OPEN ACCESS Freely available online

PLOS ONE

The Yule Approximation for the Site Frequency Spectrum after a Selective Sweep

Sebastian Bossert*, Peter Pfaffelhuber

Department of Mathematical Stochastics, Albert-Ludwigs University, Freiburg, Germany

Abstract

In the area of evolutionary theory, a key question is which portions of the genome of a species are targets of natural selection. Genetic hitchhiking is a theoretical concept that has helped to identify various such targets in natural populations. In the presence of recombination, a severe reduction in sequence diversity is expected around a strongly beneficial allele. The site frequency spectrum is an important tool in genome scans for selection and is composed of the numbers $S_1,\ldots,S_{n-1}$, where S_k is the number of single nucleotide polymorphisms (SNPs) present in k from n individuals. Previous work has shown that both the number of low- and high-frequency variants are elevated relative to neutral evolution when a strongly beneficial allele fixes. Here, we follow a recent investigation of genetic hitchhiking using a marked Yule process to obtain an analytical prediction of the site frequency spectrum in a panmictic population at the time of fixation of a highly beneficial mutation. We combine standard results from the neutral case with the effects of a selective sweep. As simulations show, the resulting formula produces predictions that are more accurate than previous approaches for the whole frequency spectrum. In particular, the formula correctly predicts the elevation of low- and high-frequency variants and is significantly more accurate than previously derived formulas for intermediate frequency variants.

Citation: Bossert S, Pfaffelhuber P (2013) The Yule Approximation for the Site Frequency Spectrum after a Selective Sweep. PLoS ONE 8(12): e81738. doi:10.1371/journal.pone.0081738

Editor: William J. Etges, University of Arkansas, United States of America

Received July 26, 2013; **Accepted** October 15, 2013; **Published** December 10, 2013

Funding: This work was funded by the project PP672/3-1 and Hu1889/1-1 of the Deutsche Forschungsgemeinschaft (DFG). 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: sebastian.bossert@stochastik.uni-freiburg.de

Introduction

Genetic hitchhiking is the cause of a severe reduction of sequence diversity in a population due to recent strong positive selection [1]. Several statistical methods are available to detect these selective sweeps. The most successful approaches include various aspects of the available data, such as the site frequency spectrum and linkage disequilibrium patterns. See e.g., [2] for a framework using a likelihood ratio test using the site frequency spectrum, [3], [4] for tests based on linkage disequilibrium and [5], who use a combination of both. The most challenging issue today is to dissect population demography from signatures of selection.

One of the most successful approaches for detecting selective sweeps is called *SweepFinder*. Here, the site frequency spectrum for a selective and a neutral model is compared for each SNP available in the data [6]. This approach highlights the necessity of making analytical predictions for site frequency spectra under strong positive selection, which is the main goal of the current manuscript. While *SweepFinder* uses a selective model with the star-like method (see e.g., [7]), here, we use a refined model.

Current theoretical investigations and predictions of the signature of strong positive selection are mostly based on a genealogical perspective. The resulting genealogy is termed coalescent in a random background and was studied by [8] and [9]. The simplest approximation for large selection coefficients is the star-like approximation from [10] and [7]. The star-like approximation assumes that all individuals from a sample taken at the time of fixation are direct descendants of the founder of the selective sweep. In addition, recombination events may have split the history of the target of selection from a linked neutral variant. [7], [11], and [12] used a marked Yule process, which has been shown to be a finer approximation by [7]. Rather than using a star-like approximation of the genealogy at the target of selection, [12] used the idea put forward by [13], which states that in the early phase of a selective sweep, the beneficial allele behaves similarly to a supercritical branching process. As a consequence, the genealogy also resembles a supercritical branching process, which turns out to be a Yule process [14].

In this manuscript, we go beyond approximating the genealogy by a marked Yule process and provide an analytical expression for the site frequency spectrum after a selective sweep. Two features of the spectrum are the most important for data analysis: an excess of singletons (which might also arise due to population expansion) and an excess of high-frequency variants (which appear to be a unique feature of sweeps; [15]). [16] already gave an approximation of the site frequency spectrum and used the excess of high frequency variants to develop a statistical test for positive selection. Using our analytical approximations, we will see that such classical approaches slightly overestimate the number of high-frequency variants, while our Yule-approximation is more accurate. In addition, intermediate-frequency variants are predicted accurately only by the marked Yule-approximation. These features of the Yule-approximation can be used to construct conservative tests for selective sweeps.

Model and Results

Consider a (diploid) population of size N which evolves under the neutral Wright-Fisher model. We will study two loci (called A- and B-locus) within this population, which recombine with probability r per generation. (We neglect recombination within loci.) At the A-locus, the population is fixed for the wild-type a before time $t=0$. The B-locus is modeled using an infinite sites model of mutation with mutation probability μ per generation (see [17]). At time $t=0$, a beneficial mutation A with fitness $1+s$ appears at the A-locus and is conditioned on eventual fixation in the whole population. Our main interest is the site frequency spectrum of the B-locus at the fixation time T of the A-allele, which we also refer to as the end of the sweep. Consider a sample of size n taken at time T, and let S_i be the number of SNPs at the B-locus where the derived variant is present in exactly i individuals. The time before $t=0$ is called the *neutral phase*, while the time between $t=0$ and $t=T$ is the *selective phase*.

Diffusion approximation and structured coalescent

To derive an approximation of the expected site frequency spectrum, we rely on a diffusion approximation for the frequency of the beneficial A-allele (see e.g., [18]) and a coalescent process in a random background as described in [9] (see also [8]). Recall (e.g., from [19]) that the frequency of the A-allele after $t=0$, when time is rescaled by a factor of $2N$, is approximately given by the solution $\mathcal{Y}=(Y_t)_{t\geq 0}$ of the stochastic differential equation

$$dY=\alpha Y(1-Y)\coth(\alpha Y)dt+\sqrt{Y(1-Y)}dW, \qquad Y_0=0, \quad (1)$$

where $\alpha=2Ns$ is the rescaled (genic) selection intensity, and s is defined by saying that $(1+s)x/(1+sx)$ is the expected number of A-alleles in the next generation if the current frequency is x. Observe that $Y_t=1$ after some random time T, which we call the fixation time of A. In the background of the path $\mathcal{Y}$, we consider a structured coalescent that evolves as follows (see Figure 1 for an illustration, where a sample of size $n=9$ is used): Set $\beta=T-t$ and start with n lines at time $\beta=0$ (i.e., $t=T$ and the end of the sweep) in the A-background. The following four transitions can occur between times $\beta=0$ and $\beta=T$, i.e., during the selective phase:

1. Coalescence of a pair of lines in the A-background: at rate $1/Y_\beta$, any pair of lines in the A-background coalesces.
2. Switching of background from A to a by recombination: at rate $\rho(1-Y_\beta)$ with $\rho:=2Nr$ (r is the recombination fraction between the selective and neutral locus within a single generation), any line in the A-background changes to the a-background.
3. Coalescence of a pair of lines in the a-background: at rate $1/(1-Y_\beta)$, any pair of lines in the a-background coalesces.
4. Switching of background from a to A by recombination: at rate ρY_β, any line in the a-background changes to the A-background.

Due to these transitions, there is a random number K_a of lines in the a-background at time $\beta=T$ and $K_A\in\{0,1\}$ lines in the A-background. (If there was two or more lines in the A-background, their coalescence rate would have been arbitrarily large by the coalescence rate $1/Y_\beta$.) The resulting $K:=K_a+K_A$ lines follow a standard neutral coalescent after time $\beta=T$, i.e., every pair of lines coalesces at rate 1 after only a single line is left and the process is stopped.

After having constructed the random tree from the coalescing lines, every line is hit by mutation events at the rate $\theta/2$, with $\theta:=4N\mu$. We call an event a *mutation of size i* if it falls on a branch leading to exactly i leaves of the tree. The number of size i mutations is called S_i, and $S_1,\ldots,S_{n-1}$ is called the site frequency spectrum, which we will approximate for large α below.

Yule approximation of the genealogy in the selective phase

In [19] and [11], the following approximation of the structured coalescent during the selective phase was developed with the limits of large α and for $\rho\ll\alpha$: As was shown, events 3. and 4. from the structured coalescent can be ignored because their probability becomes small for large α. Thus, each line undergoes at most one recombination event during the selective phase. Two lines of the

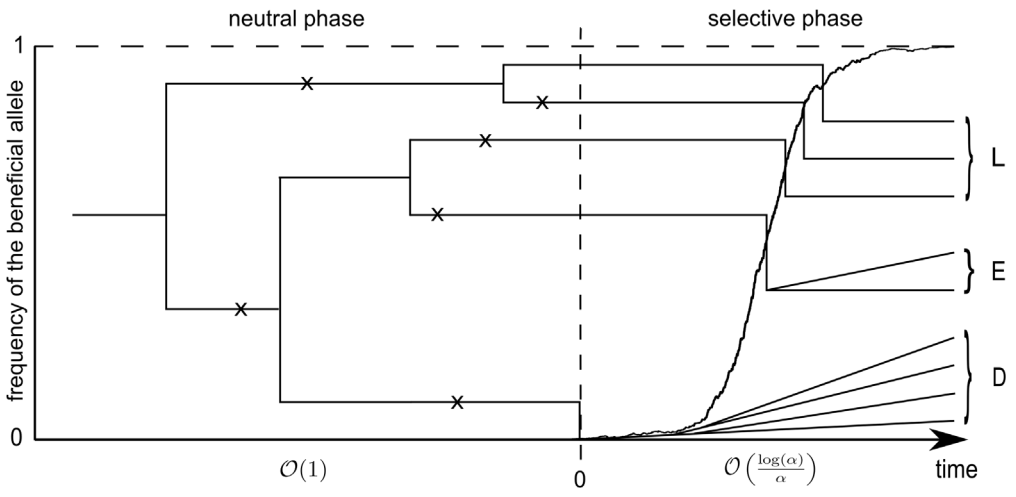

Figure 1. The structured coalescent. In the given example of the structured coalescent, we see on the right side the selective phase with a sample of size 9 at the moment of fixation and the frequency development of the beneficial allele. At time 0, there are 3 late recombinant families (labeled with L), which all have a size of 1, one early recombinant family (labeled with E) of size 2 and one nonrecombinant family (labeled with D) of size 4. These lines then start a standard coalescent in the neutral phase. The crosses illustrate SNPs in the sample.
doi:10.1371/journal.pone.0081738.g001

genealogy at time $\beta = T$ belong to the same family if they coalesce between time 0 and T. The following families are distinguished:

1. *Nonrecombinant family*: The set of individuals whose ancestral lineages never left background A.
2. *Early recombinant families*: The set of individuals whose ancestral lines have not left background A before (according to the backward time β) the first coalescence in the sample occurs, but the ancestor at time $\beta = T$ (equivalent to $t = 0$) is in background a.
3. *Late recombinant families:* The families consisting of a single individual whose ancestral line has left background A before the first coalescence in the sample, and the ancestor at time $\beta = T$ is in background a.

Note that late recombinant families are of size 1 by definition, and there can be at most one nonrecombinant family that has inherited their B-allele from the founder of the sweep.

To get an approximation formula for the genealogy at time $\beta = T$, we first need the distribution for the number and size of the different families. Recall from Theorem 1 in [19] that the genealogy consists (up to an error of probability of order ρ^2/α^2) of

- L *late recombinant* families of size 1,
- one *early recombinant* family of size E and
- one *nonrecombinant* family of size $n - L - E$.

For the joint distribution of L and E, define a random variable F, distributed according to

$$\mathbb{P}(F \le i) = \frac{(i-(n-1)) \cdots (i-1)}{(i+(n-1)) \cdots (i+1)}. \tag{2}$$

Given $F = f$, L is a binomial random variable with n trials and success probability $1 - p_f$, where

$$p_f = \exp\left(-\frac{\rho}{\alpha} \sum_{i=f}^{\lfloor 2\alpha \rfloor} \frac{1}{i} \right). \tag{3}$$

The distribution of E depends on L and on another variable S, which gives the number of lines that are affected by the early recombination at time F according to

$$\mathbb{P}(S = s) = \begin{cases} \frac{\rho n}{\alpha} \sum_{i=2}^{n-1} \frac{1}{i} & \text{for} \quad s = 1 \\ \frac{\rho n}{\alpha} \frac{1}{s(s-1)} & \text{for} \quad 2 \le s \le n-1 \\ \frac{\rho n}{\alpha} \frac{1}{n-1} & \text{for} \quad s = n. \end{cases} \tag{4}$$

(Note that the case $\frac{\rho n}{\alpha} \sum_{i=1}^{n-1} \frac{1}{i} > 1$ requires a different definition of the distribution of S, which we give in Section A of the SI.) As one or more of these S lines could experience a late recombination event, they could be kicked out of the family of early recombinants. This explains the hypergeometric distribution of E, i.e., given $S = s$ and $L = l$, the variable E is hypergeometric with

$$\mathbb{P}(E = e) = \frac{\binom{s}{e}\binom{n-s}{n-l-e}}{\binom{n}{n-l}}. \tag{5}$$

Combining these equations, a straightforward calculation (see Corollary 2.7 in [19]) leads to

$$\mathbb{P}(E = e,\, L = l) = \mathbb{E}[p_F^{n-l}(1-p_F)^l] \cdot \begin{cases} \frac{\rho n}{\alpha} \frac{(n-1)\binom{n-2}{e-2} 1\!\!1_{l+e=n} + \binom{n-1}{l}}{e(e-1)} & e \ge 2 \\ \frac{\rho n}{\alpha} \left(1\!\!1_{l+1=n} + \binom{n-1}{l} \sum_{i=2}^{n-1} \frac{1}{i} + \sum_{s=2}^{n} \frac{\binom{n-s}{l-s+1}}{s-1} \right) & e = 1 \\ \binom{n}{l}\left(1 - \frac{\rho n}{\alpha}\left(\sum_{i=1}^{n-1}\frac{1}{i} - \frac{l}{n}\sum_{i=2}^{n-1}\frac{1}{i}\right)\right) & \\ \quad + \frac{\rho n}{\alpha}\left(\frac{1}{n} 1\!\!1_{l=n} + \sum_{s=2}^{n}\binom{n-s}{n-l}\frac{1}{s(s-1)}\right) & e = 0. \end{cases} \tag{6}$$

Note that this equation corrects an error (in the case of $e = 0$) of the equation of [19]; see SI, Sections A and B. Moreover, there is a factor of 2 difference here because we assume a diffusion constant of 1 in (1).

Yule approximation of the site frequency spectrum

Our goal is to obtain an expectation of the site frequency spectrum, $\mathbb{E}[S_i]$, at the end of a selective sweep using the approximation from (6). We will assume that α is large and that no new mutations accumulate in the sample during the selective phase. Moreover, recombination between the A- and B-locus has to be in a certain range to see a non-trivial frequency spectrum. (Here, trivial would either mean that there is no variation at all if ρ is too small or a neutral site frequency spectrum if ρ is too large.) Recalling that the duration of the sweep is approximately $(2\log\alpha)/\alpha$ (see [19]), ρ must be on the order of $\alpha/\log(\alpha)$. In other words, ρ/α is on the order of $1/\log(\alpha)$ and hence small if α is large.

To get an approximation formula for the frequency spectra, the events and probabilities of the selective phase must be joined with the neutral phase. In the neutral phase, Kingman's coalescent describes the genealogy of the $K = K_a + K_A$ remaining lines. The crucial point is how to combine the approximation of the genealogy of the A-locus during the selective phase with a neutral coalescent before the onset of the sweep. A critical quantity is the number K of ancestors of the sample at the onset of the sweep. Because a mutation can only influence at most $K-1$ of these ancestors, the descendants in the selective phase depend on this number of lines. Recall that the sample size is n, $\alpha = 2Ns$ is large, $\theta/2 = 2N\mu$ is the mutation rate and $\rho = 2Nr$ is the recombination rate, with ρ/α being small. Therefore, the expected number of mutations of size i is (see SI, Section C for the proof)

$$\mathbb{E}[S_i]=\sum_{l=1}^{n}\mathbb{P}(E=0,L=l)\left(\frac{(l+1-i)\theta}{(l+1)i}1\!\!1_{l\geq i}+\frac{\theta}{l+1}1\!\!1_{n-l\leq i}\right)$$
$$+\sum_{l=0}^{n-1}\mathbb{P}(E=1,L=l)\left(\frac{(l+2-i)\theta}{(l+2)i}1\!\!1_{l+1\geq i}+\frac{\theta}{l+2}1\!\!1_{n-l-1\leq i}\right)$$
$$+\sum_{s=2}^{n-1}\sum_{l=0}^{s-2}\mathbb{P}(E=s-l,L=l)\cdot\theta\cdot$$
$$\left[\frac{(l+2-i)(l+1-i)}{(l+1)(l+2)i}1\!\!1_{l\geq i}\quad+\frac{(l+1-i+n-s)}{(l+1)(l+2)}1\!\!1_{n-s\leq i}1\!\!1_{l+n-s\geq i}\right.$$
$$\left.+\frac{(1-i+s)}{(l+1)(l+2)}1\!\!1_{s-l\leq i}1\!\!1_{s\geq i}+\frac{(i+l-n+1)}{(l+1)(l+2)}1\!\!1_{n-l\leq i}\right]$$
$$+\sum_{l=1}^{n-2}\mathbb{P}(E=n-l,L=l)\left(\frac{(l+1-i)\theta}{(l+1)i}1\!\!1_{l\geq i}+\frac{\theta}{l+1}1\!\!1_{n-l\leq i}\right)+\mathcal{O}\left(\frac{\rho^2}{\alpha^2}\right),\quad(7)$$

for $1\leq i\leq n-1$, where the probabilities of $\mathbb{P}(E=e,L=l)$ are given by (6). We note that the term $\mathcal{O}(\frac{\rho^2}{\alpha^2})$ is due to the use of the approximation formula for the selective phase.

To get an idea of how this formula is computed, consider again Figure 1. There are 3 late recombinant families, one early recombinant family of size 2 and one nonrecombinant family (labeled D) of size 4. Given these values, there are two different ways for a mutation to get to a size of 2. Either it had a size of 2 at time $t=0$ and these two lines were two late recombinant families, or it had size 1 at time $t=0$ and then was the founder of the early recombinant family, which has a size of 2 at the end of the sweep. Taking into account all possibilities, (7) arises.

Previous approximation formulas

Using simulations, we compared the Yule approximation formula (7) to two other approximation formulas for the frequency spectra. The first approximation is from [16] and will be called the *deterministic formula* because a deterministic development of the frequency of allele A is assumed in this approach. The second approximation is the *star-like approximation* (see [7] or chapter 6 in [20]).

Deterministic approximation

In [16], Fay and Wu obtained the following approximation for the site frequency spectrum after a selective sweep, building on the ideas of [1]. They obtain

$$\mathbb{E}[S_k]=\int_0^1\binom{n}{k}x^k(1-x)^{n-k}\phi(x)dx \qquad (8)$$

with

$$\phi(x)=\theta\left(\frac{1}{x}-\frac{1}{\tilde{r}}\right)1\!\!1_{[0,\tilde{r}]}(x)+\frac{\theta}{\tilde{r}}1\!\!1_{[1-\tilde{r},1]}(x),\qquad \tilde{r}:=\frac{r}{s}\log\left(\frac{1}{p_0}\right)$$

where p_0 is the starting frequency of the beneficial allele. For the numerical comparison, we use $p_0=\frac{1}{\alpha}$ because, in this situation, the length of the selective phase is $2\log(\alpha)/\alpha$, which is close to the expectation of the stochastic model.

Star-like approximation

For the classical star-like approximation, every line in the selective phase has the same independent chance to recombine and be in background a at time $t=0$. Therefore, $E=0$, and L is binomially distributed with parameters n and $1-e^{-\rho\frac{\log\alpha}{\alpha}}$, which is the probability that a single line recombines. Combining this insight with (7) leads to the equation

$$\mathbb{E}[S_i]=\sum_{l=1}^{n}\binom{n}{l}(1-e^{-\rho\log\alpha/\alpha})^l(e^{-\rho\log\alpha/\alpha})^{n-l}$$
$$\left(\frac{(l+1-i)\theta}{(l+1)i}1\!\!1_{l\geq i}+\frac{\theta}{l+1}1\!\!1_{n-l\leq i}\right)+\mathcal{O}\left(\frac{\rho}{\alpha}\right).\qquad(9)$$

Note that for small ρ/α, the approximation error is much larger than in (7).

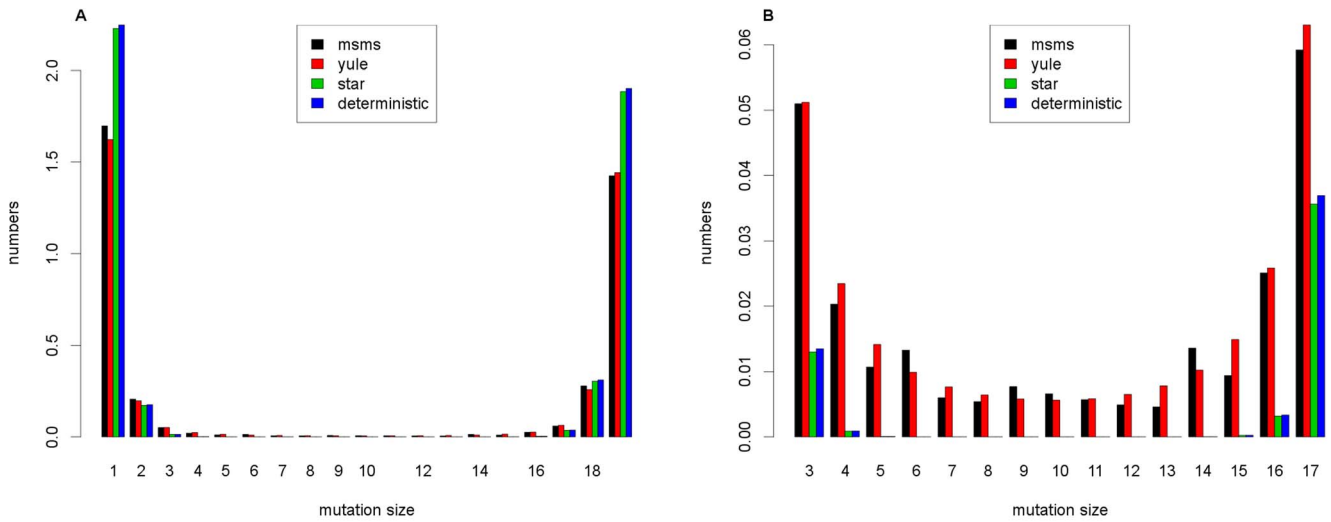

Figure 2. Comparison of the expected frequency spectra I. Comparison of the 3 approximation formulas and the results from msms for the parameters $2N=2000000$, $n=20$, $\theta=10$, $r=0.00005$ and $s=0.02$. In **A**, the whole frequency spectrum is illustrated, while in **B**, the number of the mutation sizes between 3 and 17 are enlarged.
doi:10.1371/journal.pone.0081738.g002

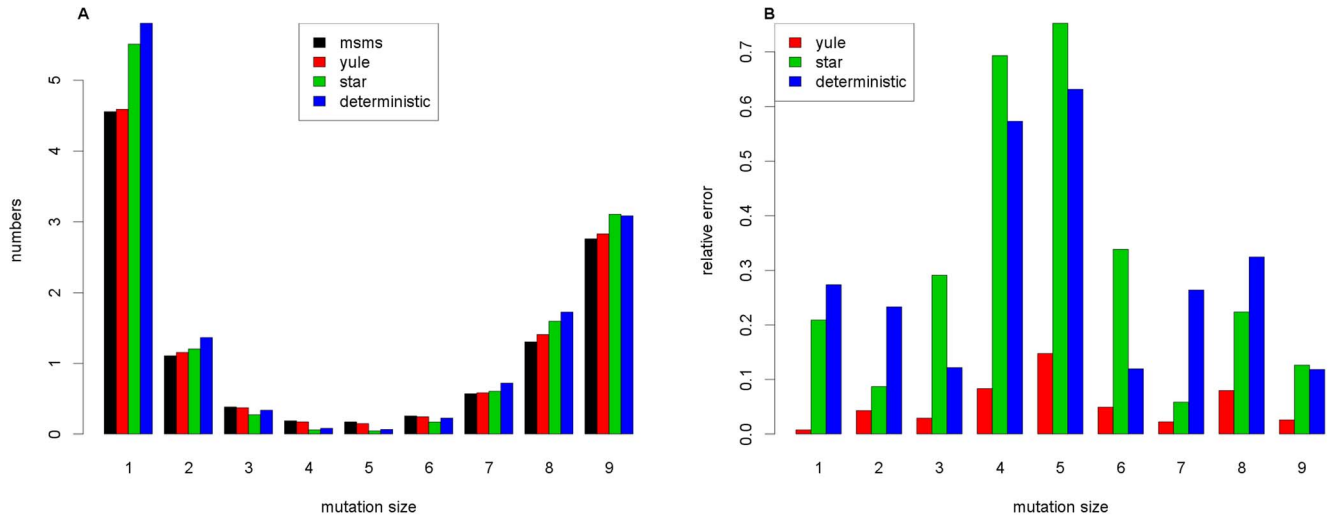

Figure 3. Comparison of the expected frequency spectra II. Comparison of the 3 approximation formulas and the results from msms for the parameters $2N = 2000000$, $n = 10$, $\theta = 10$, $r = 0.0002$ and $s = 0.01$. In **A**, we see the expected frequency spectra, and in **B**, we see the relative errors compared to the reference *msms*.
doi:10.1371/journal.pone.0081738.g003

Numerical comparison

Our goal is to compare the performance of the Formulas (7), (8) and (9) to simulations from the Wright-Fisher model. For the Wright-Fisher model, the simulation tool msms was used (which stands for *make sample mit selection*, see [21] or http://www.mabs.at/ewing/msms/index.shtml). To compare the different formulas for the expected frequency spectra, the average of 10^5 iterations was taken as a reference. Figure 2 shows the case of a high selective advantage $s = 0.02$ in a sample of size $n = 20$. Theoretically, the Yule and star-like approximations converge for large $\alpha = 2Ns$. However, while the deterministic and star-like approximations perform about equally well, the (absolute and relative) error of (7) is smaller.

In Figure 3, we used a smaller selective coefficient $s = 0.01$ and a sample of size $n = 10$. Here, the relative error of the star-like and deterministic approximations exceed 0.6. Again, the Yule approximation (7) gives the best results, with the relative error never exceeding 0.2. Reassuringly, all approximations give good results for low- and high-frequency variants that are known to be fundamental in detecting selecting sweeps in data.

In applications, the case of a high recombination rate is of particular importance. Here, (7) needs to be corrected as described in Appendix A. Because the error of all approximation formulas increases with recombination rate, it is no surprise that the errors in Figure 4 are larger than those in Figures 2 and 3. Still, the Yule approximation works best for most of the frequency classes.

Discussion

The site frequency spectrum is a basic summary statistic used for the analysis of SNP data. Theoretical predictions of the shape of

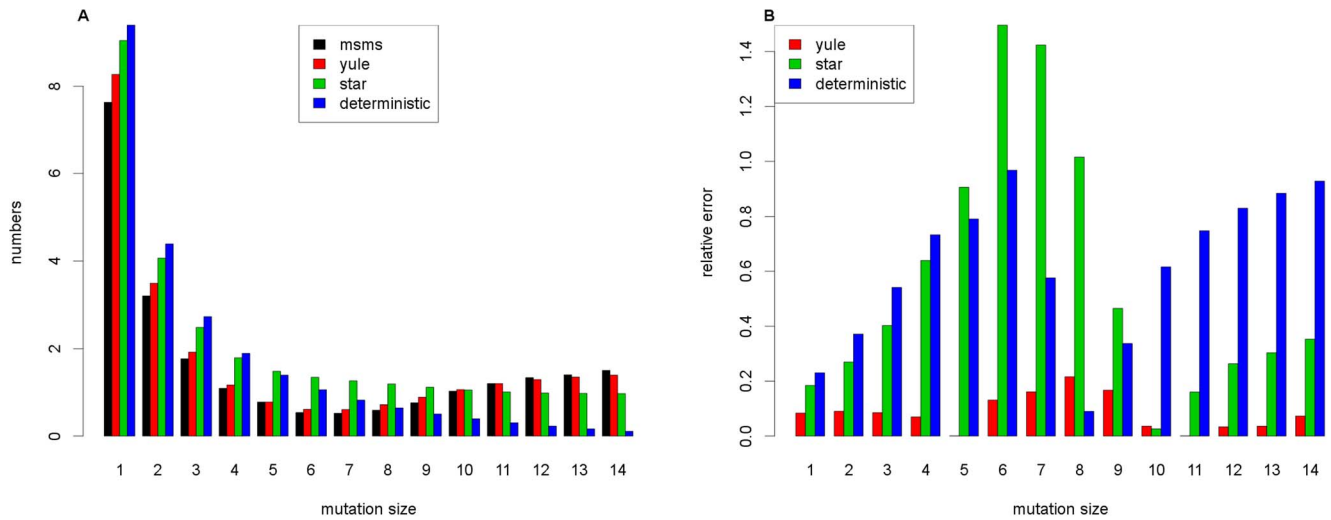

Figure 4. Comparison of the expected frequency spectra III. Comparison for the parameters $2N = 1000000$, $n = 15$, $\theta = 10$, $r = 0.003$ and $s = 0.03$, where the adjusted formula for the joint distribution according to Appendix A is needed. In **A**, the expected frequency spectra are depicted, and in **B**, the relative errors compared to the reference *msms* are illustrated.
doi:10.1371/journal.pone.0081738.g004

the frequency spectrum are most important in order to understand the evolutionary forces that have shaped the genomic data at hand. In the present paper, we have demonstrated how a recently developed approximation for selective sweeps from [7], [19], [11], [12], based on a marked Yule process, leads to such a prediction (at least for the expected site frequency spectrum). For the analytical formula, two cases have to be taken into account. If $d := \frac{\rho n}{\alpha} \sum_{i=1}^{n-1} \frac{1}{i} < 1$, the marked Yule process can be applied directly, but if $d > 1$, we have to use some normalization procedure. The latter case arises if the neutral locus has a large recombinational distance to the target of selection. In the parameter constellation of Figure 4, neither of the approximations works particularly well, with relative errors up to 20% for the Yule and deterministic approximations and over 140% for the star-like approximation. However, theoretical predictions become worse for larger ρ/α and errors are less predicable in this setting.

For smaller recombinational distances, we find that the Yule approximation outperforms the star-like approximation, especially for intermediate frequency variants (relative error up to 20% for the Yule approximation versus up to 80% for the star-like approximation, see Figure 3). In a comparison between the Yule and star-like approximations, a basic difference is that the star-like approximation forbids what we called *early recombinant families*. Such families lead to a decrease in the number of singleton mutations, which is shown in our simulations and has the greatest impact on the relative errors we reported above.

Altogether, the combination of (7) and (11) gives our analytical formula. Most importantly, compared to other approaches, such as the deterministic approach of [16] and the star-like approximation derived in [10], [7] and used e.g., in [3], the Yule process approximation has a smaller error in nearly all cases. Although the formulas derived in the Yule approximation are more involved, they can still be easily implemented for data applications to obtain a higher accuracy. Above all, such accuracy is desirable in genome scans for selective sweeps, which are frequently carried out by software such as *SweepFinder* [6].

Supporting Information

Appendix S1
Supporting Information for the article.
(PDF)

Acknowledgments

We thank Joachim Hermisson for fruitful discussions and two anonymous referees for their helpful comments.

Author Contributions

Conceived and designed the experiments: PP. Wrote the paper: SB PP. Made the simulations: SB.

References

1. Smith JM, Haigh J (1974) The hitch-hiking effect of a favorable gene. Genetic Research 23: 23–35.
2. Kim Y, Stephan W (2002) Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 160: 765–777.
3. Kim Y, Nielsen R (2004) Linkage disequilibrium as a signature of selective sweeps. Genetics 167: 1513–1524.
4. Jensen JD, Thornton KR, Bustamante CD, Aquadro CF (2007) On the utility of linkage disequilibrium as a statistic for identifying targets of positive selection in nonequilibrium populations. Genetics 176: 2371–2379.
5. Pavlidis P, Jensen JD, Stephan W (2010) Searching for footprints of positive selection in whole-genome snp data from nonequilibrium populations. Genetics 185: 907–922.
6. Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, et al. (2005) Genomic scans for selective sweeps using snp data. Genome Research 15: 1566–1575.
7. Durrett R, Schweinsberg J (2004) Approximating selective sweeps. Theoretical Population Biology 66: 129 – 138.
8. Kaplan NL, Hudson RR, Langley CH (1989) The 'hitchhiking effect' revisited. Genetics 123: 887–899.
9. Barton NH, Etheridge AM, Sturm AK (2004) Coalescence in a random background. Ann Appl Probab 14: 754–785.
10. Barton NH (1998) The effect of hitch-hiking on neutral genealogies. Genetic Research 72: 123–133.
11. Pfaffelhuber P, Haubold B, Wakolbinger A (2006) Approximate genealogies under genetic hitch-hiking. Genetics 174: 1995–2008.
12. Pfaffelhuber P, Studeny A (2007) Approximating genealogies for partially linked neutral loci under a selective sweep. J Math Biol 55: 299–330.
13. Fisher R (1930) The Genetical Theory of Natural Selection. Second edition. Oxford: Clarendon Press.
14. Evans S, O'Connell N (1994) Weighted occupation time for branching particle systems and a representation for the supercritical superprocess. Canad Math Bull 37: 187–196.
15. Stephan W (2010) Genetic hitchhiking versus background selection: the controversy and its implications. Phil Trans R Soc B 365: 1245–1253.
16. Fay JC, Wu CI (2000) Hitchhiking under positive darwinian selection. Genetics 155: 1405–1413.
17. Kimura M (1969) The number of heterozygous nucleotide sites maintained in a finite population due to steady ux of mutations. Genetics 61: 893–903.
18. Ewens WJ (2004) Mathematical population genetics. I, volume 27 of *Interdisciplinary Applied Mathematics*. New York: Springer-Verlag, second edition. Theoretical introduction.
19. Etheridge A, Pfaffelhuber P,Wakolbinger A (2006) An approximate sampling formula under genetic hitchhiking. Ann Appl Probab 16: 685–729.
20. Durrett R (2008) Probability models for DNA sequence evolution. Probability and its Applications (New York). New York: Springer, second edition.
21. Ewing G, Hermisson J (2010) Msms: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics 26: 2064–2065.