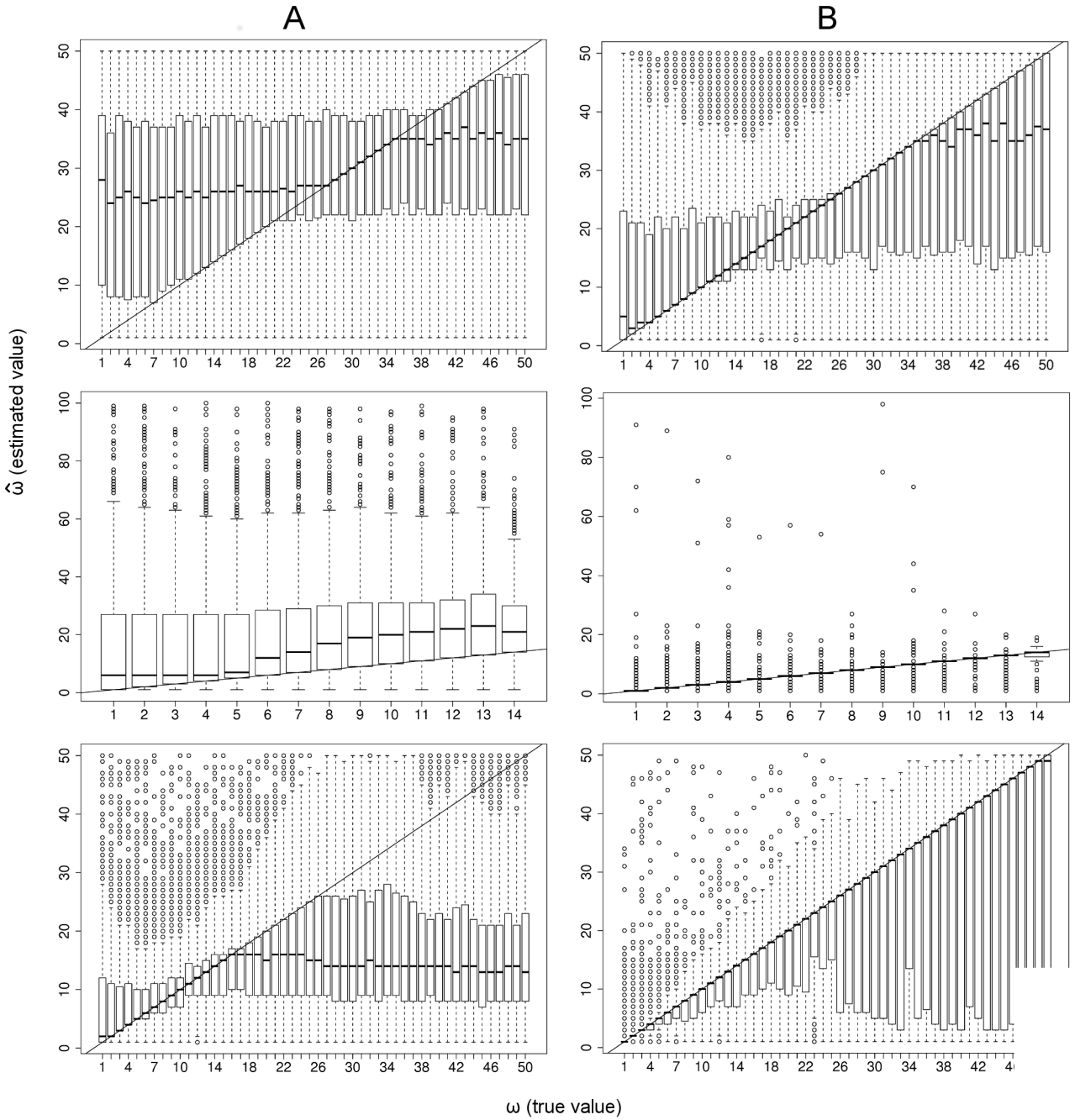

Figure 4. Estimation of Ω. A: estimation of Ω_0 by $\widehat{\Omega}_0$. B: estimation of Ω_0 by $\widehat{\Omega}_0 \cap I$. First row: standard neutral model. Second row: Selective sweep; estimation of Ω_0 at distance $1kb$ from selected site. Third row: Selective sweep; distance $5kb$ from selected site. Parameters: $N = 10^5$; $n = 100$ (top and bottom row); $n = 200$ (middle row); $\theta = 40$; $s = 0.005$; $\tau = 10^{-4}$.
doi:10.1371/journal.pcbi.1003060.g004

figure, we obtain the following estimate for the correlation half-life

$$L_{half} \approx \frac{5s}{\log(\alpha)cn}. \qquad (10)$$

For the parameters used in Figure 3, we have $L_{half} \approx 3300$ bp, which agrees well with the simulation result.

In contrast to tree height and length, tree topology at the selected site does not necessarily differ from a neutral tree; only when moving away from the sweep site, and with recombination, topology may drastically change. In fact, given a shallow tree, recombination leads with high probability to an increase of tree height and to unbalanced trees [15]. Thus, recombination events next to the selected site tend to increase tree height (see sketch in Figure 1B and C) and to create a bias in favour of unbalanced trees, i.e. trees with small Ω_0 (Figure 4A). The expected *proximal* distance x_p from the selected site of such a recombination event can be estimated as

$$x_p \approx 1/r_u, \tag{11}$$

where $r_u = c\ n\ t^*/2$, c is the per site recombination rate, and $n\ t^*$ is the length of a star-like phylogeny; the factor $1/2$ accounts for the fact that it is more likely to recombine with an ancestral chromosome (thereby increasing tree height) as long as these are more abundant than the derived chromosomes carrying the selected allele. Roughly, this is the case during the first half of the fixation time t^*. Assuming instead of the star phylogeny a random tree topology of average length $a_{n-1}\ t^*$ at the selected site, one obtains the larger (call it *distal*) estimate

$$x_d \approx 1/r_l, \tag{12}$$

where $r_l = c\ a_{n-1}\ t^*/2$.

Unbalanced trees tend to have strongly elongated root branches and harbor an over-abundance of high frequency derived SNP alleles [6,16]. With microsatellites it is usually not possible to determine the ancestral and derived states of an allele, because they mutate at a high rate and possibly undergo back-mutation. However, under the symmetric single step mutation model, the expected distance between a pair of alleles (in terms of motif copy numbers) behaves as the distance in a one-dimensional symmetric random walk and therefore increases at a rate proportional to the square root of the scaled mutation rate θ (see Methods). Thus, alleles which are separated by long root branches tend to form two distinct allele clusters.

Estimating Ω

Tree topology is ususally not directly observable and has to be estimated from data. We focus on estimating Ω_i, $i \le k < h$, from microsatellite data. Given a sample of n microsatellite alleles with tandem repeat counts $|A_j|$, $1 \le j \le n$, we use UPGMA [24] to construct a hierarchical cluster diagram. If subtree topology within a particular cluster node should not be uniquely re-solvable, for instance if alleles are identical, we randomly assign the alleles of the subtree under consideration to two clusters with equal probability. This gives preference to clusters of balanced size in case of insufficient resolution. We then use the inferred tree topology $\widehat{\mathbf{\Omega}}_i$ to estimate Ω_i of the true tree. This procedure is conservative for the test statistics described below, since it gives preference to large values $\widehat{\omega}$ when the true value ω is small (Figure 4, column A). For a cluster pair C_1, C_2, define the distance as

$$\mathrm{dist} = \min_{i \in C_1, j \in C_2} |(|A_i| - |A_j|)|. \tag{13}$$

We find that UPGMA clustering gives good estimates of Ω_0 when clusters are clearly separated from each other, i.e. when $\mathrm{dist} > 1$. Let I be the indicator variable for this event. Then, we have for the median

$$\mathrm{med}(\widehat{\mathbf{\Omega}}_0 \cap I | \omega_0) \approx \omega_0,$$

(Figure 4, column B). Without requiring $\mathrm{dist} > 1$ the estimate $\widehat{\mathbf{\Omega}}_0$ is more biased. In part, this is due to the conservative UPGMA strategy mentioned above. However, estimation of Ω_0 is very accurate when root branches are strongly elongated, i.e. under conditions of selective sweeps or certain bottlenecks (Figure 4, bottom).

Application: Testing the neutral evolution hypothesis

We now turn to an application of the above results and explain how a new class of microsatellite based tests of the neutral evolution hypothesis can be defined.

Consider a sample of n alleles at a microsatellite marker and record their motif repeat numbers. Applying UPGMA clustering to the alleles, we obtain estimates $\widehat{\mathbf{\Omega}}_i$, $i \le k$ as described above. These are transformed to $\widehat{\mathbf{\Omega}_i^*} = 2\widehat{\mathbf{\Omega}}_i/n_i$. Then, we determine the following test statistics

$$T_k^{(\mathrm{sum})} := \widehat{S_k} = \sqrt{\frac{12}{k+1}} \sum_{i=0}^{k} \left(\widehat{\mathbf{\Omega}_i^*} - \frac{1}{2} \right) \tag{14}$$

$$T_k^{(\mathrm{product})} := \prod_{i=0}^{k} \widehat{\mathbf{\Omega}_i^*} \tag{15}$$

$$T_0^{(\mathrm{dist})} := \widehat{\mathbf{\Omega}_0^* \cap I} \tag{16}$$

Thus, the test variable $T_k^{(\mathrm{sum})}$ in eq (14) is the estimate of S_k given in eq (8). Similarly, $T_k^{(\mathrm{product})}$ and $T_0^{(\mathrm{dist})}$ are the estimates of the product $\Pi_{i=0}^k \Omega_i^*$ and of $\Omega_0^* \cap I$.

We now test the null hypothesis $T^{(\cdot)} > q$ for a critical value $q = q(\alpha)$. For a given level α we obtain the critical value q for $T^{(\mathrm{sum})}$ from the standard normal distribution and for $T^{(\mathrm{product})}$ from the uniform product distribution in eq (7) (Table 1). For $T^{(\mathrm{dist})}$ we use the critical value of the normalized version of eq (1). Generally, these critical values are conservative, since Ω_i^* tends to over-estimate Ω_i, when small (Figure 4). In particular, statistic $T^{(\mathrm{dist})}$ is very conservative due to the additional condition on the distance. The true critical values for level α would be larger than those shown in Table 1.

False positive rates and power. First, we analyzed the false positive rates under the standard neutral scenario (i.e., constant N) for different mutation rates θ and varying sample sizes n. As reference parameter settings for simulations with `msmicro` (see Methods) we use sample size $n = 200$, microsatellite mutation rate $\theta = 40$ and recombination rate $r = 400$. The latter corresponds to a recombination rate of 10^{-8} per bp per chromosome, when one assumes a population size of $N = 10^5$ and a size of the investigated genomic region of 10^5 bp ($r = 4N \cdot 10^{-8} \cdot 10^5$). We placed 15 microsatellite markers at positions 1, 10, 30, 60, 70, 80, 85, 87, 88, 89, 90, 91, 92, 95, 100 kb. As expected, we find that the false positive rates remain below their theoretical expectation for all parameter choices θ and n (Figure 5 top; Tables 2 and 3). For the simulations with selection we assumed that a site at position 89 kb was undergoing a selective sweep with selection coefficient $s = 0.005$ or $s = 0.01$. The time since completion of the sweep

Table 1. Critical values for the tests considered in eqs (14)–(16).

α	$T_2^{(\mathrm{sum})}$	$T_2^{(\mathrm{product})}$	$T_0^{(\mathrm{dist})}$
0.01	−2.32635	0.0002235	0.01
0.05	−1.64485	0.0018441	0.05

doi:10.1371/journal.pcbi.1003060.t001

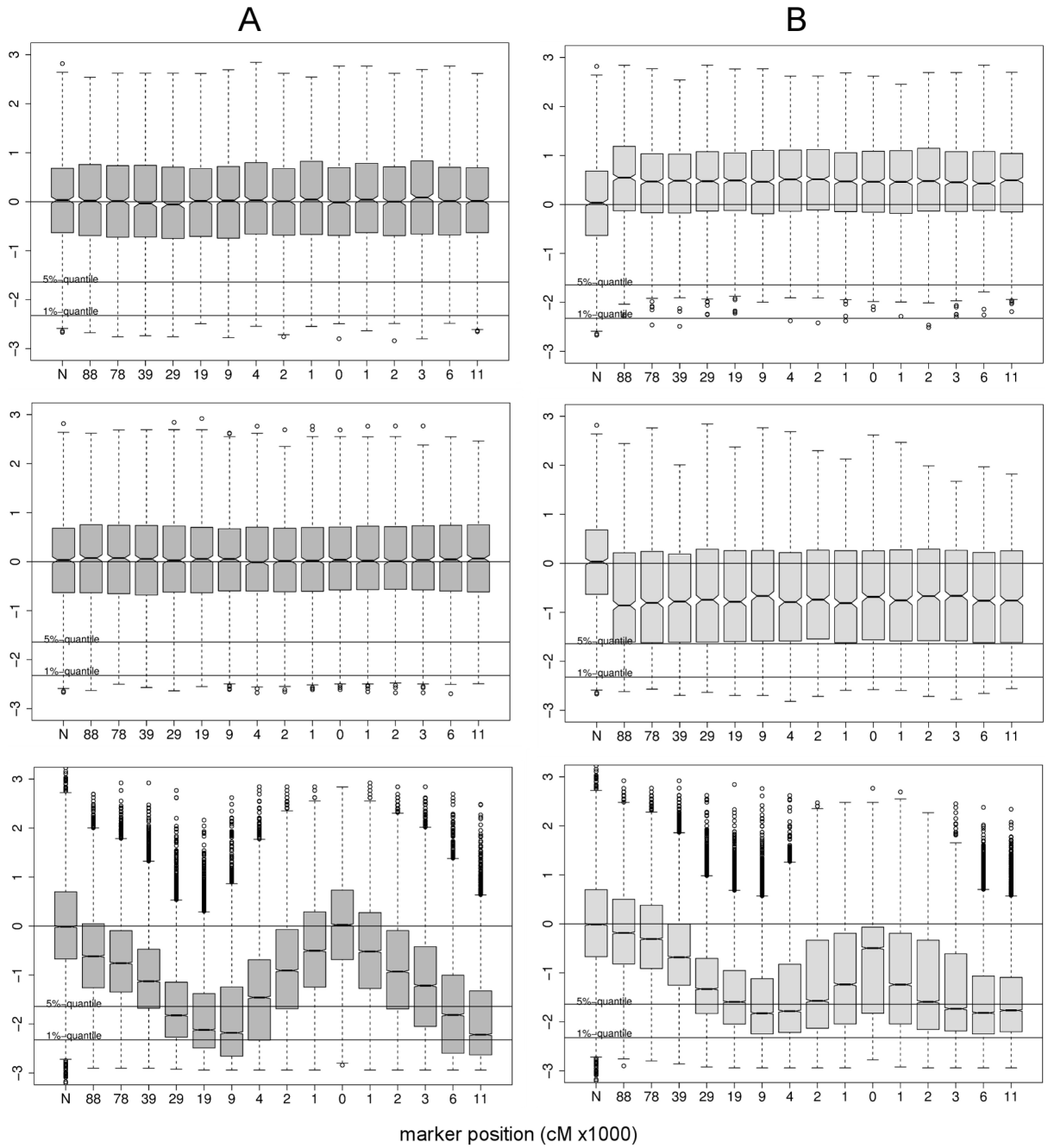

Figure 5. Profile of S_2 and $\hat{S}_2$ along a recombining chromosome. Plots in column A show the distribution of $S_2 = \sum_{i=0}^{2} 2(\Omega_i^* - 1/2)$, i.e. when the tree topology is known. Plots in column B show the distribution of the estimate $\hat{S}_2 = \sum_{i=0}^{2} 2(\widehat{\Omega_i^*} - 1/2)$ when the tree topology is unknown, but estimated from microsatellite polymorphism data. Each boxplot corresponds to one of 15 marker loci located at the positions indicated on the $x-$axis. The regions spans 100 kb in total. Symmetric step-wise mutation model with $\theta = 40$. Other parameters: $n = 200$, $N = 10^5$ and recombination rate per *bp* $c = 10^{-8}$ (corresponding to 1 cM/Mb). First row: standard neutral model with constant N. Second row: bottleneck model with severity 1 and onset $\tau = 0.01$. Third row: Selective sweep at locus $x = 0$ with $s = 0.005$ which was completed $\tau = 10^{-4}$ time units ago. For comparison with the theoretical expectation, the leftmost boxplot in each panel shows the standard normal distribution (labeled 'N').
doi:10.1371/journal.pcbi.1003060.g005

was an adjustable parameter τ, with the reference setting $\tau = 10^{-4}$. We simulated hard selective sweeps, i.e. the selected allele is introduced as a single copy and fixed with probability about $2s$. The test statistic $T_2^{(\mathrm{sum})}$ is shown in Figure 5 and power profiles for all three tests in Figure 6. We find that maximum power of the tests is attained within the interval given by eqs (11) and (12) (Figure 6 and Tables 4 and S1). Depending on the strength of selection, maximum power is close to the upper interval bound at x_d ($s = 0.005$, Table S1), or removed from x_d towards the interior of the interval ($s = 0.01$, Table 4). This is in agreement with the expectation that only very strong selective sweeps generate a star-like phylogeny, which lead to the proximal estimate x_p in eq (11). Thus, the location of the power maximum depends on the strength of selection and the details of the tree topology at the selected site.

Maximum power for the compound tests $T_2^{(\mathrm{sum})}$ and $T_2^{(\mathrm{product})}$ is more removed from the selected site than for the simple test $T_0^{(\mathrm{dist})}$. The latter measures imbalance only at the root node v_0 and is most sensitive to single recombination events between marker and selected site, while multiple events blur the effect. The power of all tests is sensitive to the mutation rate and to sample size (Tables S2 and S3). For the parameters tested, the power of the simple $T_0^{(\mathrm{dist})}$ increases when θ or n increase. For $T_2^{(\mathrm{sum})}$, maximum power is reached for $\theta \approx 20$. Very small, as well as very high, mutation rates produce little power. Realistic mutation rates in insects and vertebrates are between $\theta = 5$ and 50 [25–27], thus within the powerful domain. Importantly, power can be increased by increasing sample size: all of the above tests become more powerful for large samples (Tables S3, S4 and S5). Since the tests consistently underscore the theoretical false positive rate, relaxed singnificance levels (for instance $\alpha = 0.05$) can be applied. At level $\alpha = 0.05$ test $T_2^{(\mathrm{sum})}$ has power of more than 80% to detect recent selective sweeps (Figure 6 and Table 4). For intermediate mutation rates power of test $T_2^{(\mathrm{sum})}$ is somewhat higher than of $T_2^{(\mathrm{product})}$ (Table S2). Generally, power profiles of $T_2^{(\mathrm{sum})}$ and $T_2^{(\mathrm{product})}$ follow qualitatively the same pattern. In contrast, power of test $T_0^{(\mathrm{dist})}$ may be quite different. Interestingly, $T_0^{(\mathrm{dist})}$ performs better than $T_2^{(\mathrm{sum})}$ or $T_2^{(\mathrm{product})}$ when selection is only moderately strong. Unsurprisingly, power of all tests depends heavily on the strength of selection. Also, the time since completion of the selective sweep influences power. Reasonable power can be reached if $\tau < 10^{-3}$ in coalescent units.

We also examined how much the tests are confounded by deviations from the standard neutral model. First, we determined the false positive rates under a population bottleneck. From other studies it is known that bottlenecks with a severity (duration divided by depth) around 1 are particularly problematic [16,28]. We find that tests $T_2^{(\mathrm{sum})}$ and $T_2^{(\mathrm{product})}$ can produce substantially more false positives than expected, in particular if bottlenecks are recent (Table S6). Interestingly, test $T_0^{(\mathrm{dist})}$ is very robust against these disturbances and the false positive rate remains clearly under the theoretical value for all onset parameters tested (Table S6).

We note that the false positive rates of $T_2^{(\mathrm{sum})}$ and $T_2^{(\mathrm{product})}$ depend strongly on the bottleneck duration even when the severity is kept fixed (Table S7). Very short (duration 0.001), but heavy reductions of N are more disturbing for $T_2^{(\mathrm{sum})}$ and $T_2^{(\mathrm{product})}$ than long, but shallow bottlenecks (duration 0.1). In contrast, $T_0^{(\mathrm{dist})}$ is fairly insensitive to changes of bottleneck duration (Table S7).

Under a model of fast population expansion (expansion rate 10), all tests remain below, or close to, their theoretical false positive rate. Again, test $T_0^{(\mathrm{dist})}$ is unsensitive to population expansion and varying onset times (Table S8).

We expected that our topology based tests would yield many false positives under a model of population subdivision. As a potentially critical case we examined sampling from a population divided into two sub-populations which split $2N$ generations ago and which exchange migrants at rate m. We analyzed both varying migration rates and varying sampling schemes (Tables S9 to S12). The false positive rate for tests $T_2^{(\mathrm{sum})}$ and $T_2^{(\mathrm{product})}$ remains clearly under its theoretical expectation, even if sampling is heavily biased (sample size of sub-population 1 was $n_1 = 195$ and of sub-population 2 was $n_2 = 5$; Table S9). In contrast, test $T_0^{(\mathrm{dist})}$, which only measures tree imbalance at the root node, is more vulnerable to biased sampling from a sub-divided population. The false-positive rate grows up to 17% if $n_1 = 195$ and $n_2 = 5$. In general, we find test $T_0^{(\mathrm{dist})}$ to be less vulnerable to population bottlenecks, but tests $T_2^{(\mathrm{sum})}$ and $T_2^{(\mathrm{product})}$ to be more robust under population substructure.

Finally, we examined how deviation from the single step mutation model would influence our tests. We modified the mutation model and allowed occasional jumps (probability p) of larger steps. We tested jumps of step size 2 (Table S13) and 7 (Table S14). All tests, eminently the compound tests $T_2^{(\mathrm{sum})}$ and $T_2^{(\mathrm{product})}$, remain clearly below their theoretical false positive rate.

Case study

Emergence of drug resistance in malaria parasites is among the best documented examples for recent selective sweeps. We re-analyzed 16 microsatellite markers surrounding a well studied drug resistance locus of malaria parasites [29] (Figure 7). The signature of recent positive selection is consistently detected by all tests on two markers somewhat downstream of the drug resistance locus *pfmdr1* (marker l–35 and l–37 in the notation of [29]; Table 5). Highest significance is reported by test $T_2^{(\mathrm{product})}$ (p-value close to 0.001). $T_0^{(\mathrm{dist})}$ reports a p-value of 0.006 and $T_2^{(\mathrm{sum})}$ reports p-values slightly above 0.010. In addition, $T_2^{(\mathrm{product})}$ reports locus l–29 (located upstream of *pfmdr1*) to be significant at $p = 0.025$. This locus is also detected by $T_0^{(\mathrm{dist})}$ ($p = 0.038$). Other four loci are reported only by $T_0^{(\mathrm{dist})}$ (l–30 ($p = 0.006$), l–31 ($p = 0.025$), l–32 ($p = 0.006$), l–30 ($p =$) and l–40 ($p = 0.031$)). Discrepancies in the test results are due to their different sensitivities to various parameters. The simple and compound tests have different power profiles with power peaks at different positions from the selected site (Figure 6). Plasmodium in South-East Asia is most likely expanding and sub-structured; however, there is only limited knowledge about the details.

As shown above, $T_0^{(\mathrm{dist})}$ is quite sensitive to biased sampling from different sub-populations. Some of the significant results of $T_0^{(\mathrm{dist})}$ may be inflated due to sub-structure. There is also some disagreement between tests $T_2^{(\mathrm{sum})}$ and $T_2^{(\mathrm{product})}$ regarding significance, although both test imbalance at tree nodes v_0, v_1 and v_2. In fact, the cases reported by the two tests may still differ in their details. Comparing the three components $\hat{\omega}_0^*$, $\hat{\omega}_1^*$ and $\hat{\omega}_2^*$ with respect to their maximum and minimum, we find that the cases reported as significant by $T_2^{(\mathrm{sum})}$ have a $\max(\hat{\omega}_0^*,..,\hat{\omega}_2^*) < 0.4$ and a $\min(\hat{\omega}_0^*,..,\hat{\omega}_2^*)$ up to 0.1. In contrast, for $T_2^{(\mathrm{product})}$, the maximum is close to 1.0 while the minimum tends to be less than 0.04 (Figure S4). Thus, test $T_2^{(\mathrm{sum})}$ is more restrictive in the sense that all components $\hat{\omega}_0^*$, $\hat{\omega}_1^*$ and $\hat{\omega}_2^*$ have to be small to yield a significant result. $T_2^{(\mathrm{product})}$ is more permissive and accepts that one of the three components may be large.

All tests agree on significance of two markers close to a site which was previously shown to have experienced a selective sweep. They also agree all on strongly increased p-values in the immediate vicinity of the selected site (l–33, l–34). Together, these results confirm the accuracy and practical utility of our tests.

Discussion

The binary coalescent has a number of well-studied combinatoric and analytic properties [1,30,31]. Here we only concentrate on tree topology and use a classic result of Tajima [19] to define a simple measure, Ω_i, of tree balance. It is the minimum of the left and right subtree sizes under internal node v_i. Its normalized version is approximately uniform on the unit interval and the

Table 2. Empirical false positive rate for varying θ.

	$\alpha = 0.01$			$\alpha = 0.05$		
θ	$T_2^{(\mathrm{sum})}$	$T_2^{(\mathrm{product})}$	$T_0^{(\mathrm{dist})}$	$T_2^{(\mathrm{sum})}$	$T_2^{(\mathrm{product})}$	$T_0^{(\mathrm{dist})}$
0.1	0.00006	0.00327	0.0001	0.00035	0.01323	0.00047
0.5	0.00523	0.01703	0.00109	0.01724	0.07931	0.00431
1.0	0.01142	0.01749	0.002	0.0463	0.08957	0.00887
1.5	0.01251	0.01414	0.00281	0.06365	0.08425	0.01145
2.0	0.01145	0.01127	0.00355	0.06736	0.07399	0.01354
2.5	0.00933	0.00843	0.00421	0.06579	0.06571	0.01549
3.0	0.00756	0.00663	0.00458	0.06042	0.05718	0.01781
4.0	0.00559	0.00478	0.00534	0.04936	0.04348	0.01884
5.0	0.00415	0.00315	0.0057	0.04073	0.03455	0.0208
10.0	0.00145	0.00131	0.00616	0.0244	0.01889	0.02433
20.0	0.00069	0.00049	0.00632	0.01411	0.01032	0.02685
30.0	0.00064	0.00038	0.00656	0.01018	0.00805	0.02744
40.0	0.00043	0.00035	0.00647	0.00839	0.00651	0.02702
50.0	0.00027	0.00031	0.006	0.00828	0.00631	0.02754
100.0	0.00028	0.00033	0.00615	0.00693	0.00591	0.02846
120.0	0.00024	0.00027	0.00614	0.00666	0.00571	0.02806
150.0	0.00024	0.00028	0.00593	0.00699	0.00548	0.02876
200.0	0.00034	0.00026	0.00624	0.00641	0.005	0.02844

Neutral model N constant, $n = 200$, $r = 400$. Significance levels α are based on theoretical formulae according to eqs (7) and (8) (reference value indicated in bold).
doi:10.1371/journal.pcbi.1003060.t002

summation over internal nodes v_i, $i = 1,..k$, is close to normal. Another summary statistic of tree balance is Colless' index C [32]. It also depends on the sizes of left- and right subtrees of the internal nodes, but its distribution is more complicated. C has received attention in the biological literature before [33] and, more recently, in theoretical studies, for instance by Blum&Janson [34]. A problem with Colless' index is that it is difficult to estimate if the true tree structure is unknown. But, limiting attention to the tree structure close to the root, we show that the balance measure Ω can be estimated, for instance, from microsatellite allele data by a clustering method. We found that a version of UPGMA clustering gives most reliable results.

Coalescent trees for linked loci are not independent. However, correlation dissipates with recombinational distance. In fact, under neutral conditions only about ten recombination events are sufficient to reduce correlation in tree topology by 50%. Thus, estimating tree imbalance at multiple microsatellites can be performed independently for each marker, if they are sufficienty distant from each other. Conversely, with a very small number of recombination events, Ω is not drastically altered on average [15]. Thus, when working with SNPs, one may afford to consider haplotype blocks containing a few more recombination events than segragting sites and still be able to reconstruct a reliable gene genealogy. This possibility will be explored in more detail elsewhere.

Table 3. Empirical false positive rate for varying sample size n.

	$\alpha = 0.01$			$\alpha = 0.05$		
n	$T_2^{(\mathrm{sum})}$	$T_2^{(\mathrm{product})}$	$T_0^{(\mathrm{dist})}$	$T_2^{(\mathrm{sum})}$	$T_2^{(\mathrm{product})}$	$T_0^{(\mathrm{dist})}$
10	N/A	N/A	0.21417	0.0099	N/A	0.21417
20	0.00035	0	0.09527	0.01215	0.00035	0.09527
50	0.00055	0.00003	0.03318	0.0094	0.00286	0.03318
100	0.00052	0.00022	0.0151	0.00925	0.00527	0.02778
150	0.00044	0.00033	0.00902	0.00934	0.00609	0.02411
200	0.00039	0.00033	0.00592	0.00943	0.00684	0.02666
300	0.00038	0.00042	0.00388	0.00976	0.00828	0.02282
500	0.00042	0.00055	0.00394	0.01009	0.0093	0.02169
1000	0.00044	0.00104	0.00474	0.01107	0.01148	0.02057

Neutral model N constant, $\theta = 40$, $r = 400$. Significance levels α are based on theoretical formulae according to eqs (7) and (8) (reference value indicated in bold).
doi:10.1371/journal.pcbi.1003060.t003

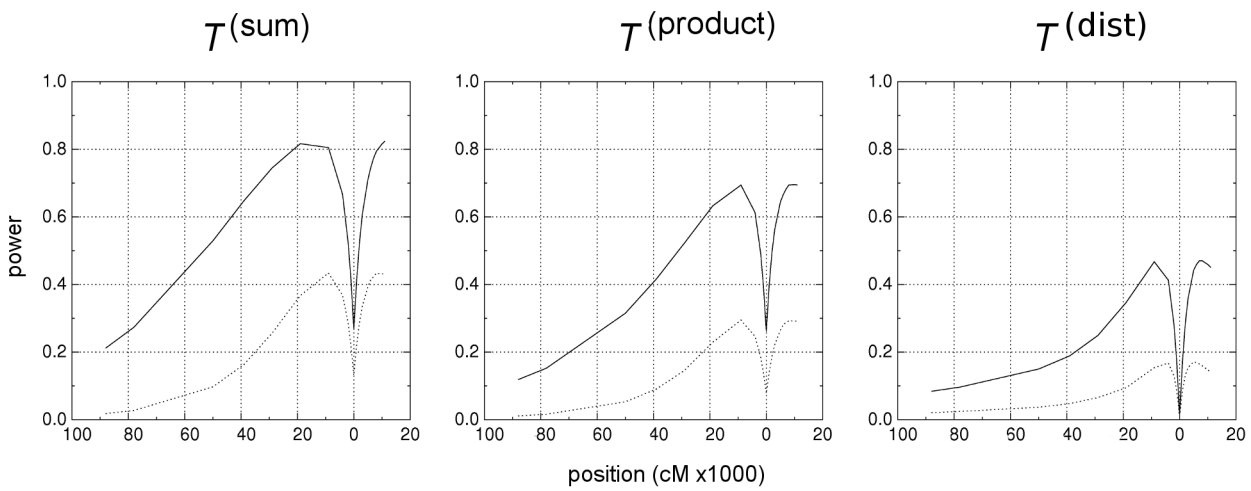

Figure 6. Power to detect loci under recent selection by the three tests defined in eqs (14) to (16). Parameters: level $\alpha=0.05$ (solid) and $\alpha=0.01$ (dotted); selection coefficient $s=0.01$; time since fixation $\tau=10^{-4}$; sample size $n=200$; mutation rate $\theta=40$; recombination rate $c=10^{-8}$. The $x-$axis shows positions to the left (negative values) and right (positive values) of the locus under selection at position $x=0$. Scale is in cM x1000, corresponding here to kb.
doi:10.1371/journal.pcbi.1003060.g006

Microsatellites have been used before as markers for selective sweeps. Schlötterer et al. [35] have proposed the lnRH statistic to detect traces of selection and Wiehe et al. [28] have shown that a multi-locus vesion of lnRH for linked markers can yield high power while keeping false positive rates low. However, a severe practical problem with the lnRH statistic is that it requires data from two populations, and for each of them two additional and independent sets of neutral markers for standardization. There are a few methods to detect deviations from the standard neutral model based on single microsatellite locus data from one

Table 4. Power of $T^{(\mathrm{sum})}$, $T^{(\mathrm{product})}$ and $T^{(\mathrm{dist})}$ in dependence of distance to selected site.

	$\alpha=0.01$			$\alpha=0.05$			
distance (kb)	$T_2^{(\mathrm{sum})}$	$T_2^{(\mathrm{product})}$	$T_0^{(\mathrm{dist})}$	$T_2^{(\mathrm{sum})}$	$T_2^{(\mathrm{product})}$	$T_0^{(\mathrm{dist})}$	**SKD***
−88.0	0.01794	0.01085	0.02051	0.21164	0.11855	0.08392	0.8468
−78.0	0.02708	0.01613	0.02416	0.27325	0.15216	0.09606	0.8873
−50.0	0.09714	0.0528	0.03672	0.52898	0.31465	0.1497	0.9353
−39.0	0.16291	0.09047	0.04749	0.64722	0.41612	0.1887	0.9440
−29.0	0.25581	0.14603	0.06525	0.74461	0.52288	0.24893	0.9435
−19.0	0.3671	0.22901	0.09412	0.81644	0.63249	0.34637	0.9161
−9.0	0.4339	0.29439	0.15377	0.80504	0.69404	0.46718	0.7931
−4.0	0.3679	0.24615	0.16738	0.66659	0.6127	0.41263	0.5969
−2.0	0.28585	0.18531	0.13051	0.52249	0.49368	0.28243	0.4535
−1.0	0.21826	0.1383	0.08574	0.41239	0.39426	0.16926	0.3657
0.0	0.13085	0.07765	0.00971	0.26692	0.26055	0.01182	0.2670
1.0	0.21972	0.13898	0.08428	0.41424	0.39615	0.17043	0.3601
2.0	0.28701	0.18614	0.12943	0.52215	0.49205	0.28118	0.4600
3.0	0.33496	0.22226	0.1548	0.6066	0.56351	0.35927	0.5366
5.0	0.39321	0.26452	0.1706	0.71037	0.64409	0.44278	0.6549
6.0	0.41455	0.28116	0.16901	0.74565	0.66566	0.45932	0.6928
7.0	0.42253	0.28768	0.1645	0.77335	0.68215	0.47001	0.7407
8.0	0.4334	0.29241	0.15955	0.79386	0.69415	0.47023	0.7652
10.0	0.43149	0.29145	0.14693	0.81588	0.69532	0.45901	0.8091
11.0	0.43158	0.2914	0.13982	0.82358	0.69425	0.45046	0.8336

Selective sweep with $s=0.01$, $\tau=10^{-4}$, $\theta=40$, sample size $n=200$.
*SKD-test by Schlötterer *et al.* [37].
doi:10.1371/journal.pcbi.1003060.t004

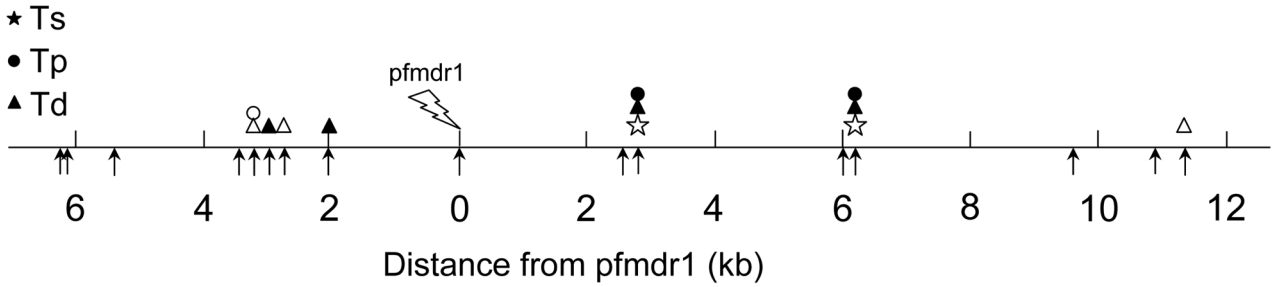

Figure 7. Traces of selection around a drug resistance locus in Plasmodium. Results of tests $T^{(\mathrm{sum})}$ (stars), $T^{(\mathrm{product})}$ (circles) and $T^{(\mathrm{dist})}$ (triangles) applied to a 17 kb region sorrounding the *pfmdr1* locus in *P.falciparum*. Shown are significant results on the 5% (open symbols) and 1% (filled symbols) levels. Positions of the examined microsatellite markers are indicated by arrows. Data from [29].
doi:10.1371/journal.pcbi.1003060.g007

population. For instance, the test by Cornuet and Luikart [36], which compares observed and expected heterozygosity, is designed to detect population bottlenecks. A test by Schlötterer et al. [37] uses the number of alleles at a microsatellite locus and determines whether an excess of the number of alleles is due to positive selection (SKD test). However, as the authors pointed out, the test depends critically on a reliable locus-specific estimate of the scaled mutation rate. We have compared SKD and the test proposed here with respect to power and false positive rates. While the SKD-test is generally more powerful, especially at larger distances from the selected site (Table 4 and Suppl. Tables S1, S5), it has higher false positive rates than the tests proposed here, in particular when compared to $T_0^{(\mathrm{dist})}$ (Suppl. Table S6), and for non-standard mutation models (Suppl. Tables S13, S14). Note also that under population sub-structure SKD yields up to 100 times more false positives than our tests (Suppl. Tables S9 to S12).

It should be emphasized that it is the topology of the underlying genealogical tree, not the genetic variation, which constitutes the basis for the test statistics proposed here. The two steps, estimating topology, and performing the test are two distinct tasks. The quality of the tests hinges on the quality of the re-constructed genealogy. With a perfectly re-constructed genealogy the false positive rates are completely independent from any evolutionary mechanisms which do not affect the average topology, such as historic changes of population size. However, simulations show that power would still remain under 100% in this case. The robustness of topology based tests with respect to demographic changes has been shown before by Li [16] for a similar test which uses SNP data to reconstruct Ω_0. But Li's test can only be performed if an additional non-topological criterion is satisfied and thus can only test a subset of trees with Ω_0. The tests $T^{(\mathrm{sum})}$ and $T^{(\mathrm{product})}$ defined here rely only on topological properties of the

Table 5. Test statistics and *p*-values for the empirical data set of *P.falciparum*.

pos		$T_2^{(\mathrm{sum})}$	*p*-value	$T_2^{(\mathrm{product})}$	*p*-value	$T_0^{(\mathrm{dist})}$		
						$\widehat{\omega_0}$	*n*	dist[1]
I–25	953,644	−0.1906	0.4244	0.0569	0.4537	146	324	1
I–26	953,768	0.5591	0.7120	0.1235	0.6520	148	322	2
I–27	954,506	−0.7872	0.2156	0.0282	0.3085	95	320	2
I–28	956,456	−1.2289	0.1096	0.0069	0.1268	11	324	3
I–29	956,686	−0.8912	0.1864	0.0007	0.0254*	6*	314	4
I–30	956,917	−0.6710	0.2511	0.0019	0.0521	1**	325	3
I–31	957,169	−1.3464	0.0891	0.0030	0.0706	4*	322	4
I–32	957,861	−0.3083	0.3789	0.0024	0.0598	1**	325	31
I–33	**959,894**	0.8260	0.7956	0.0101	0.1629	147	326	2
I–34	962,445	−0.0498	0.4801	0.0611	0.4706	140	326	2
I–35	962,699	−2.1600	0.0154*	1.9e-5	0.0014**	1**	326	23
I–36	965,905	−0.7369	0.2306	0.0337	0.3415	36	326	2
I–37	966,096	−2.2470	0.0123*	1.4e-5	0.0010**	1**	326	9
I–38	969,495	0.3941	0.6533	0.1713	0.7402	117	323	2
I–39	970,775	0.1901	0.5754	0.0528	0.4366	17	322	3
I–40	971,251	−0.8336	0.2023	0.0025	0.0616	5*	323	2

Given are the theoretical *p*-values based on the standard normal (for $T_2^{(\mathrm{sum})}$) and on the product uniform (for $T_2^{(\mathrm{product})}$) distributions. Values for $T_0^{(\mathrm{dist})}$ are given as raw data ($\widehat{\omega_0}$, *n*, δ). The *p*-value is $2\widehat{\omega_0}/n$. 5% (single star) and 1% (double star) significance are indicated. Marker positions are taken from [29]. The region analyzed (about 17 kb) corresponds to about 1 cM (site under selection in bold).
[1]defined in eq (13).
doi:10.1371/journal.pcbi.1003060.t005

genelaogy and we argue that multi-allelic markers, such as microsatellites, help estimating the true genealogy and improving test results. Although our analyses and simulations are based on the binary Kingman [1] coalescent, we expect that the new test statistics should be robust also under more general coalescent models, for instance when multiple mergers during the selective sweep phase are allowed [38].

Despite a shift to high throughput sequencing technologies in the last decade, microsatellite typing continues to be a cost-efficient and fast alternative to survey population variability in many experimental studies. This is in particular true for projects directed towards parasite typing, e.g. of Plasmodium, and projects with non-standard model organisms, e.g. social insects [39,40], but also for many biomedical studies.

Methods

Coalescent simulations

We simulated population samples under neutral and hitchhiking models with modified versions of the procedures described by Kim and Stephan [41] and Li and Stephan [42] and of `ms` [43], termed `msmicro`. In the modified versions we incorporated evolution of microsatellite loci under the symmetric, single step and multi-step mutation models. Microsatellite mutations are modeled as changes to the number of motif repeats, where only numbers but not particular sequence motifs are recorded. Output data comprise coalescent trees in Newick format and the state of microsatellite alleles for each of n sequences. With `msmicro` also multiple linked microsatellites can be modeled. Coalescent simulations were run under different evolutionary conditions: neutral with constant population size ($N=10^5$), neutral with bottleneck (bottleneck severity $\frac{\text{duration}}{\text{depth}}=\frac{0.001}{0.001}=1$, time since bottleneck $\tau=0.01$), population size expansion (growth rate 10), neutral two-island model with migration, and hard selective sweeps (selection $s=0.01$ and $s=0.005$, time since fixation of sweep allele $\tau=10^{-4}$).

Tree topology

Realizations ω_i of the 'true' random variables Ω_i, $0\le i\le k$ were extracted from the simulation results. Estimation of $\hat{\omega}_i$ was performed by UPGMA hierarchical clustering. If a cluster node could not be uniquely resolved then we gave preference to a bipartite partition in which the left and right subtrees were of equal or similar size. This was accomplished by randomly assigning alleles to two clusters with equal probability. To estimate $\hat{\omega}_0$ we also explored a simple clustering method which works in the following way: we first sorted alleles by size; then we divided the sorted list into two halfs. The separator was placed between those two alleles which had maximal distance (in terms of microsatellite repeat units) from each other. If this was not unique, the separator was placed between those two alleles that resulted in two sets of most similar size. While this clustering method is very effective in estimating ω_0, it is less accurate than UPGMA clustering for ω_i, $i>0$.

Distance between microsatellite alleles

The single step symmetric mutation model behaves as a one-dimensional symmetric random walk of step size one. The theory of random walks (e.g. [44]) tells that the average distance between the origin of the walk and the current position scales with the square root of the number k of steps. More precisely,

$$E_{\text{dist}}=\sqrt{\frac{2k}{\pi}}.$$

The variance is linear in k. Here, steps are represented by mutational events occuring at rate θ. Thus, $E_{\text{dist}}=\sqrt{2\theta/\pi}$ and $V_{\text{dist}}\approx\theta/e$, where e is Euler's constant. The empirical distance between two clusters C_1 and C_2 can be calculated as

$$\text{dist}(C_1,C_2)=\min_{A_i\in C_1,A_j\in C_2}\text{dist}(A_i,A_j).$$

Supporting Information

Figure S1 Agreement of S_k with the standard normal. Shown are the distribution functions for the standard normal distribution (green line), and for (see eq (8)) $S_k=\sqrt{12/(k+1)}\sum_{i=0}^{k}(\Omega_i^*-1/2)$, $k=2$ (red line) and $k=0$ (blue line). The latter is uniform on $-1.73, 1.73$. Obviously, already for $k=2$ the agreement between the standard normal and S_k is quite good.
(EPS)

Figure S2 Average number of recombination events in neutral coalescent trees. (A) in dependence of sample size n ($4Nr=10$) and (B) of the scaled recombination rate $4Nr$ ($n=100$). Red: simulation results obtained from 1000 replicates of `ms` [43]. Shown are average (bullets) and standard deviation (whiskers). Black: theoretical value $E(n_r)=4Nra_{n-1}$.
(EPS)

Figure S3 Distance from sweep site to first recombination site. Given that the rate of the first recombination event adjacent to a selective sweep site is $r_l=a_{n-1}ct_f/2$ (in case of a neutral topology) or $r_u=nct_f/2$ (in case of a star phylogeny) the distance between the selected site and the 'first' recombination event is described by a Poisson process with rate $r_l x$ or $r_u x$. Shown is the probability that the Poissson variable is 0 (i.e., for a "recombination free zone") for r_l (upper curve) and r_u (lower curve).
(EPS)

Figure S4 Differences between tests (A) $T_2^{(\text{sum})}$ and (B) $T_2^{(\text{product})}$. Given a test is significant at level $\alpha=0.01$, the plots show the maximum (x-axis) and the minimum (y-axis) of the three terms ω_1^*, ω_2^* and ω_3^*, which enter into the sum and product in $T_2^{(\text{sum})}$ and $T_2^{(\text{product})}$, respectively. The sum- and product-tests may yield different results, because the summands are differently constrained (here (A), the maximum $\lesssim 0.4$) than the factors (here (B), the maximum may reach almost 1, but the minimum is smaller than in the sum-test).
(PDF)

Table S1 Power of $T^{(\text{sum})}$, $T^{(\text{product})}$ and $T^{(\text{dist})}$ in dependence of distance to selected site. Moderate selection strength.
(PDF)

Table S2 Power of $T^{(\text{sum})}$, $T^{(\text{product})}$ and $T^{(\text{dist})}$ in dependence of mutation rate θ.
(PDF)

Table S3 Power of $T^{(\text{sum})}$, $T^{(\text{product})}$ and $T^{(\text{dist})}$ in dependence of sample size n.
(PDF)

Table S4 Power of $T^{(\mathrm{sum})}$, $T^{(\mathrm{product})}$ and $T^{(\mathrm{dist})}$ in dependence of distance to selected site. Small sample size.
(PDF)

Table S5 Power of $T^{(\mathrm{sum})}$, $T^{(\mathrm{product})}$ and $T^{(\mathrm{dist})}$ in dependence of distance to selected site. Large sample size.
(PDF)

Table S6 Empirical false positive rate. Bottleneck model with varying onset τ of the bottleneck. Strength is fixed at $0.01N$. Significance levels α are based on theoretical formulae according to eqs (7) and (8).
(PDF)

Table S7 Empirical false positive rate. Bottleneck model with varying duration of the bottleneck. Severity (duration divided by strength) is fixed at 1. Significance levels α are based on theoretical formulae according to eqs (7) and (8).
(PDF)

Table S8 Empirical false positive rate. Population expansion with varying onset τ of the expansion. Expansion rate is fixed at 10.
(PDF)

Table S9 Empirical false positive rate. Population substructure with two sub-populations, split time $t=1$ in the past and sampling scheme $n_1=195$, $n_2=5$. Varying migration rate m per generation per $4N$ individuals. Significance levels α are based on theoretical formulae according to eqs (7) and (8).
(PDF)

Table S10 Empirical false positive rate. Population substructure with two sub-populations, split time $t=1$ in the past and sampling scheme $n_1=190$, $n_2=10$. Varying migration rate m per generation per $4N$ individuals. Significance levels α are based on theoretical formulae according to eqs (7) and (8).
(PDF)

Table S11 Empirical false positive rate. Population substructure with two sub-populations, split time $t=1$ in the past and sampling scheme $n_1=180$, $n_2=20$. Varying migration rate m per generation per $4N$ individuals. Significance levels α are based on theoretical formulae according to eqs (7) and (8).
(PDF)

Table S12 Empirical false positive rate. Population substructure with two sub-populations, split time $t=1$ in the past and sampling scheme $n_1=150$, $n_2=50$. Varying migration rate m per generation per $4N$ individuals. Significance levels α are based on theoretical formulae according to eqs (7) and (8).
(PDF)

Table S13 Empirical false positive rate. Mutation model with jumps of size 2. Varying probability p for a step of size 2. With probability $1-p$ the step size is 1. Significance levels α are based on theoretical formulae according to eqs (7) and (8).
(PDF)

Table S14 Empirical false positive rate. Mutation model with jumps of size 7. Varying probability p for a step of size 7. With probability $1-p$ the step size is 1. Significance levels α are based on theoretical formulae according to eqs (7) and (8).
(PDF)

Acknowledgments

We wish to thank R. Fürst for contributing the UPGMA clustering software for microsatellite alleles, A. Schlizio for help in preparing figures and T. Anderson for sharing the raw data of a set of microsatellites from a population of *P. falciparum* from Thailand. Further, we would like to thank M. Hasselmann for discussion and comments.

Author Contributions

Conceived and designed the experiments: HL TW. Performed the experiments: HL TW. Analyzed the data: HL TW. Contributed reagents/materials/analysis tools: HL TW. Wrote the paper: HL TW. Obtained microsatellite data from T. Anderson: 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, volume 7. Oxford University Press. pp. 1–44.
3. Wakeley J (2009) Coalescent theory – an introduction. Greenwood Village, Colorado: Roberts&Company.
4. Ewens W (1972) The sampling theory of selectively neutral alleles. Theor Popul Biol 3: 87–112.
5. Tajima F (1989) The effect of change in population size on DNA polymorphism. Genetics 123: 597–601.
6. Fay J, Wu C (2000) Hitchhiking under positive Darwinian selection. Genetics 155: 1405–1413.
7. Fu Y (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics 147: 915–925.
8. Ferretti L, Perez-Enciso M, Ramos-Onsins S (2010) Optimal neutrality tests based on the frequency spectrum. Genetics 186: 353–365.
9. Griffiths RC, Marjoram P (1996) Ancestral inference from samples of DNA sequences with recombination. J Comput Biol 3: 479–502.
10. Hudson R, Kaplan N (1988) The coalescent process in models with selection and recombination. Genetics 120: 831–840.
11. Kaplan NL, Hudson RR, Langley CH (1989) The "hitchhiking effect" revisited. Genetics 123: 887–899.
12. McVean G, Cardin N (2005) Approximating the coalescent with recombination. Philos Trans R Soc Lond B Biol Sci 360: 1387–1393.
13. Wiuf C, Hein J (1999) Recombination as a point process along sequences. Theor Popul Biol 55: 248–259.
14. Eriksson A, Mahjani B, Mehlig B (2009) Sequential markov coalescent algorithms for population models with demographic structure. Theor Popul Biol 76: 84–91.
15. Ferretti L, Disanto F, Wiehe T (2013) The effect of single recombination events on coalescent tree height and shape. PLoS ONE 8(4): e60123.
16. 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.
17. Walsh J, Lynch M (2013) Evolution and Selection of Quantitative Traits. Sinauer Associates. Available: http://nitro.biosci.arizona.edu/zbook/NewVolume-2/pdf.
18. Wedderburn JHM (1922) The functional equation g(x2) = 2 x + [g(x)]2. The Annals of Mathematics 24: pp. 121–140.
19. Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. Genetics 105: 437–460.
20. Hudson R, Kaplan N (1985) Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111: 147–164.
21. Zivkovic D, Wiehe T (2008) Second-order moments of segregating sites under variable population size. Genetics 180: 341–357.
22. Reich D, Cargill M, Bolk S, Ireland J, Sabeti P, et al. (2001) Linkage disequilibrium in the human genome. Nature 411: 199–204.
23. Barton N (2000) Genetic hitchhiking. Philos Trans R Soc Lond B Biol Sci 355: 1553–1562.
24. Sokal RR, Sneath PHA (1963) Principles of Numerical Taxonomy. New York: W. H. Freeman and Co.
25. Schug MD, Mackay TFC, Aquadro CF (1997) Low mutation rates of microsatellite loci in *Drosophila melanogaster*. Nature Genetics 15: 99–102.
26. Fernando Vazquez J, Perez T, Albornoz J, Dominguez A (2000) Estimation of microsatellite mutation rates in *Drosophila melanogaster*. Genet Res 76: 323–326.
27. Ellegren H (2004) Microsatellites: simple sequences with complex evolution. Nat Rev Genet 5: 435–445.
28. Wiehe T, Nolte V, Zivkovic D, Schlötterer C (2007) Identification of selective sweeps using a dynamically adjusted number of linked microsatellites. Genetics 175: 207–218.
29. Nair S, Nash D, Sudimack D, Jaidee A, Barends M, et al. (2007) Recurrent gene amplification and soft selective sweeps during evolution of multidrug resistance in malaria parasites. Mol Biol Evol 24: 562–573.

30. Griffiths RC (1984) Asymptotic line-of-descent distributions. J Math Biol 21: 67–75.
31. Berestycki N (2009) Recent progress in coalescent theory. Ensaios matemáticos. Sociedade Brasileira De Matemática.
32. Colless DH (1982) Review: [untitled]. Systematic Zoology 31: pp. 100–104.
33. Kirkpatrick M, Slatkin M (1993) Searching for evolutionary patterns in the shape of a phylogenetic tree. Evolution 47: pp. 1171–1181.
34. Blum MGB, Franois O, Janson S (2006) The mean, variance and limiting distribution of two statistics sensitive to phylogenetic tree balance. The Annals of Applied Probability 16: pp. 2195–2214.
35. Schlötterer C (2002) A microsatellite based multi-locus screen for the identification of local selective sweeps. Genetics 160: 753–763.
36. Cornuet J, Luikart G (1996) Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. Genetics 144: 2001–2014.
37. Schlötterer C, Kauer M, Dieringer D (2004) Allele excess at neutrally evolving microsatellites and the implications for tests of neutrality. Proc R Soc Lond B 271: 869–874.
38. Neher R, Hallatschek O (2013) Genealogies of rapidly adapting populations. Proc Natl Acad Sci U S A 110: 437–442.
39. Stolle E, Wilfert L, Schmid-Hempel R, Schmid-Hempel P, Kube M, et al. (2011) A second generation genetic map of the bumblebee bombus terrestris (linnaeus, 1758) reveals slow genome and chromosome evolution in the apidae. BMC Genomics 12: 48.
40. Behrens D, Huang Q, Gessner C, Rosenkranz P, Frey E, et al. (2011) Three QTL in the honey bee apis mellifera l. suppress reproduction of the parasitic mite varroa destructor. Ecol Evol 1: 451–458.
41. Kim Y, Stephan W (2002) Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 160: 765–777.
42. Li H, Stephan W (2006) Inferring the demographic history and rate of adaptive substitution in drosophila. PLoS Genet 2: e166.
43. Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337–338.
44. Feller W (1968) An Introduction to Probability Theory and Its Applications, volume 1. Wiley.