ELSEVIER

Stochastic Processes and their Applications 93 (2001) 1–24

stochastic processes and their applications

www.elsevier.com/locate/spa

On the quantity and quality of single nucleotide polymorphisms in the human genome

Richard Durrett [*,1], Vlada Limic [2]

Department of Mathematics, Cornell University, Ithaca NY 14853, USA

Received 24 April 2000; received in revised form 4 October 2000; accepted 20 October 2000

Abstract

Single nucleotide polymorphisms (SNPs) are useful markers for locating genes since they occur throughout the human genome and thousands can be scored at once using DNA microarrays. Here, we use branching processes and coalescent theory to show that if one uses Kruglyak's (Nature Gen. 12 (1999) 139–144) model of the growth of the human population and one assumes an average mutation rate of 1×10^{-8} per nucleotide per generation then there are about 5.7 million SNP's in the human genome, or one every 526 base pairs. We also obtain results for the number of SNPs that will be found in samples of sizes $n\geqslant2$ to gain insight into the number that will be found by various experimental procedures.

1. Introduction

The information in DNA is encoded in a sequence of nucleotides, four chemicals that are usually referred to by the first letters of their names: A, C, G, and T. Single nucleotide polymorphisms (SNPs) are as the name suggests, single nucleotides in a genome that are polymorphic, i.e., in which each allele has a frequency of less than 99% in the population as a whole. SNPs are of interest as genetic markers for locating genes. To look for a gene that causes a disease like type I diabetes, one would take a sample of several hundred individuals with and without the disease and then look for a correlation (in genetics this is called linkage disequilibrium) between the disease state of individual and the state of these markers. A significant correlation in one region of the genome would then suggest that it contains the disease causing gene and further sequencing efforts would be concentrated there.

Given that the human genome consists of about 3 billion nucleotides, it is clear that this strategy will require a large number of SNPs, but it is an important question to determine the order of magnitude of the number required. Current technology allows DNA microarrays to be constructed so that state of thousands, or perhaps tens of

* Corresponding author. Fax: +1-607-255-8282.

E-mail address: rtdl@cornell.edu (R. Durrett).

[1] Partially supported by NSF grant DMS 98777066 through the Probability program and a supplement to C.F. Aquadro's NIH grant GM36431 through the program for the study of Complex Biological Systems.

[2] Partially supported by an NSF Postdoctoral Fellowship.

0304-4149/01/$ - see front matter
PII: S0304-4149(00)00090-9

thousands of SNPs can be determined in a single procedure, (Landegren et al., 1998) but is that enough? In a recent article in *Nature Genetics*, Kruglyak (1999) used simulations based on the coalescent to suggest the range at which SNPs have useful levels of linkage disequilibrium (i.e., correlation) are unlikely to extend beyond 3 kilobases (kb) in the general population. Since the human genome consists of 3 billion bases, this means that even if SNPs are evenly spaced, approximately 500,000 SNP's are needed, a very depressing number for makers of DNA chips.

The purpose of this article is to do a mathematical analysis of two related questions: "How many SNPs are there in the human genome?" and "What percentage of these will a sample of size n find?" The second question in the case $n = 6$ is related to Celera's original strategy for sequencing the human genome, as described by a lecture of Gene Myers at a Cornell Theory Center Symposium on October 14, 1999. At that point, he said that they would achieve "10 times coverage", i.e., each nucleotide will be sequenced on the average 10 times. To make the assembly process easier, one individual will be used for the first 6 times coverage of the human genome. Since humans are diploid organisms (i.e., have two copies of the genetic information) this will with probability $\frac{31}{32}$ lead to a sample of size 2 at each site. To find more SNPs, Celera will then use multiple individuals for the final 4 times coverage. This gives an average of a little less than four samples per nucleotide, making a total of about 6.

The exact details of Celera's strategy are not important here. We will derive results that are valid for all $n \geqslant 2$, since they allow us to make predictions about the results of other sampling strategies. For example, $n = 20$ is related to experimental work of Wang et al. (1998) and $n = 5$ is related to the strategy Celera ended up using. To answer the questions posed above, we need a model of the growth of the human population. Following Kruglyak (1999) we will assume that humans had a constant population size of 10,000 individuals until 5000 generations ago and then expanded at a constant exponential rate to its present day size of 5 billion. Solving the equation

$$\mu^{5000} = (5 \times 10^9)/(10{,}000) = 500{,}000, \tag{1.1}$$

we find $\mu = 1.00263$. For those who might complain that the current population is actually 6 billion, we note that this changes the answer to 1.00266. Another possible objection is that according to the World Book encyclopedia, "the world's population grew slowly before AD1, then almost doubled by the year 1000. At its present rate of growth the world's population doubles every 41 years". Taking the estimated world population of 138 million in AD1, using a human generation time of 20 years leads to the new equation

$$\mu^{4900} = (138 \times 10^6)/(10{,}000) = 13{,}800, \tag{1.2}$$

which solves to give $\mu = 1.00194$. Since all three computations lead to roughly the same growth rate, we will choose $\mu = 1.0026$ to keep the closest connection with Kruglyak's work.

For convenience, we will let $T = 5000$ and index generations by integers $m \leqslant T$ so that the expansion began at time $m = 0$. Recalling that humans are diploid organisms,

the number of copies of the nucleotide under consideration in generation m is then

$$N_m = \begin{cases} 20{,}000 & \text{for } m \leqslant 0, \\ 20{,}000\mu^m & \text{for } m \geqslant 0. \end{cases} \tag{1.3}$$

Kruglyak builds his genealogical relationships by working backwards in time and using the discrete time coalescent with a varying population size (see e.g., Griffiths and Tavaré, 1994). In words, each of the N_m nucleotides in generation m picks its parent uniformly from the possible choices in generation N_{m-1}. Of course, all of the choices at one time are made independent of each other and independent of what has already been done at times $T, T-1, \ldots, m+1$.

In addition to working backwards in time, we will find it convenient to work forward from time 0, using a branching process in which each individual in generation t gives birth to an independent number of children in generation $t+1$ with mean μ. To see what distribution to take for the number of children, we note that in the coalescent a given nucleotide in generation t will be chosen with probability $1/N_t$ by each of the N_{t+1} nucleotides in generation $t+1$, and N_t is large, so the number of descendants will have roughly a Poisson distribution with mean $\mu = N_{t+1}/N_t$. For readers who might complain that observed human family size distributions are not Poisson, we note that (i) this choice of distribution is needed to keep a close connection with the coalescent, and (ii) the assumption can easily be dropped. As we will indicate in Section 2, the answers depend only on the first two moments of the number of offspring X, $\mu = EX$ and $a = EX(X-1)/2$.

While working forward in time we are only interested in individual nucleotides that have offspring alive at the present, time $T = 5000$. It is a well-known fact in the theory of branching processes (this and other "well-known" facts can be found in Chapter 1 of Athreya and Ney (1972)) that if we let p_k be the probability of k children and define the generating function $\phi(\theta) = \sum_{k=0}^{\infty} p_k \theta^k$ then the probability a family has died out by generation k is $\sigma_k = \phi^k(0)$, and if $\mu > 1$ then as $k \to \infty$, σ_k converges to σ, the unique solution of $\phi(\sigma) = \sigma$ that lies in $[0,1)$. In the case of interest here, the Poisson distribution with mean μ has generating function

$$\phi(\theta) = \sum_{k=0}^{\infty} \mathrm{e}^{-\mu} \frac{\mu^k}{k!} \theta^k = \exp(-\mu(1-\theta)), \tag{1.4}$$

so the fixed point equation is $\sigma = \exp(-\mu(1-\sigma))$.

For a given value of μ Eq. (1.4) can only be solved numerically. However, our μ is close to 1, so expanding ϕ to second order in a Taylor series around 1, we can solve it approximately with the simple result that the survival probability $\rho = 1 - \sigma$ has

$$\rho \approx \frac{\mu - 1}{a} = 0.0052. \tag{1.5}$$

With a little help from a computer one can find that $\rho = 0.00518203$.

Let Z_m be the number of individuals in generation m in the branching process. The expected value $EZ_m = Z_0\mu^m$, so if we let $\hat{Z}_m^T$ be the number of individuals in generation m with offspring at time T and ρ_k be the probability an individual in generation 0 has offspring in generation k, we have

$$E\hat{Z}_m^T = Z_0\mu^m\rho_{T-m}, \tag{1.6}$$

a result that was derived earlier by Griffiths and Pakes (1988). To compare this result with the prediction of the coalescent let $\hat{Y}_m^T$ be the number of individuals in generation m that have offspring in generation T. In Section 6 we will show that

Theorem. *If $T \to \infty$ and $M \to \infty$ then*

$$\max_{M \leqslant m \leqslant T} \left| \frac{\hat{Y}_m^T}{N_m \rho_{T-m}} - 1 \right| \to 0 \quad \textit{in probability.} \tag{1.7}$$

Having proved the equivalence between the coalescent in an exponentially growing population and the corresponding Poisson branching process, we will feel free to use either process to investigate mutations at positive times. This and the ordinary coalescent in a population of constant size are the three models we will consider here. All of our estimates of the number of SNPs are based on an estimate of the per nucleotide per generation probability, u, of a mutation, so we make the

Important announcement. *To remove the mutation probability from later calculations, we will instead calculate the expected total time in the genealogy.*

The reader can then multiply by their favorite estimate of the mutation rate to get a concrete estimate of the number of SNPs. Along the way we will do this with our favorite estimate $u = 1 \times 10^{-8}$ which comes from Drake et al. (1998).

Having announced our plan, we have $8 = 2 \times 2 \times 2$ things to do. We have to compute the total time in the genealogy at positive times and at negative times, in the whole population and in a sample of size n, and in addition for these four combinations we have to compute the expected amount of "good time," times when the mutation will be a SNP, i.e., have frequency between 1% and 99%.

1.1. Results for the entire population

1.1.1. Total time for $t \geqslant 0$

The expected total time in the tree between times 0 and T is

$$\sum_{m=0}^{T} N_0 \mu^m \rho_{T-m} = N_T \sum_{k=0}^{T} \mu^{-k} \rho_k. \tag{1.8}$$

Using our approximation $\rho \approx 0.0052$ and recalling $\mu^{-T} = 1/500{,}000$ we have

$$\rho \sum_{k=0}^{T} \mu^{-k} = \rho \, \frac{1 - \mu^{-(T+1)}}{1 - \mu^{-1}} \approx \frac{\rho \mu}{\mu - 1} \approx 2.$$

The second part of the sum must be evaluated numerically. Stopping the first time the survival probability $\rho_k < 0.0052$, which occurs at $k = 2178$, we have

$$\sum_{k=0}^{T} \mu^{-k} (\rho_k - 0.0052)^+ \approx 8.78.$$

Combining the last two results with (1.8), and recalling that the number of the copies of the nucleotide at time T is $N_T = 2(5 \times 10^9)$ gives

$$\textit{the expected total time in the tree in } 0 \leqslant t \leqslant T \textit{ is } 1.078 \times 10^{11}. \tag{1.9}$$

Taking $u = 1 \times 10^{-8}$ as our estimate for the mutation rate, it follows that the expected number of mutations per nucleotide is approximately

$$(1 \times 10^{-8}) \cdot 2(5 \times 10^9) \cdot 10.78 = 1078.$$

This is a huge number of mutations. However, an average of $(1 \times 10^{-8})2(5 \times 10^9) = 100$ of these mutations occurred in the most recent generation. In a moment, we will see that almost all of the 1078 mutations per site exist at very small frequencies.

1.1.2. Good time for $t \geqslant 0$

Our next step is to calculate the probability that a mutation will have a frequency greater than 1% in the population today. Suppose, for simplicity, that the mutation occurs at time 0. Our estimate of the survival probability in the branching process implies that on the average a fraction 0.0052 of the 20,000 individuals at time 0, or 104, will have descendants alive at time T. To estimate the probability that a mutation at time 0 will have a frequency greater than 1% at time T, we use a result of Jagers (1975), see (2.2) below, to conclude that since our branching process is close to critical, the number of descendants at time T, conditioned to be positive, and divided by its mean, has approximately an exponential distribution.

If we ignore the variability of the total of the 104 normalized family sizes, an assumption we will justify in Section 2 by computing the exact distribution, then it follows from Jagers' result that the probability of ending up at a frequency greater than 1% is approximately $\exp(-1.04) = 0.3534$. As one moves forward to generation m, the number of individuals with offspring alive at time T grows to $N_0 \mu^m \rho_{T-m} \geqslant 104 \mu^m$. Using the lower bound, we see that the probability a mutation will have a frequency greater than 1% is approximately $\exp(-1.04\mu^m)$, which decays to 0 very rapidly. The last fact implies that the difference between ρ_{T-m} and ρ is unimportant in this case.

Summing we see that the number of opportunities in generations $0 \leqslant t \leqslant T$ for a mutation with frequency of at least 1% is

$$\approx \sum_{t=0}^{T} 104\mu^t \exp(-1.04\mu^t) = 13{,}595. \tag{1.10}$$

Since any mutation at a positive time will be contained inside one of the 104 families at time 0, there is only a very small probability that the mutation will end up with 99% of the population, and we will ignore this. Multiplying (1.10) by our mutation rate estimate of $u = 1 \times 10^{-8}$ gives a per nucleotide probability for SNPs of

$$p = 1.3595 \times 10^{-4} \quad \text{or 1 SNP from } [0, T] \text{ every 73,556 bp.}$$

As some readers may have noticed, this density is much less than the figure of 1 SNP per Kilobase (Kb) that is often quoted. (See Wang et al., 1998; Lai et al., 1998; Brookes, 1999.) There is no contradiction, however. As we will soon see, most of the mutations that are SNPs occurred at times $t < 0$.

1.1.3. Total time for $t < 0$

To count the expected number of mutations at times $t < 0$ we will use the theory of the coalescent. To follow the arguments below, the reader need know only that if time is written in units of N_0 generations, then in the limit as $N_0 \to \infty$ the number of lineages in the coalescent decreases from k to $k-1$ at rate $k(k-1)/2$ (see e.g., Kingman, 1982a; Hudson, 1990). To begin to compute the expected number of mutations at times $t < 0$, we note that each of the original $N_0 = 20,000$ nucleotides will have offspring at time T with probability 0.0052, so the number that succeed has approximately a Poisson distribution with mean 104. When there are K success in the population we have to work backwards in time until their lineages coalesce. The result cited above implies that the time required, when measured in units of N_0 generations, has mean

$$\sum_{k=2}^{K} k \frac{2}{k(k-1)} \approx 2 \ln K.$$

We will argue in Section 3 that it is permissible to replace K by its mean in the formula above. However, as the reader can easily check by computing each side of the equation for $K = 104$, to get an accurate answer one must use the sum rather than its approximation. Multiplying by $N_0 = 20,000$ we arrive at the following estimate for the total time in the tree at times $t < 0$ for the population:

$$20{,}000 \sum_{k=2}^{104} \frac{2}{k-1} = 20{,}000 \times 10.43358 = 208{,}672.$$

Taking $u = 1 \times 10^{-8}$, the expected number of mutations per nucleotide at times $t < 0$ is

$$p = 2.08672 \times 10^{-3} \quad \text{or one mutation from } t < 0 \text{ every 479 bp.}$$

1.1.4. Good time for $t < 0$

To determine the frequency of the mutations at times $t < 0$ in the population at the present time T, we use Ewens, (1972) sampling formula in Section 3 to conclude, see (3.11), that on the average 3.3268×10^{-4} mutations per nucleotide fail to end up with a frequency between 1% and 99% of the population. This reduces the probability given above to

$$p = 1.75345 \times 10^{-3} \quad \text{or 1 SNP from } t < 0 \text{ every 570 bp.}$$

Dividing by our mutation rate estimate $u = 1 \times 10^{-8}$ we see that

$$\textit{the expected amount of good time in the tree for } t < 0 \textit{ is } 175{,}345. \tag{1.12}$$

Adding the 13,595 from (1.10) for the good times $t \in [0, T]$, we get an expected total good time in the tree of 189,940. From this we get

Our main result. *Assuming a mutation rate of $u = 1 \times 10^{-8}$ gives an estimate for the density of SNPs of $p = 1.8994 \times 10^{-3}$ or 1 SNP every 526 bp.*

Dividing 3×10^9 by 526 gives our estimate of 5.7 million SNPs in the human genome. These results in (1.9)–(1.12) are summarized in Table 1 for comparison with

Table 1
Summary of computations

	Population	Sample
Total time in $[0,T]$	1.078×10^{11}	29,891
Good time in $[0,T]$	13,595	458
Total time for $t<0$	208,672	89,074
Good time for $t<0$	175,345	82,881
Total time, total		119,595
Good time, total	189,940	83,356

1.2. Results for a sample of size n

As we have mentioned earlier, we will begin by considering the special case $n=6$, because of its connections to Celera's strategy. A second reason is that restricting our attention to $n=6$ will give us the opportunity to have concrete numerical answers in addition to our sometimes complicated formulas.

1.2.1. Total time for $t \geqslant 0$

As in the case of the entire population, we will begin at the present time T and work backwards. To get an upper bound on the size of the genealogy of a sample of size six, we can suppose that the six lines stay distinct until time 0, after which they coalesce in the usual way. The total time in the genealogical tree between 0 and T will then be $6 \cdot 5000 = 30{,}000$. The last result is an upper bound, but it turns out to be quite a good one. Let $X(t)$ be the number of lineages surviving to time t. One can recursively compute (even with a small computer) the probabilities $P(X(t)=k)$ working backwards from time T to conclude

$$\sum_{t=0}^{T-1} EX(t) = 29{,}891. \tag{1.13}$$

This result shows that the naive upper bound of 30,000 is very good. This outcome is no surprise since genealogies in exponentially growing populations are known to be "star-shaped", see e.g., Slatkin and Hudson (1991).

1.2.2. Good time for $t \geqslant 0$

Using the reasoning that led to (1.10) we can conclude that for the sample of size 6, the expected amount of good time (i.e., instances at which a mutation will be a SNP at time T) in the tree during $[0,T]$ is

$$\sum_{t=0}^{T-1} EX(t) \exp(-1.04\mu^t) = 458. \tag{1.14}$$

Multiplying by 1×10^{-8} leads to an estimate of 4.58×10^{-6} SNPs per nucleotide from mutations in $[0,T]$ or 1 every 218,340 nucleotides. However, this number was doomed to be disappointing by the corresponding computation for the whole population: the

expected number of polymorphic mutations from $[0,T]$ is only 1 SNP from $[0,T]$ every 73,556 nucleotides.

1.2.3. Total time for $t<0$

To begin, we note that if all the six lineages do not coalesce before time 0, then by the reasoning that led to (1.11) the expected total time in the tree before 0 will be

$$2\left(\frac{1}{5}+\frac{1}{4}+\frac{1}{3}+\frac{1}{2}+1\right)20{,}000=91{,}333.$$

Note that even though there are six individuals rather than 104 as in (1.11), the total time here is about 44% of the time 208,672 for the whole population given in (1.12).

The first correction that must be made in the previous calculation is to realize that the six lineages from time T will undergo some coalescence during $[0,T]$. Let $X(t)$ be the number of distinct lineages at time t. Recursively computing $P(X(t)=k)$ starting from $P(X(T)=6)=1$ gives

k	6	5	4	$\leqslant 3$
$P(X(0)=k)$	0.74881	0.22742	0.02284	9.2×10^{-4}.

Taking this into account, however, does not make a big change. The expected total time in the tree drops only a small amount to 89,074, a loss of about 2.5%.

1.2.4. Good time for $t<0$

The next factor to consider is that not all mutations will have a frequency of 1%. Using a result of Joyce and Tavaré (1987) that relates the coalescent to the binary branching process, and some elementary computations with the distribution of order statistics, see Section 5, we can compute that the expected total amount of good time in the tree at times $t<0$ to be 82,881. See Table 1 which summarizes our results for the population and the sample of size 6.

Our mutation rate estimate is 1×10^{-8} per nucleotide for the human genome, which has 3×10^{9} nucleotides, so there are an average of 30 mutations per generation. Multiplying the total time 119,595 by 30 gives our prediction that a sample of size 6 will have 3.58 million variable nucleotides. Of course, only the mutations at good times will produce SNP's so

$$\frac{83{,}566}{119{,}565}=70\%$$

of the variable nucleotides or 2.51 million are SNPs.

Having worked to do computations for the special case of a sample of size 6, it is now straightforward to generalize to samples of any size. The results are given in Table 2.

During the time it took to write this paper and have it refereed, Celera sequenced the entire human genome 4.6 times (rather than their initially proposed 10) using DNA from five individuals, three females and two males who have identified themselves as Hispanic, Asian, Caucasian, or African American. To provide an overestimate of the number of SNPs they found, we will assume that they have 5 times coverage of the genome with each nucleotide sequences from five different chromosomes. They

Table 2
Results for varying sample size

	Good time in $[0,T]$	Good time for $t<0$	Total good time	Sample vs. population	Fraction that are SNPs
2	157	36,786	36,943	0.196	0.750
3	234	55,180	55,413	0.293	0.751
4	309	67,180	67,489	0.357	0.735
5	384	75,984	76,369	0.404	0.718
6	458	82,881	83,339	0.441	0.701
7	531	88,514	89,045	0.471	0.684
8	603	93,251	93,854	0.497	0.668
9	674	97,321	97,995	0.519	0.653
10	745	100,876	101,620	0.538	0.639
12	882	106,834	107,717	0.570	0.612
14	1,017	111,679	112,697	0.596	0.589
16	1,149	115,732	116,881	0.619	0.567
18	1,277	119,193	120,471	0.638	0.547
20	1,403	122,197	123,600	0.654	0.529

claim to have found 2.4 million SNPs. (This information comes from press releases on Celera's web page: www.celera.com.) Consulting the table our prediction is that they have found 40.4% of the 5.7 million SNPs in the human genome or 2.3 million.

At the top of Table 2 we see that in the case $n=2$, 75% of differences between two chromosomes are SNPs, and that one individual already has about 20% of the SNPs in the genome. At the other end of Table 2, the case $n=20$ gives results relevant to the experimental set up of Wang et al. (1998), who screened genetic material from 10 humans to look for SNPs in 26,568 sequence tagged sites (STSs) used in the construction of a physical map of the human genome at the Whitehead Institute. As Table 2 indicates, if one were to simply accept sites that were polymorphic in the sample then one would find 65.4% of the SNPs in the region surveyed, but one would also find an almost equal number of variable nucleotides that are not SNPs.

Wang et al. (1998) did not use this naive experimental design. To quote from their paper: "Each STS was amplified from four samples: three individual samples and a pool of 10 individuals. Candidate SNPs were declared when two alleles were seen among a subsample of three individuals, with both alleles present at a frequency of greater than 30%". In principle, one could also use our methods to compute the probability of success with that strategy, however, we have not yet attempted to wrestle with the details. Of course, no algorithm can find SNPs that are not variable in the sample so the 65.4% figure is an upper bound on the performance of any selection algorithm. Wang et al. (1998) found 279 "candidate SNPs" after screening 279,165 nucleotides, which corresponded to a rate of one SNP per 1001. Multiplying the good time in the tree for a sample of size 20 given in Table 2 times our mutation rate 1×10^{-8} we conclude that the per nucleotide probability of a good SNP in this region of the genome is 1.236×10^{-3} or one every 809 bp, in good agreement with what Wang et al. (1998) found.

Up to this point all of our calculations have been done using the 1% definition of polymorphism. However the basic computational machinery generalizes to other cutoff

Table 3
Varying the threshold for polymorphism

%	Good time for $[0, T]$	Good time for $t < 0$	Relative to 1%
1	13,593	175,345	1.000
5	42	117,342	0.621
10		87,461	0.463
15		68,931	0.365
20		54,970	0.291
25		43,430	0.230
30		33,340	0.176

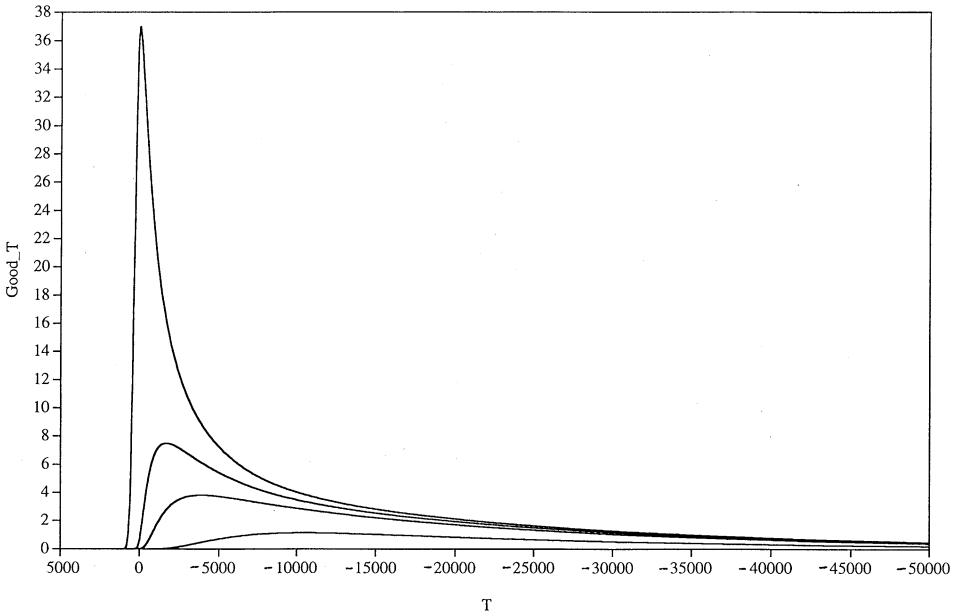

Fig. 1. Good time in the tree as a function of generation number for the 1, 5, 10 and 30% definition of SNP.

levels. Table 3 gives results for cutoffs of 5–30% compared to the number present with the 1% definition. In the first column numbers that are < 1 are left blank. Note that when the threshold is increased to 0.05, 38% of the 0.01 level SNPs are no longer considered polymorphic. At the other extreme insisting on between 30% and 70% reduces the total number to 17.6% of the original collection. Using $u = 1 \times 10^{-8}$ for a mutation rate we have a prediction of one SNP every 526 bp, so using SNPs where the most common allele is at most 70% would produce one every 2988 bp, which matches the density Kruglyak (1999) says we need.

The methods used to generate Table 3 can without any additional effort give us how the good time in the tree is distributed over time. Fig. 1 shows results for the 1%, 5%, 10%, and 30% definitions of SNPs. Comparison with the histograms in Kruglyak (1999) shows that our analytical approach provides more refined results than simu-

lation alone. Combining this information with an estimate of the recombination rate per generation one can obtain an estimate of the range over which there is significant linkage disequilibrium, however we have not carried out the details of the calculation.

The remainder of the paper is devoted to justifying the claims and doing the computations reported in the introduction. We will study the number of mutations in the whole population in $[0, T]$ in Section 2, and the number of mutations at times $t < 0$ in Section 3. Sections 4 and 5 are devoted to the number of mutations at corresponding times in a sample. Finally in Section 6, we prove (1.7) to justify our claim that the branching process and the coalescent approaches give the same answer for the expected number of mutations.

2. Mutations in the population at times $0 \leqslant t \leqslant T$

Let X have a Poisson distribution with mean μ. Differentiating the definition of the generating function

$$\phi'(1) = EX = \mu \quad \text{and} \quad a \equiv \phi''(1)/2 = E(X(X-1)/2).$$

Expanding in a Taylor series about 1, we have

$$\phi(1-x) = \phi(1) - \phi'(1)x + \phi''(1)\frac{x^2}{2} + \cdots = 1 - \mu x + ax^2 + \cdots.$$

Setting $1 - x = \phi(1-x)$ and rearranging we have for $\mu \approx 1$

$$(\mu - 1)x \approx ax^2 \quad \text{so } x \approx \frac{\mu - 1}{a}.$$

Letting $\rho_k = 1 - \sigma_k$ denote the probability of the family line surviving for k generations and $\rho = 1 - \sigma$ be the limit of $\rho_k = P(Z_k > 0)$, this says that in the Poisson case with mean $\mu = 1.0026$

$$\rho \approx \frac{\mu - 1}{a} \approx 0.0052. \tag{2.1}$$

Our next task is to compute the distribution of the frequency at time T of a mutation that occurs at time t. We begin with special case $t=0$. Each individual at time 0 starts a copy of the branching process Z_t. Well-known results imply that when $\mu > 1$ as $t \to \infty$, $Z_t/\mu^t \to W$ a random variable with $EW = 1$ and $P(W > 0) = P(Z_t > 0 \text{ for all } t)$. A less widely known result, but one very useful for us is Theorem 3.3.1 of Jagers (1975).

Theorem. *For any $\alpha > 0$ let $\mathscr{K}_\alpha$ be a class of Galton–Watson processes with reproduction variances less than α and uniformly convergent second reproduction moments (i.e., for each $\varepsilon > 0$ it should be possible to choose k_ε so that*

$$\sum_{k > k_\varepsilon} k^2 p_k < \varepsilon \tag{2.2}$$

for all reproduction laws in the class). Suppose that the number 1 belongs to the closure of the set of reproduction means of processes in $\mathscr{K}_\alpha$. Write $a = \phi''(1)/2$ and

interpret $c_{\mu,n} = (1-\mu^{-n})/(\mu-1)$ *as* n *for* $\mu = 1$. *Then uniformly for all process in* $\mathscr{K}_\alpha$ *as* $n \to \infty$ *and* $\mu \to 1$

(a) $ac_{\mu,n}P(Z_n > 0) \to 1$,
(b) $E(Z_n/ac_{\mu,n}|Z_n > 0) \to 1$,
(c) $P(Z_n/ac_{\mu,n} > u|Z_n > 0) \to \mathrm{e}^{-u}$.

If we apply this result with $\mathscr{K}_\alpha$ equal to the Galton–Watson processes in which the offspring distribution is Poisson with a mean $\mu \in [\frac{1}{2}, 2]$, it follows that when μ is close to 1, the limit distribution has $P(W = 0) = 1 - \rho$ and

$$P(W > x) \approx \rho \mathrm{e}^{-\rho x}. \tag{2.3}$$

As should be clear from the formulation of Jagers' theorem, (2.2), this result holds for a general distribution, but the value of the survival probability ρ given in (2.1) changes.

Let K be the number of individuals in generation 0 with offspring at time T. Each of the original $N_0 = 20{,}000$ nucleotides, will have offspring at time T with probability 0.0052, so the number that succeed has approximately a Poisson distribution with mean 104. If we let $V_1, V_2, \ldots$ be independent with $P(V_i > x) = \mathrm{e}^{-x}$ then the fraction of offspring in family 1 has approximately the same distribution as $V_1/\sum_{i=1}^{K} V_i$. Writing $S = \sum_{k=2}^{K} V_i$ and $V = V_1$ we have

$$\begin{aligned} P\left(\frac{V}{V+S} > x\right) &= P\left(V > \frac{x}{1-x} S\right) \\ &= \int_0^\infty \mathrm{e}^{-xs/(1-x)} P(S = s)\, \mathrm{d}s = E\mathrm{e}^{-xS/(1-x)}. \end{aligned} \tag{2.4}$$

Now $E\mathrm{e}^{-\theta V_i} = 1/(1+\theta) = 1-x$ when $\theta = x/(1-x)$. Writing λ instead of 104 to prepare for later generalizations, and summing over the possible values of our Poisson random variable K, except for $K = 0$ which has probability e^{-104}, we have

$$E\mathrm{e}^{-xS/(1-x)} = \sum_{k=1}^{\infty} \mathrm{e}^{-\lambda} \frac{\lambda^k}{k!} (1-x)^{k-1} = \frac{\mathrm{e}^{-\lambda x} - \mathrm{e}^{-\lambda}}{1-x}. \tag{2.5}$$

Note that as $x \to 1$, the right-hand side does not go to 0 but to $\lambda \mathrm{e}^{-\lambda} = P(K = 1)$.

Changing variables $\lambda x = y$ leads to

$$P\left(\frac{V}{V+S} > y/\lambda\right) = \frac{1}{1-y/\lambda}(\mathrm{e}^{-y} - \mathrm{e}^{-\lambda}). \tag{2.6}$$

We will use this formula with $y/\lambda = 0.01$, so we will (i) ignore the first factor which is 1/0.99 and (ii) ignore $\exp(-\lambda) = \mathrm{e}^{-100y}$ which for $y \geqslant 1$ is much smaller than e^{-y}. Implementing these two ideas, we have the very simple conclusion

$$P\left(\frac{V}{V+S} > \frac{y}{E(V+S)}\right) \approx \mathrm{e}^{-y}. \tag{2.7}$$

We could have arrived at this end much more easily if we had simply replaced $V+S$ in the denominator by its mean. However, now we know that this simple approximation is accurate for the values of y and λ we are concerned with.

To illustrate the use of (2.6) we note that at time 0, $E(V+S)=104$, so if we are interested in families that are more than $x=0.01$ of the population then $y=1.04$ and the probability is about $e^{-1.04}=0.3534$. As we mentioned in the introduction, we don't worry about the probability that $V/(V+S)>0.99$ which by (2.6) with $\lambda=104$ and $y/\lambda=0.99$ is

$$100(e^{-102.96}-e^{-104}).$$

As we move forward from time 0 the probability of being larger than 1% drops to 0 very quickly. The population doubles every 267 generations, so in generation $267r$, we have $\lambda=104\times 2^r$. The value $x=0.01$ corresponds to $y=1.04\times 2^r$, so the probability of ending up with at least 1% of the population at the rth stage is $\exp(-1.04\times 2^r)$. Numerical results show that this rapidly gets very small

r	0	1	2	3	4
probability	0.3534	0.1249	0.0156	2.4×10^{-4}	5.9×10^{-8}.

In words, each time the population doubles the probability of success is squared.

The last calculation shows that there will be a negligible contribution from times $t\geqslant 1000$ so by a remark in the calculations used to evaluate the sum (1.8), we can replace ρ_k by its limit ρ. Thus, to compute the number of opportunities for mutations at times $t\geqslant 0$ we have to evaluate

$$\sum_{t=0}^{T} 104\mu^t \exp(-1.04\,\mu^t). \tag{2.8}$$

Replacing the sum by an integral from 0 to ∞ and then changing variables $x=\mu^t$, $dx=(\ln\mu)\mu^t\,dt$ we have that sum above is

$$\approx \frac{100}{\ln\mu}\int_1^\infty 1.04\exp(-1.04x)\,dx=\frac{100}{0.0026}e^{-1.04}=13,595. \tag{2.9}$$

Of course, one can skip the approximation and the calculus, by evaluating the sum in (2.8) numerically to find that it is 13,593.70.

3. Mutations in the population at times $t<0$

To count the expected number of mutations at times $t<0$ we can use coalescent theory. Let K be the number of individuals at time 0 that have offspring alive at time T. Well-known results about the coalescent imply that as we work backwards from time 0, the amount of time required for coalescence, when measured in units of N_0 generations, has mean

$$\sum_{k=2}^{K}\frac{2}{k-1}\approx 2\ln K. \tag{3.1}$$

We would like to replace K by its mean. To justify this we expand $f(x)=2\ln x$ in Taylor series around the point $x_0=EK$ to conclude that

$$E(2\ln K)=2\ln EK+\frac{2}{EK}E(K-EK)-\frac{1}{(EK)^2}E(K-EK)^2+\cdots. \tag{3.2}$$

We have $E(K - EK) = 0$, $E(K - EK)^2 = 104$, and $EK = 104$, so the first correction term vanishes and the second is $\frac{1}{104} = 0.0096$ compared with $2 \ln 104 = 9.2888$, so we can safely ignore it.

The difference of the two sides of (3.1), which some readers will recognize as roughly two times Euler's constant $\gamma \approx 0.577$, is not small enough to ignore, so we will instead evaluate the sum

$$\sum_{k=2}^{104} \frac{2}{k-1} = 10.43358 \tag{3.3}$$

and conclude that the expected total time in the tree at times $t < 0$ is approximately

$$10.43358 \times 20{,}000 = 208{,}672. \tag{3.4}$$

Taking $u = 1 \times 10^{-8}$ for our estimate of the mutation rate, the expected number of mutations per nucleotide is 1.043358×10^{-3} or one every 958 bp.

To determine the distribution of the frequencies of these mutations, we use Ewens, (1972) sampling formula. Recalling that our N is the total number of copies and letting

$$\theta = 2Nu = 2 \times (2 \times 10^4) \times (1 \times 10^{-8}) = 4 \times 10^{-4},$$

it says that the probability of an allelic partition $(a_1, a_2, \ldots a_n)$ is given by

$$\frac{n!}{\theta_{(n)}} \prod_{j=1}^{n} \left(\frac{\theta}{j}\right)^{a_j} \frac{1}{a_j!}. \tag{3.5}$$

Here a_i is the number of alleles with i representatives in the sample of size n and $\theta_{(n)} = \theta(\theta+1)\cdots(\theta+n-1)$. When $a_n = 1$ this becomes

$$\frac{1}{\theta+1} \frac{2}{2+\theta} \cdots \frac{(n-1)}{(n-1)+\theta}, \tag{3.6}$$

which is the probability that coalescence always comes before the next mutation. Plugging in $\theta = 2 \times 10^{-4}$ and $n = 104$, we see that the probability of no mutation is

$$p_0 = 0.997915590 = 1 - (2.084410 \times 10^{-3}).$$

The reader will see the reason for the high degree of precision in a minute.

Consider now the case in which $a_j = a_{n-j} = 1$ for some $j < 52$ and $a_i = 0$ otherwise. Replacing the product in (3.6) by its value computed above, (3.5) becomes

$$p_0 \theta \frac{n}{j(n-j)}. \tag{3.7}$$

When $j = 52$ the answer is $\frac{1}{2}$ of this. Summing the probability of a j to $n-j$ split for $j = 1$ to 52 gives that the probability of one mutation is

$$p_1 = 2.082367 \times 10^{-3}. \tag{3.8}$$

Using this with the value of p_0, we see that the probability of two or more mutations is

$$q_2 = 2.043 \times 10^{-6}.$$

Note that q_2 is much smaller than p_1, the probability of one mutation, consistent with the observation that "in humans, tri-allelic and tetra-allelic SNPs are rare to the point of nonexistence". See e.g., Brookes (1999).

Table 4
Probabilities that the total of j out of 104 families is at most 1% of population at time T

j	$f(j,104)$	j	$f(j,104)$
1	0.64839	6	6.58×10^{-4}
2	0.27902	7	9.11×10^{-5}
3	0.08687	8	1.09×10^{-5}
4	0.02088	9	1.16×10^{-6}
5	0.00405	10	1.11×10^{-7}

Suppose now that j of the K lineages at time 0 have the mutation. Here we are assuming that K is fixed, as would occur if we were looking at the conditional distribution of the good time $t < 0$ conditional on the number of families at time 0 that have offspring at time T. We have argued earlier that variability in K can be ignored so we will set $K = 104$. Our next step is to compute $f(j,K) =$ the probability a fixed set of j of the K families end up with at most 1% of final population. To do this let $\xi_1, \ldots, \xi_K$ be independent mean one exponentials, and let $S_j = \xi_1 + \cdots + \xi_j$.

It is a standard fact that $\{S_j/S_K, 1 \leqslant j < K\}$ has the same distribution as the order statistics from a sample of $K-1$ random variables uniform on $(0,1)$. The last observation leads easily to the following formula for the density function. If $1 \leqslant j < K$ then

$$P(S_j/S_K = x) = (K-1)\binom{K-2}{j-1} x^{j-1}(1-x)^{K-1-j}. \tag{3.9}$$

A little calculus shows that

$$P(S_j/S_K \leqslant y) = 1 - \sum_{i=0}^{j-1} \binom{K-1}{i} y^i (1-y)^{K-1-i}, \tag{3.10}$$

which can be checked by induction or by noting that the right hand side is the probability that at least j particles will end up in $(0,y)$ when we throw $K-1$ at the unit interval.

Setting $y = 0.01$ and $K = 104$ in (3.10) we can compute Table 4.

In our numerical computations we will suppose that this probability $f(j,104) = 0$ for $10 \leqslant j \leqslant 52$, and thus make a very small error in our computations. Taking this into account, the loss from the mutation probability is

$$\sum_{j=1}^{10} \theta \frac{104}{j \cdot (104-j)} f(j,104) = 1.66634 \times 10^{-4}, \tag{3.11}$$

which reduces the previous frequency of 10.4227×10^{-4} given in (3.6) to

$$p = 8.76724 \times 10^{-4} \quad \text{or 1 every 1140 bp.} \tag{3.12}$$

Dividing by our mutation estimate $u = 5 \times 10^{-9}$, we see that the expected good time in the tree for $t < 0$ is 175,345.

4. Mutations in the sample at times $0 \leqslant t \leqslant T$

Our first task in this section is to compute the expected number of mutations hitting a genealogy of six individuals. To get an upper bound, we can suppose that the six lines stay distinct until time 0, after which they coalesce in the usual way. The total time in the tree will then be $6 \times 5000 = 30,000$ between 0 and T. To get an exact result we have to compute how much coalescence occurs between the six lineages in $[0, T]$. When there are k lineages at time $t+1$ a coalescence will occur at time t with probability

$$\binom{k}{2} \Big/ N_t + \mathrm{O}(1/N_t^2).$$

From this it follows that if we represent the time interval $[t, t+1]$ as a segment of length $1/N_t$ then on the new time scale, our process is almost the continuous time coalescent in which k lineages coalesce after an exponentially distributed amount of time with mean $1/\binom{k}{2}$.

This idea which is due to Kingman (1982b), see page 104, allows us to reduce our computation for a population of variable size to one for the ordinary coalescent run for an amount of time

$$\tau = \sum_{t=0}^{T-1} \frac{1}{N_t} = \frac{1}{N_0} \sum_{t=0}^{T-1} \mu^{-t} = \frac{1-\mu^{-T}}{N_0 \cdot (1-\mu^{-1})} \approx \frac{\mu}{N_0(\mu-1)} = \frac{1.0026}{52}. \tag{4.1}$$

Reindexing time so that time 0 is the present, and so that s represents s units of time in the past, let T_k be the first time at which there are only k lineages. Since $\binom{6}{2} = 15$, it is clear that the probability of no coalescence is

$$P(T_5 > \tau) = \exp(-15\tau) = 0.7488538. \tag{4.2}$$

Since $\binom{5}{2} = 10$, breaking things down according to the value of T_5 we have that the probability of ending up with 5 lineages at time τ is

$$\begin{aligned} P(T_4 > \tau > T_5) &= \int_0^\tau 15\mathrm{e}^{-15r}\mathrm{e}^{-10(\tau-r)}\,\mathrm{d}r = 3\mathrm{e}^{-10\tau}\int_0^\tau 5\mathrm{e}^{-5r}\,\mathrm{d}r \\ &= 3(\mathrm{e}^{-10\tau} - \mathrm{e}^{-15\tau}) = 0.2285897. \end{aligned} \tag{4.3}$$

This already accounts for 97.6% of the probability but we can go further by noting

$$P(T_4 > s) = P(T_5 > s) + P(T_4 > s > T_5) = 3\mathrm{e}^{-10s} - 2\mathrm{e}^{-15s} \tag{4.4}$$

and $\binom{4}{2} = 6$ so the probability of ending up with 4 lineages at time τ is

$$\begin{aligned} P(T_3 > \tau > T_4) &= \int_0^\tau 30(\mathrm{e}^{-10r} - \mathrm{e}^{-15r})\mathrm{e}^{-6(\tau-r)}\,\mathrm{d}r = 30\mathrm{e}^{-6\tau}\int_0^\tau (\mathrm{e}^{-4r} - \mathrm{e}^{-9r})\,\mathrm{d}r \\ &= 30\mathrm{e}^{-6\tau}\left(\frac{1}{4}(1-\mathrm{e}^{-4\tau}) - \frac{1}{9}(1-\mathrm{e}^{-9\tau})\right) = 0.0228590, \end{aligned} \tag{4.5}$$

which now accounts for 99.9% of the probability.

To compute the effect coalescence has on reducing the tree between times 0 and T, we suppose that there is a constant coalescence probability $15/N_t$. If a reduction in the

Table 5
Probability that j of the 6 lineages survive to time 0

j	Coalescent	Recursion
6	0.74885	0.74881
5	0.22736	0.22742
4	0.02285	0.02284
3		9.1×10^{-4}
2		1.4×10^{-5}
1		5.3×10^{-8}

number of lineages occurs at time k then we have lost one lineage for $k+1$ times $(0,1,\ldots k)$ so the expected loss is at most

$$\frac{15}{N_0}\sum_{k=0}^{T-1}\mu^{-k}(k+1). \tag{4.6}$$

Using the fact that the mean of geometric with success probability $p=1-\mu^{-1}$ is $1/p$ the above is

$$\approx \frac{15}{N_0(1-\mu^{-1})^2} = \frac{15\mu^2}{52\times 0.0026} = 111.52. \tag{4.7}$$

Subtracting this from the upper bound of 30,000 gives an adjusted estimate of 29,888.48 for the total time in the tree of the six individuals during $[0,T]$.

Using the logic that led to (2.7), we can compute the number of opportunities at times $t\geqslant 0$ for mutations that lead to polymorphic SNP's, by evaluating

$$\sum_{t=0}^{T-1} EX(t)\exp(-1.04\times\mu^t),$$

where $X(t)$ is the number of lineages at time t. To do this we have written a program to work backwards from time T to time 0 computing $P(X(t)=k)$ by the discrete time recursion

$$P(X(t-1)=k) = \left(1-\frac{\binom{k}{2}}{N(t-1)}\right)P(X(t)=k)+\frac{\binom{k+1}{2}}{N(t-1)}P(X(t)=k+1). \tag{4.8}$$

Table 5 gives the values of $P(X(0)=k)$ and compares with the values computed earlier using the continuous time coalescent.

The small discrepancy between the two sets of answers is due mainly to the fact that the coalescent computation is for the continuous time limit, while the recursion happens in discrete time. However, there is also some round off error which effects the sixth significant digit in these computations.

Using the values of $P(X(t) = k)$, from the discrete time recursion our program computes

$$\sum_{t=0}^{T-1} EX(t) = 29{,}891.08,$$

$$\sum_{t=0}^{T-1} EX(t)\exp(-1.04 \times \mu^t) = 458.13. \tag{4.10}$$

The first result shows that our lower bound of 29,888.48 from (4.7) is very accurate. The second answer is considerably lower than the 13,595 for the total population given in (2.9), however, the most important source of mutations is yet to come.

5. Mutations in the sample at times $t \leqslant 0$

If there are exactly six lineages at time 0 then using the reasoning that led to (3.1) the expected total time in the genealogy before time 0 will be

$$2 \times \left(\frac{1}{5} + \frac{1}{4} + \frac{1}{3} + \frac{1}{2} + 1\right) \times 20{,}000 = 91{,}333. \tag{5.1}$$

Taking into account the possibility there may not be six lineages at time 0 and using the coalescence probabilities computed in the previous section, we then arrive at the corrected value for the expected total time for the sample before time 0,

$$2 \times \left(2.28333 - \sum_{k=1}^{5} P(T_k \leqslant \tau)\frac{1}{k}\right) \times 20{,}000 = 89{,}074, \tag{5.2}$$

which is about a 2.5% reduction from the previous value.

As in Section 3, not all of the mutations will lead to an allele that has a frequency of 1% at time T. To begin to attack this problem we will first give a new more complicated solution of the problem of computing the expected number of mutations for the whole population at times $t < 0$. The key is the following

Lemma. *Consider the coalescent starting with ℓ lineages. If we pick one of them at random when there are $k \leqslant \ell$ lineages then the probability it will contain m of the ℓ starting lineages is*

$$s(k,m) = \binom{\ell - m - 1}{k - 2} \Big/ \binom{\ell - 1}{k - 1}. \tag{5.3}$$

Proof. It is well known and easy to check that we can build up the relationships between particles in the continuous time coalescent by running a continuous time binary branching process in which each particle splits into two at rate 1. (See Joyce and Tavaré, 1987.) Consulting p. 109 of Athreya and Ney (1972) we see that starting from a single particle at time 0, the number of particles in a binary branching process at time t has a geometric distribution with success probability $p = \mathrm{e}^{-t}$. Let Z_t^i, $1 \leqslant i \leqslant k$

be independent copies of the branching process. If $j_1,\dots j_k$ be positive integers that add up to ℓ then

$$P(Z_t^1 = j_1,\dots Z_t^k = j_k) = (1-p)^{\ell} p^k.$$

Recalling that there are $\binom{\ell-1}{k-1}$ possible vectors $(j_1,\dots,j_k)$ of positive integers that add up to ℓ it follows that

$$P\left(Z_t^1 = m \,\middle|\, \sum_{j=1}^{k} Z_t^j = \ell\right) = \binom{\ell-m-1}{k-2} \Big/ \binom{\ell-1}{k-1}. \qquad \square$$

When the mutant has m copies at time 0, the probability it will end up at most 1% of the population is given by the quantity $f(m,104)$ defined at the end of Section 3 for $m \leqslant 52$. We also have to worry about the possibility of the mutation ending up with more than 99% of the population, so we will set $f(104-m,104) = f(m,104)$ for $1 \leqslant m < 52$. Breaking things down according to the number of particles and using (5.3) we see that the expected amount of time $t \leqslant 0$ where a mutation will produce a polymorphic SNP is

$$20{,}000 \sum_{k=2}^{104} \frac{2}{k(k-1)} k \sum_{m=1}^{105-k} s(k,m)[1 - f(m,104)], \tag{5.4}$$

where we have kept the redundant factor of k to prepare for the analogous formula given in (5.6) for the sample. Evaluating the sum gives 175,345 which is identical to the answer found at the end of Section 3.

To compute the answer for the sample, we begin with the case in which there is no coalescence of the six lineages during $[0,T]$. To do the computation in this case, it is useful to think of the original 104 lineages as consisting of six green particles (the sample) and 98 white particles, and that the coalescence of a green particle with another particle (green or white) yields a green particle. Let X_k be the number of green particles when only k of the original 104 lineages remain. Our first step is to compute $p_{k,j} = P(X_k = j)$ starting from $p_{104,6} = 1$, using the discrete time recursion

$$p_{k,i} = p_{k+1,i}\left(1 - \frac{(i-1)i}{k(k+1)}\right) + p_{k+1,i+1}\frac{i(i+1)}{k(k+1)}. \tag{5.5}$$

Once this is done we can compute that the amount of time $t \leqslant 0$ where a mutation will produce a SNP is

$$20{,}000 \sum_{k=2}^{104} \frac{2}{k(k-1)} \mu(k) \sum_{m=1}^{105-k} s(k,m)[1 - f(m,104)], \tag{5.6}$$

where $\mu(k) = \sum_{i=2}^{6} i \cdot p_{k,i}$. We have excluded the term $p_{k,1}$ from the sum defining the "mean" $\mu(k)$ since the mutation must be polymorphic in the sample to be detected. Doing the sum in (5.6) yields the answer 84,919. To check our work we replaced $1 - f(m,104)$ by 1 in (5.6) to compute that the expected total time in the tree for times $t < 0$ was indeed 91,333.33 as we computed in (5.1).

The calculations in the last paragraph are for starting with exactly 6 lineages at time 0. To get the expected value for 6 particles at time T, we have to consider what happens starting with 5 or 4 particles at time 0. The results are given in Table 6.

Table 6
Components of the final answer for times $t < 0$ for sample size 6

Particles	Probability	Total time	Good time	Size bias
6	0.748812	91,333	84,919	88,831
5	0.227422	83,333	77,712	81,142
4	0.023764	73,333	68,587	71,486
	Average	89,073	82,881	86,658

6. Equivalence of the coalescent and branching process approaches

In this section we will prove (1.7). When moving back from generation m to generation $m-1$, each of the $\hat{Y}^T_m$ surviving lineages will pick each of the $1/N_{m-1}$ ancestors with equal probability. Noting that the expected number of ancestors chosen is N_{m-1} times the probability a given ancestor is selected we have

$$E(\hat{Y}^T_{m-1}|\hat{Y}^T_m) = N_{m-1}\left[1 - \left(1 - \frac{1}{N_{m-1}}\right)^{\hat{Y}^T_m}\right]. \tag{6.1}$$

Using the facts that $N_m/N_{m-1} = \mu$ and for large k, $(1-1/k)^y \approx \exp(-y/k)$ this formula can be written as

$$E\left(\frac{\hat{Y}^T_{m-1}}{N_{m-1}}\middle|\frac{\hat{Y}^T_m}{N_m}\right) \approx 1 - \exp\left(-\mu\frac{\hat{Y}^T_m}{N_m}\right) = \psi\left(\frac{Y^T_m}{N_m}\right), \tag{6.2}$$

where $\psi(x) = 1 - e^{-\mu x} = 1 - \phi(1-x)$ and ϕ is the generating function for the Poisson given in (1.4). Thus the expected fraction of surviving lines almost satisfies the same recursion that the survival probability ρ_k does and it follows by induction that $E\hat{Y}^T_{T-k}/N_{T-k} \approx \rho_k$. Our next task is to show that if all of the population sizes are large then the approximation is good and furthermore the fraction observed stays close to its mean.

Theorem. *If $T \to \infty$ and $M \to \infty$ then*

$$\max_{M \leqslant m \leqslant T}\left|\frac{\hat{Y}^T_m}{N_m\rho_{T-m}} - 1\right| \to 0 \quad \textit{in probability.} \tag{6.3}$$

Proof. We begin by computing second moments. Let $q_i = (1 - i/N_m)^{\hat{Y}^n_{m+1}}$ be the probability that i individuals specified in advance are all not chosen by the Y^n_{m+1} lineages in generation $m+1$. Writing Y^T_m as a sum of indicator random variables

$$E((\hat{Y}^n_m)^2|\hat{Y}^n_{m+1}) = N_m(1-q_1) + N_m(N_m-1)[1-(2q_1-q_2)].$$

Subtracting the square of the mean, $[N_m(1-q_1)]^2$, we have

$$\mathrm{var}(\hat{Y}^n_m|\hat{Y}^n_{m+1}) = N_m q_1(1-q_1) + N_m(N_m-1)[q_2 - q_1^2].$$

Simple algebra shows $q_2 < q_1^2$, so the off diagonal terms are negative, and it follows that

$$\mathrm{var}\left(\frac{\hat{Y}^n_m}{N_m}\middle|\hat{Y}^n_{m+1}\right) \leqslant \frac{1}{4N_m}. \tag{6.4}$$

If we let $f_m(x) = 1 - (1 - 1/N_m)^{x\mu N_m}$ then it follows from (6.1) and (6.4) that

$$P\left(\left|\frac{\hat{Y}^n_m}{N_m} - f_m\left(\frac{\hat{Y}^n_{m+1}}{N_{m+1}}\right)\right| > N_m^{-1/3}\right) \leqslant \frac{\mathrm{var}(\hat{Y}^n_m/N_m|\hat{Y}^n_{m+1})}{N_m^{-2/3}} \leqslant \frac{1}{4N_m^{1/3}}. \tag{6.5}$$

Therefore if we define the "good event"

$$G_{M,n} = \left\{\left|\frac{\hat{Y}^n_m}{N_m} - f_m\left(\frac{\hat{Y}^n_{m+1}}{N_{m+1}}\right)\right| \leqslant N_m^{-1/3} \text{ for all } M \leqslant m \leqslant n\right\},$$

we have a good estimate for its failure probability

$$P(G^c_{M,n}) \leqslant \sum_{m=M}^{n-1} \frac{1}{4N_m^{1/3}} \leqslant \frac{1}{4N_M^{1/3}(1-\mu^{-1/3})}. \tag{6.6}$$

To prove (6.3) now we define iterated functions

$$g_{n,m}(x) = f_m(g_{n,m+1}(x)) \quad \text{for } m < n$$

with $g_{n,n}(x) = f_n(x)$ and then write

$$\begin{aligned}\frac{\hat{Y}^n_m}{N_m} - \psi^{n-m}(1) = \frac{\hat{Y}^n_m}{N_m} - f_m\left(\frac{\hat{Y}^n_{m+1}}{N_{m+1}}\right) + f_m\left(\frac{\hat{Y}^n_{m+1}}{N_{m+1}}\right) - f_m(g_{n,m+1}(1)) \\ + g_{n,m}(1) - \psi^{n-m}(1).\end{aligned} \tag{6.7}$$

The first difference on the right is controlled by (6.6). Our next step is to estimate the third difference.

Lemma. *There is a constant C_o so that if $N_m \geqslant 2$ then*

$$\sup_{x\in[0,1]} |f_m(x) - \psi(x)| \leqslant C_o\mu/N_m, \tag{6.8}$$

$$\sup_{x\in[0,1]} |f'_m(x) - \psi'(x)| \leqslant C_o(\mu + \mu^2)/N_m. \tag{6.9}$$

Proof. Expanding $-\ln(1-x) = x + x^2/2 + \cdots$, we have

$$\frac{-\varepsilon^{-1}\ln(1-\varepsilon) - 1}{\varepsilon} \to \frac{1}{2} \quad \text{as } \varepsilon \to 0.$$

Using $1 - x \leqslant \mathrm{e}^{-x}$ it follows that there is a C_o so that

$$0 \leqslant -\varepsilon^{-1}\ln(1-\varepsilon) - 1 \leqslant C_o\varepsilon \quad \text{if } 0 \leqslant \varepsilon \leqslant \tfrac{1}{2}. \tag{6.10}$$

To estimate the difference in (6.8), we observe that $(1 - 1/N_m)^{N_m} \leqslant e^{-1}$ so

$$f_m(x) \geqslant 1 - \exp(-\mu x) = \psi(x).$$

To bound $f_m(x) - \psi(x)$ and hence prove (6.8), we note that

$$\begin{aligned}\mathrm{e}^{-\mu x} - \mathrm{e}^{\mu x N_m \ln(1-1/N_m)} &= \int_x^{-xN_m\ln(1-1/N_m)} \mu\mathrm{e}^{-\mu y}\,\mathrm{d}y \\ &\leqslant \mu x(-N_m\ln(1-1/N_m) - 1) \leqslant C_o\mu x/N_m.\end{aligned}$$

To prove the second result, we note that differentiating the definitions

$$f_m'(x) = -\mu N_m \ln\left(1 - \frac{1}{N_m}\right)\left(1 - \frac{1}{N_m}\right)^{x\mu N_m}$$

and $\psi'(x) = \mu e^{-\mu x}$. Adding and subtracting $\mu f_m(x)$, then using (6.10) and (6.8) we have

$$|f_m'(x) - \psi'(x)| \leqslant \frac{C_o\mu}{N_m} f_m(x) + \mu|f_m(x) - \psi(x)| \leqslant \frac{C_o(\mu + \mu^2)}{N_m}$$

which proves (6.9). □

We are now ready to tackle the third term on the right in (6.7).

Lemma. *If $N_m \geqslant 2$ then*

$$|g_{n,m}(1) - \psi^{n-m}(1)| \leqslant C_o\mu \sum_{k=m}^{n-1} \frac{1}{N_m}. \tag{6.11}$$

Proof. Using the triangle inequality

$$|g_{n,m}(1) - \psi^{n-m}(1)| \leqslant |f_m(g_{n,m+1}(1)) - \psi(g_{n,m+1}(1))| + |\psi(g_{n,m+1}(1)) - \psi(\psi^{n-m-1}(1))|.$$

The first term $\leqslant C_o\mu/N_m$ by (6.8). To estimate the second difference we note that $\psi(x) = 1 - \exp(-\mu x)$ is increasing and concave, so we have $\psi'(x) \leqslant \psi'(\rho) < 1$ for all $x \geqslant \rho$. $\psi^k(1) \downarrow \rho$, the positive fixed point of $\psi(x)$, so $\psi^{n-m-1}(1) \geqslant \rho$. To handle the other term in the second difference, we note that $f_m(x) \geqslant \psi(x)$ so by induction it follows that $g_{n,m+1}(1) \geqslant \psi^{n-m-1}(1) \geqslant \rho$. Combining our observations, the second term is bounded by $|g_{n,m+1}(1) - \psi^{n-m-1}(1)|$ and the result in (6.11) follows by induction. □

It remains to estimate the middle term on the right in (6.7). The first step is to note that results in the proof of (6.11) imply that

We can pick $\delta_o > 0$ so that $\psi'(x) \leqslant 1 - \delta_o$ when $x \in [\rho - \delta_o, 1]$. (6.12)

Thus if we let $\delta_1 = \rho - \phi(\rho - \delta_o)$ which is $< \delta_o$, then for $x \geqslant \rho - \delta_o$,

$$\psi(x) \geqslant \psi(\rho - \delta_0) = \rho - \delta_1. \tag{6.13}$$

Lemma. *If M is large then on the good set of outcomes $G_{M,n}$*

$$\frac{\hat{Y}^n_m}{N_m} \geqslant \rho - \delta_o \quad \textit{for } M \leqslant m \leqslant n. \tag{6.14}$$

Proof. We proceed by induction backwards. The conclusion is trivially true when $m = n$. Suppose now that $\hat{Y}^n_{m+1}/N_{m+1} \geqslant \rho - \delta_o$. Using $f_m \geqslant \psi$, our choice of δ_1 implies that

$$f_{m+1}(\hat{Y}^n_{m+1}/N_{m+1}) \geqslant \rho - \delta_1,$$

so on the good set $G_{M,n}$ we will have

$$\frac{\hat{Y}^n_m}{N_m} \geqslant \rho - \delta_1 - N_m^{-1/3} \geqslant \rho - \delta_o,$$

if M is chosen large enough so that $N_M^{-1/3} \leqslant \delta_o - \delta_1$. □

The next result takes care of the second term in the right in (6.7) and thus will complete the proof of our main result, (6.4).

Lemma. *If M is large then on $G_{M,n}$ we have for $M \leqslant m < n$*

$$\left| f_m\left(\frac{\hat{Y}^n_{m+1}}{N_{m+1}}\right) - f_m(g_{n,m+1}(1)) \right| \leqslant \left| \frac{\hat{Y}^n_{m+1}}{N_{m+1}} - g_{n,m+1}(1) \right| \leqslant \sum_{k=m+1}^{n-1} N_k^{-1/3}. \tag{6.15}$$

Proof. To prove the first inequality, observe that by (6.9) and the choice of δ_o, if M is large then $0 \leqslant f_m'(x) \leqslant 1$ when $x \geqslant \rho - \delta_o$. Using (6.14) now and the triangle inequality it follows that on $G_{M,n}$ if $M \leqslant k < n$ then

$$\left| \frac{\hat{Y}^n_k}{N_k} - f_{n,k}\left(\frac{\hat{Y}^n_{k+1}}{N_{k+1}}\right) \right| + \left| f_k\left(\frac{\hat{Y}^n_{k+1}}{N_{k+1}}\right) - f_k(g_{n,k+1}(1)) \right|$$

$$\leqslant N_k^{-1/3} + \left| \frac{\hat{Y}^n_{k+1}}{N_{k+1}} - g_{n,k+1}(1) \right|.$$

and the desired result follows by induction. □

References

Athreya, K.B., Ney, P.E., 1972. Branching Processes. Springer, New York.

Brookes, A.J., 1999. The essence of SNPs. Gene 234, 177–186.

Drake, J.W., Charlesworth, B., Charlesworth, D., Crow, J.F., 1998. Rates of spontaneous mutation. Genetics 148, 1667–1686.

Ewens, W.J., 1972. The sampling theory of selectively neutral alleles. Theoret. Pop. Biol. 7, 212–220.

Griffiths, R.C., Pakes, A.G., 1988. An infinite-alleles version of the simple branching process. Adv. Appl. Probab. 20, 489–524.

Griffiths, R.C., Tavaré, S., 1994. Sampling theory for neutral alleles in a varying environment. Phil. Trans. Roy. Soc. London B. 344, 403–410.

Hudson, R.R., 1990. Gene genealogies and the coalescent process. In: Futuyama, D., Antonovic, J. (Eds.), Oxford Surveys in Evolutionary Biology, Vol. 1, pp. 1–44.

Jagers, P., 1975. Branching Processes with Biological Applications. Wiley, New York.

Joyce, P., Tavaré, S., 1987. Cycles, permutations, and the structure of the Yule process with immigration. Stoch. Proc. Appl. 25, 309–314.

Kingman, J.F.C., 1982a. The coalescent. Stoch. Proc. Appl. 13, 235–248.

Kingman, J.F.C., 1982b. Exchangeability and the evolution of large populations. In: Koch, G., Spizzichino, F. (Eds.), Exchangeability in Probability and Statistics. North-Holland, Amsterdam, pp. 97–112.

Kruglyak, L., 1999. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nature Gen. 22, 139–144.

Lai, E., Riley, J., Purvis, I., Roses, A., 1998. A 4-Mb high-density single nucleotide polymorphism-based map around human APOE. Genomics 54, 31–38.

Landegren, U., Nilsson, M., Kwok, P.Y., 1998. Reading bits of genetic information: methods for single nucleotide polymorphism analysis. Genome Res. 8, 769–776.

Slatkin, M., Hudson, R.R., 1991. Pairwise comparison of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics 129, 555–562.

Wang, D.G, et al., 1998. Large-scale, identification, mapping, and genotyping of single nucleotide polymorphisms in the human genome. Science 280, 1077–1082.