Dating Rare Mutations from Small Samples with Dense Marker Data

Luke C. Gandolfo,∗†‡ Melanie Bahlo,†‡ and Terence P. Speed†‡§

Abstract

We present a method for estimating the age of a mutation based on the genetic length of ancestral haplotypes shared between individuals carrying the mutation. The method can be reliably applied to small samples, typical of situations involving rare mutations, and makes effective use of modern high-density SNP data, thus overcoming two of the limitations with existing methods. The method provides age estimates and confidence intervals without the use of asymptotic theory, and is applicable to genealogies where the data is independent or correlated. In the correlated case we estimate the correlation directly from the data, rather than relying on a model for the genealogy. To demonstrate the method’s efficacy, we provide simulation results and compare it to other methods. The length data is obtained with a simple procedure, and an R script is available for performing the calculations.

INTRODUCTION

Estimating the age of mutations, i.e. “dating” mutations, has become a fascination for geneticists, particularly since a number of classic studies in the 1990s (e.g. Serre et al. 1990; Risch et al. 1995; Stephens et al. 1998) through to the present-day (e.g. Holm et al. 2011; Fu et al. 2012). This interest has stimulated the development of a number of methods for dating mutations. The main limitations with existing methods, however, are that: (1) they cannot be applied reliably in situations involving rare mutations where sample sizes are typically small; and (2) they cannot make effective use of today’s high-density SNP data (derived from genotyping arrays or genome sequencing).

Common methods for dating mutations fall into three broad categories: methods based on the frequency of the mutation, methods based on gene trees, and methods based on the decay of linkage disequilibrium (LD) around the mutation (see Slatkin and Rannala 2000;
Rannala and Bertorelle 2001; Colombo 2008). Frequency methods model the way a mutation’s population frequency changes over time, and use the mutation’s current frequency to estimate its age. The most well-known method of this sort is that proposed by Kimura and Ohta (1973), which assumes a constant population size and an absence of natural selection. Other frequency methods include Thompson (1976) and Griffiths and Tavaré (1998). Frequency methods are not well suited to small sample situations involving rare mutations since obtaining an accurate frequency estimate in such situations is difficult.

Gene tree methods infer mutation age from a sample of mutation bearing chromosomes using observed variation in sequence or marker data (at polymorphic loci) completely linked to the mutation. A mutation model is proposed for the process which generates the observed variation (e.g. a Poisson process model with an infinitely many sites assumption), and a model (e.g. a coalescent process) is proposed for the genealogy of the sample (e.g. Thomson 1998; Griffiths and Tavaré 1999; Markovtsova et al. 2000) or the genealogy is inferred directly using distance or maximum parsimony methods (e.g. Morral et al. 1994). The observed variation (presumed to have arisen over time) is then used to infer the mutation’s age. Gene tree methods are not well suited to dating rare mutations using SNP data. Rare mutations are relatively young, and so sufficient time will usually not have elapsed since the mutation’s creation to generate the required variation (see Rannala and Bertorelle 2001). If enough time has elapsed for variation to arise, this variation will typically not be observable in SNP data (sequence or rapidly evolving microsatellite data would be the only viable options).

Methods based on LD decay make use of the fact that the mutation will be in LD with nearby marker alleles at polymorphic loci, and the decay of this LD over time (through recombination) provides information about the mutation’s age. The most well-known method of this sort is given in Serre et al. (1990). Variations on the Serre method include those of Stephens et al. (1998) and Reich and Goldstein (1999). Serre’s method (and its variations) estimates a mutation’s age by using information from a single marker locus. In order to utilize information available from other marker loci, some have applied the Serre method to each locus separately, then combined each of the resulting age estimates into a single estimate, e.g. by averaging (see Holm et al. 2011). The efficacy of the Serre approach ultimately rests on the asymptotic properties of the estimate. We will show, however, that these properties do not hold for small samples.

Methods have also been proposed which combine the frequency approach with one of the other approaches: some methods combine the frequency and gene tree approaches (e.g. Slatkin and Rannala 1997; Stephens 2000), and some methods combine the frequency and LD approaches (e.g. Rannala and Reeve 2003). Since these methods require the mutation’s frequency, they have the same problem mentioned above for plain frequency methods: they are not suited to small sample situations.

An alternative LD approach is based on the following biological picture: present day mutation “carriers” will have inherited the mutation as segments of the ancestral chromosome which originally acquired the mutation, where these segments will differ in size due
to random recombination events (Figure 1). These inherited ancestral segments (harboring the mutation) decrease in size over time, and so their lengths can be used to estimate the mutation’s age. Genin et al. (2004) and McPeek and Strahs (1999) (originally developed for fine mapping, but also provides a method for dating mutations) are two methods based on this approach. While the Genin method is applicable to small samples, it can only utilize data from highly polymorphic loci (e.g. microsatellite loci), not high-density SNP data. The McPeek and Strahs method is also unsuited to high-density SNP data, not because it cannot be applied to SNP loci, but because the method “infers” the ancestral segment lengths using a hidden Markov model, and the required computations rapidly become impractical as the data density increases to today’s standards. The McPeek and Strahs method also makes extensive use of asymptotic techniques. However, we will show that these techniques are not effective for small samples. While others have studied the lengths of ancestral segments segregating in a population (e.g. Stam 1980; Chapman and Thompson 2002; and Baird et al. 2003), the two methods just discussed are the only approaches to use the lengths of segments harbouring a mutation to estimate the mutation’s age.

The method we propose, which we call the Gamma method, is also based on ancestral segment lengths. Unlike the methods above, however, we utilize high-density SNP data, inferring the segment lengths with a simple procedure. The method provides age estimates and confidence intervals without the use of asymptotic theory, making it applicable to

Figure 1 Biological basis for alternative LD dating methods. A mutation $m$ is created on an ancestral chromosome. After a period of time, what remains in a number of carriers are different sized segments of the ancestral chromosome.
small samples and thus suitable for typical studies involving rare mutations. To establish the method’s efficacy, we provide simulation results and compare it to other LD methods.

We make two important comments before proceeding. Firstly, all methods considered below are concerned with dating mutations on recombining parts of the genome (i.e. not mitochondrial and Y chromosome mutations), where the term “mutation” refers to any kind of variation of DNA sequence, i.e. a non-synonymous substitution, an insertion, or a deletion (care may be needed when dating large insertions or deletions). Secondly, the term “mutation age” has become ambiguous in the literature: it can mean the time since the mutation first came into existence, or it can mean the time since the most recent common ancestor (MRCA) of the sample (which is less than or equal to the time in the first sense). The term is near universally used to refer to the time since the MRCA, since the majority of methods estimate the age in that sense, and so we will also use the term this way. Often the aim is to estimate age in the MRCA sense anyway (e.g. cases involving a founder effect). However, if the mutation arose in the population from which a random sample is taken, then estimating the MRCA age will be approximately equivalent to estimating the time since the mutation came into existence.

**Materials and Methods**

**Age estimation and confidence intervals**

We begin with a statistical model describing the genetic length $l$ of an ancestral segment (measured in Morgans) and the way that length changes over time (measured in generations). Assuming only the Haldane recombination model, the length of the left and right “arms” of the ancestral segment (measured from the mutation) both have independent Exponential($\tau$) distributions, and so

$$l \sim \text{Gamma}(2, \tau), \quad (1)$$

where $\tau$ denotes the age of the mutation (see Appendix A). The same result was obtained by McPeek and Strahs (1999), by other means. This model implies that the mean ancestral segment length is $2/\tau$, i.e. the mean length decreases over time, consistent with our expectations. The essential idea of our approach is to obtain ancestral segment lengths from a sample of “unrelated” individuals carrying the mutation, then estimate the $\tau$ parameter in (1) by maximum likelihood. In reality the sampled individuals are not unrelated, since the mutation is tacitly assumed to be identical by descent, and so the term “unrelated” is used here to mean individuals who are separated by a reasonable number of meiosis events, e.g. the individuals should not be siblings or close cousins. We develop two versions of the method: one which applies to samples assumed to have an independent genealogy (i.e. sampled individuals have independent recombinational histories since their MRCA), and one which applies to samples assumed to have a correlated genealogy (i.e. a “tree-like”
genealogy where subsets of the sample have common ancestry earlier than the MRCA for the entire sample, thus producing correlated observations).

Firstly, assume that the sample has an independent genealogy. In that case, the segment lengths \( l_1, l_2, \ldots, l_n \) of a sample of size \( n \) are independent and identically distributed with \( l_i \sim \text{Gamma}(2, \tau) \), from which it follows that the maximum likelihood estimate (MLE) of \( \tau \) is

\[
\hat{\tau}_{\text{MLE}} = \frac{2n}{\sum_{i=1}^{n} l_i}. \tag{2}
\]

The MLE is a biased estimate of \( \tau \), particularly for small samples, and so for our purposes it is desirable to remove this bias. Properties of the Gamma distribution imply that

\[
\hat{\tau}_{\text{MLE}}^{-1} \sim \text{Gamma}(2n, 2n\tau), \tag{3}
\]

from which we can obtain the minimum variance unbiased estimate of \( \tau \), given by

\[
\hat{\tau} = \frac{b(n)2n}{\sum_{i=1}^{n} l_i}, \tag{4}
\]

where \( b(n) = (2n - 1)/2n \) is the bias correction factor. Properties of the Gamma distribution also imply that

\[
\tau/\hat{\tau} \sim \text{Gamma}(2n, 2nb(n)), \tag{5}
\]

i.e. \( \tau/\hat{\tau} \) is a pivotal quantity from which exact confidence intervals for \( \tau \) are easily derived (see Appendix B).

Now assume that the sample has a correlated genealogy. In this case, the segment lengths \( l_1, l_2, \ldots, l_n \) are dependent random variables. We assume that all lengths have the same pairwise correlation \( \rho \), i.e. \( \text{Corr}(l_i, l_j) = \rho \) for all \( i \neq j \). Note that this assumption is far weaker than it appears, since for our purposes \( \rho \) will act (roughly) as an average or summary of the correlation present amongst the observations. To estimate \( \tau \) in this setting we begin indirectly by estimating \( \tau^{-1} \) with

\[
\hat{\tau}^{-1} = \frac{\sum_{i=1}^{n} l_i}{2n}, \tag{6}
\]

since this remains an unbiased estimate of \( \tau^{-1} \) regardless of the dependence amongst the \( l_i \) variables. We now seek the distribution of \( \hat{\tau}^{-1} \). Inspired by the result in the case involving independent variables, i.e. result (3), we suppose that \( \hat{\tau}^{-1} \) has approximately a \( \text{Gamma}(\alpha, \beta) \) distribution, for some \( \alpha \) and \( \beta \). To obtain expressions for \( \alpha \) and \( \beta \), we first observe (see Appendix C) that the mean and variance of \( \hat{\tau}^{-1} \) is

\[
\mathbb{E}(\hat{\tau}^{-1}) = \tau^{-1} \quad \text{and} \quad \mathbb{V}(\hat{\tau}^{-1}) = \frac{1 + (n - 1)\rho}{2n\tau^2}, \tag{7}
\]
and so equating these expressions to the mean and variance of a Gamma(\(\alpha, \beta\)) random variable, i.e. \(\alpha/\beta\) and \(\alpha/\beta^2\), respectively, we easily obtain
\[
\alpha = \frac{2n}{1 + (n-1)\rho} \quad \text{and} \quad \beta = \frac{2n\tau}{1 + (n-1)\rho}.
\]
Thus, we arrive at the following distribution for \(\hat{\tau}^{-1}\):
\[
\hat{\tau}^{-1} \sim \text{Gamma}(2n^*, 2n^*\tau),
\]
where
\[
n^* = \frac{n}{1 + (n-1)\rho}.
\]
We refer to this as the Gamma approximation. We use the \(n^*\) notation to simultaneously highlight the similarity with the case under independence, i.e. result (3), and to suggest an enlightening interpretation for the effect of the dependence: it effectively reduces the sample size, i.e. the estimate is calculated with a sample of size \(n\) but behaves as if it came from a smaller sample of size \(n^*\) (since \(n^* \leq n\)). We emphasize that (9) is not an asymptotic result valid only for large \(n\); it is an approximation valid for all sample sizes.

Given the Gamma approximation, by using arguments similar to those leading to (4) and (5) above (described in Appendix B), we arrive at the following further results. Firstly,
\[
\hat{\tau} = \frac{b(n^*)2n}{\sum_{i=1}^{n} l_i}
\]
is an unbiased estimate of \(\tau\), where \(b(n^*)\) is the same bias correction factor as in the case under independence, except that \(n\) is now replaced by \(n^*\) in the formula. Secondly,
\[
\tau/\hat{\tau} \sim \text{Gamma}(2n^*, 2n^*b(n^*)),
\]
from which confidence intervals for \(\tau\) are easily obtained.

The correlation \(\rho\) can be estimated directly from the sample of lengths \(l_1, l_2, \ldots, l_n\) using the following estimate:
\[
\hat{\rho} = \frac{nl\bar{l}^2 - S^2(2n + 1)}{nl\bar{l}^2 + S^2(n - 1)},
\]
where \(\bar{l}\) and \(S^2\) are the sample mean and variance, respectively (see Appendix D). While \(\hat{\rho} \in [-1, 1]\), some values of \(\hat{\rho}\) produce inappropriate values of \(n^*\), and in that event the calculation of \(n^*\) is adjusted accordingly (see Appendix D).

Obtaining the data

To identify ancestral segments in sampled individuals, we apply the following procedure: line up high density SNP data around the mutation locus, look for continuous haplotype
Figure 2  Haplotype sharing in sampled individuals. The green regions represent continuous haplotype sharing between two or more individuals after the data has been lined up around the mutation \( m \). These regions are identified as ancestral segments. While the two longest arms on either side of the mutation stop sharing at the same point, typically the true ancestral segment in one of the individuals will continue some distance further, as represented by the pink sections (each pink section could belong to the other longest arm, but this is of no concern in our analysis). Since these extra sections cannot be detected by the sharing criterion, they are estimated from the data.

sharing between \textit{two or more} individuals, then take this sharing to define the ancestral segments (Figure 2). The segment lengths are obtained from the genetic map positions of the outermost markers. This procedure is reasonable since such sharing will typically involve hundreds of consecutive markers, which can only be explained if the shared haplotypes are inherited from a common ancestor, i.e. if they are ancestral segments.

To assess haplotype sharing for individuals with a \textit{single} copy of the mutation (e.g. with a dominant mutation), the data needs to be \textit{phased} in order to resolve the single underlying ancestral haplotype. Similarly, for individuals with copies of two distinct mutations at the same gene locus (i.e. with a compound heterozygous mutation), the data also requires phasing to reveal the two distinct ancestral haplotypes (both mutations can then be dated). Phasing is easily achieved if parental data is collected, or with modern statistical techniques (see review by Browning and Browning 2011). We note that other LD methods also require data to be phased in these situations. However, for individuals with \textit{two} copies of an identical recessive mutation, the situation is particularly simple: since such individuals inherit two ancestral haplotypes via an inbreeding loop, producing homozygosity along the haplotype, continuous haplotype sharing becomes equivalent to continuous sharing of \textit{identical homozygous markers}, i.e. phasing is not required (if either ancestral haplotype shortens while in the inbreeding loop, then the shorter copy will be revealed by this criterion).
Since haplotype sharing is only defined between two or more individuals, the two longest arms on either side of the mutation stop sharing at the same point, but the true ancestral segment in one of the individuals will almost always continue some distance further (Figure 2). This extra length cannot be detected by the sharing criterion, and so results in a biased age estimate, with the bias being more pronounced for small samples. To correct this, the missing length is estimated from the data. Let \( a_{\text{missing}} \) denote the length of one of these missing arm sections. Since the ancestral segment arm lengths have an Exponential(\( \tau \)) distribution, and since the Exponential distribution is memoryless, it follows that the distribution of \( a_{\text{missing}} \) is Exponential(\( \tau \)). Thus, we can estimate \( a_{\text{missing}} \) by estimating its expected length, which is \( \tau^{-1} \). Using the inferred arm lengths \( a_1, a_2, \ldots, a_{2n} \), we can estimate \( \tau^{-1} \) using

\[
\hat{\tau}^{-1} = \hat{a}_{\text{missing}} = \frac{\sum_{i=1}^{2n} a_i}{2n},
\]

which, apart from the two truncated observations, is an unbiased estimate of \( \tau^{-1} \), regardless of whether the data is independent or correlated. Thus, to correct the bias problem, the sum of the lengths in the age estimate formulas (4) and (11), i.e. \( \sum_{i=1}^{n} l_i \), is replaced with

\[
L = 2 \cdot \hat{a}_{\text{missing}} + \sum_{i=1}^{n} l_i,
\]

where the \( 2 \cdot \hat{a}_{\text{missing}} \) accounts for the two missing lengths.

There are three other factors which can effect the length data, obtained using the procedure above. Firstly, chance haplotype sharing is possible at the very end of each segment, having the effect of inflating the inferred lengths. If high density data with a range of allele frequencies is used, the amount of chance sharing is extremely small. In most circumstances the effect is unnoticeable; a small effect only becomes noticeable when the mutation is very old. Nevertheless, we build in a simple correction which covers these situations (see Appendix E). Secondly, the effect of mutation and errors in the marker data is unnoticeable. Modern marker data (i.e. SNP data) has low mutation rates, and modern genotyping methods (e.g. SNP arrays) produce highly reliable genotype calls of marker loci. Moreover, with the density of modern data, mutations and errors are nearly always obvious: continuous sharing interrupted by one of these features will typically be followed by another large run of continuous sharing. These features may not be so obvious near segment ends, but in that case they have only a minor influence. Finally, in correlated genealogies there is a possibility that some sampled individuals have extra haplotype sharing (around the ancestral haplotype) as a result of common ancestry more recent than the entire sample. Provided the mutation is not too young and/or the sample is not very large, the chances of a random sample including individuals with common ancestry recent enough to produce this problem are low. For correlated genealogies, we assume that recombination has removed any excess sharing of this sort in individuals who have common ancestry more recent than the entire sample.
Simulation and comparison

To evaluate the Gamma method, we apply it to ancestral segment lengths simulated using SNP data from HapMap 3 (The International HapMap 3 Consortium 2010). Phased chromosome 1 data from the CEPH population (Utah residents with ancestry from northern and western Europe) was used, with ancestral segment lengths simulated according to the Haldane recombination model around a nominated mutation locus. We selected loci which overlapped those targeted by a standard genotyping array (an Illumina 1M array); this represented \( \frac{2}{3} \) or \( \sim 68 \)K loci targeted by the array on that chromosome.

To generate a sample with an independent genealogy, each ancestral segment is independently simulated using the same HapMap chromosome (designated the ancestral chromosome), then the resulting segments are substituted into a set of unrelated HapMap chromosomes at corresponding loci. For correlated genealogies, we wanted to simulate under weak assumptions and so we use a simple forward simulation approach. To generate a sample with a correlated genealogy we begin with a single HapMap chromosome (designated the ancestral chromosome), then ancestral segments are simulated as follows: (1) extant segments in each generation produce a Poisson distributed (rate \( \lambda = 2.5 \)) number of offspring; (2) each segment is then passed on to its offspring with probability \( \frac{1}{2} \), each time undergoing a round of recombination; (3) this is continued for the required number of generations, then a random sample of segments is taken from the resulting population and substituted into a set of unrelated HapMap chromosomes at corresponding loci. For both the independent and correlated cases, the ancestral segment lengths are inferred from the resulting chromosomes using the procedure described above. Then the method is applied to these inferred lengths, using the version of the method which applies to the assumed kind of genealogy.

We compare our approach to other LD methods. We do not include frequency or gene tree methods in our comparison for two reasons: (1) our approach is based solely on LD (the mutation’s frequency is not part of our data); (2) our concern is dating rare mutations from small samples using dense marker data (frequency and gene tree methods are not suited to these situations, as explained above). We compare to the McPeek and Strahs (1999) method, since this is the only method based on haplotype lengths which can be applied to dense marker data, and we compare to the Serre et al. (1990) method, since this is the paradigm LD decay method and by far the most widely used.

Using a sample of ancestral segment lengths \( l_1, l_2, \ldots, l_n \), the McPeek and Strahs age estimate (for both independent and correlated genealogies) is given by

\[
\hat{\tau} = \frac{2n}{\sum_{i=1}^{n} l_i},
\]  

which is the MLE of \( \tau \) in the Gamma model for ancestral segment length. Since the data density makes the use of their software unfeasible, we use the lengths inferred with our
procedure. They construct confidence intervals using asymptotic MLE theory: the result that
\[
\frac{\hat{\tau} - \tau^{-1}}{\text{se}} \sim N(0,1) \quad (\text{as } n \to \infty)
\]
(16)
is used for independent genealogies, where \(\text{se} = 1/(\sqrt{2n\hat{\tau}})\) is the estimated Fisher information for \(\tau^{-1}\), and the result that
\[
\frac{\hat{\tau} - \tau^{-1}}{\text{se} \cdot f} \sim N(0,1) \quad (\text{as } n \to \infty)
\]
(17)
is used for correlated genealogies, where \(\text{se}\) is the same and \(f = \sqrt{1 + (n-1)p}\), which assumes that all lengths have identical pairwise correlation \(p\). They model correlated genealogies using a conditional coalescent process. Under that model they show that
\[
\rho = \frac{2(n-2)!(n+1) \times \sum_{i=0}^{\infty} (-1)^i \binom{n+i-1}{n-k}^{-1}}{(n-1)(n-k+1)(n-k+2)(n-k)(k-1)!/(n-k-2)!}.
\]
(18)
The Serre age estimate (regardless of the assumed genealogy) is given by
\[
\hat{\tau} = \frac{1}{\ln(1-\theta)} \ln \left( \frac{\hat{f}_{ma} - f_\alpha}{1 - f_\alpha} \right),
\]
(19)
where \(\theta\) is the recombination frequency between the mutation \(m\) and a marker allele \(\alpha\) (assumed to be ancestral, i.e. linked when the mutation was created) at a polymorphic locus, \(f_\alpha\) is the population frequency of \(\alpha\) on normal chromosomes (assumed known and constant over time), and \(\hat{f}_{ma}\) is the proportion of sampled mutant chromosomes with \(\alpha\) at the locus. For sensible estimates we require \(\hat{f}_{ma} < 1\) (otherwise \(\hat{\tau} = 0\)), and we require \(\hat{f}_{ma} > f_\alpha\) (otherwise \(\hat{\tau}\) is not a real number). For the single locus version of the method we apply this formula to randomly selected loci, and for the multi-locus version we apply the formula to every locus where it can be applied then average the results. To be as fair as possible, we only use loci within the bounds of the longest segment arms (alleles at these loci are the only ones guaranteed to be ancestral), and the proportion \(\hat{f}_{ma}\) is calculated using the knowledge of which allele is ancestral. In practice, the point at which loci should stop being used to estimate age will be unknown, and it will not always be obvious which of the alleles at a particular locus is ancestral (particularly with small samples). There are currently no proposals in the literature for calculating confidence intervals which take into account all sources of variation. However, for the single locus version assuming an independent genealogy, since \(\hat{f}_{ma}\) is the MLE under binomial sampling, confidence intervals can be calculated using the delta method:
\[
\frac{\hat{\tau} - \tau}{\text{se}} \sim N(0,1) \quad (\text{as } n \to \infty)
\]
(20)
\(^1\)This is the corrected expression reported in errata to their paper.
where
\[
\hat{s}_e = \frac{1}{\ln(1 - \theta)(\hat{f}_{ma} - f_\alpha)} \left| \sqrt{ \frac{\hat{f}_{ma}(1 - \hat{f}_{ma})}{n}} \right|.
\]

Bootstrap methods for confidence intervals cannot be employed effectively, since the requirement that \( \hat{f}_{ma} > f_\alpha \) places significant restrictions on the loci which could be used for dating: we would require \( f_\alpha > 1/n \) to guarantee that \( \hat{\tau} \) can be calculated for all possible bootstrap resamples. As far as we know, this is the first reported simulation study of the Serre method.

We evaluate performance of the methods by calculating the mean of the point estimates (to assess bias), nominal confidence interval coverage (we nominated 95% coverage), and confidence interval length (which is related to the variance of the estimate, and of relevant practical interest). By assessing performance on these dimensions, we capture even more than would be learned from the MSE (mean square error) of the estimators.

**RESULTS**

We show simulation results for the true age \( \tau = 50 \) generations (\( \approx 1,000-1,500 \) years old, assuming 20-25 year average generation span), with 1,000 replicates for each of six sample sizes \( n = 5, 10, 15, 20, 25, 30 \) (Figure 3).

For both independent and correlated genealogies, the Gamma method performs well: the age estimate is unbiased and the confidence interval has the desired coverage. The only exception is for independent genealogies with sample size \( n = 5 \), where there is a very small upward bias and a slight decrease in confidence interval coverage. Importantly, for small samples sizes, the Gamma method performs better than the McPeek and Strahs and Serre methods: (1) for small samples the age estimates of these methods have a large bias, and are only without bias for \( n > 30 \); (2) the methods do not achieve the desired confidence interval coverage (an exception being the McPeek and Strahs method for \( n \geq 30 \) in the independent genealogy case); (3) the confidence interval length for the Gamma method is typically much smaller than for the other methods. Similar results are obtained for \( \tau = 20 \) and \( \tau = 100 \) (see Figures S1 and S2).

The primary reason the McPeek and Strahs method does not perform as well, for these sample sizes, is its reliance on the asymptotic properties of the MLE used to estimate age. To compound this problem, for correlated genealogies the correlation is not accurately estimated by their formula, which explains why the confidence intervals in that case do not achieve the correct coverage. Similarly, the Serre method does not perform well with independent genealogies because it relies on asymptotic properties of the age estimate. The performance of the Serre method for correlated genealogies is not totally unexpected, since it has already been reported that the method is biased in this case (see Slatkin and Rannala 2000). We have now shown, however, that there is almost as much bias for the case involving independent genealogies.
Figure 3  Simulation results for $\tau = 50$. Top row (A,B,C) shows results for independent genealogies, and bottom row (D,E,F) shows results for correlated genealogies: (A,D) shows the mean of the estimates (horizontal grey line marks $\tau = 50$, and error bars represent 2 standard errors); (B,E) shows confidence interval coverage (horizontal grey line marks 95% coverage); and (C,F) shows boxplots for confidence interval length.
The Gamma method overcomes two of the limitations with existing methods for dating mutations: (1) the method can be reliably applied in small sample situations, making it ideal for studies involving rare mutations; (2) the method makes effective use of modern high-density SNP data. Since the method is not based on asymptotic theory, it performs better than other methods for small samples. The method provides estimates and confidence intervals for both independent and correlated genealogies. This is an advantage over other LD methods where confidence intervals are either not provided or only provided under the assumption of an independent genealogy (the McPeek and Strahs method is a notable exception). For correlated genealogies, since the method estimates the correlation directly from the data, instead of relying on idealized models of the genealogy (e.g. coalescent models), it performs better in those situations. The fact that the method does not depend on a population genetics model for the genealogy represents an advantage when applying the method in practice.

The Gamma method has other advantages. Firstly, the method not does require knowledge of the population growth rate or selective forces acting on the mutation, two factors which can significantly influence other approaches. Population growth influences the structure of correlated genealogies, but since the correlation is estimated directly from the data in the Gamma method, it can take these effects into account. Natural selection can effect a mutation’s frequency, but there is no plausible mechanism whereby it can effect the lengths of ancestral segments (other LD methods are similarly unaffected). Some methods take these factors into account (e.g. Slatkin and Rannala 1997; Rannala and Reeve 2003), requiring them as user specified parameters, but this can be a disadvantage in practice, since their magnitude is often poorly known. We note here that some have inferred the presence of selection by comparing a frequency estimate and an LD estimate of age (e.g. Stephens et al. 1998), the former being affected by selection while the latter is not. It is also worth speculating that LD age estimates might provide a basis for making inferences about a mutation’s frequency; LD is related to age, and age is related to frequency.

Secondly, the Gamma method has some flexibility regarding the data it can be applied to. Unlike some LD methods, the Gamma method is not limited to one data type, e.g. it can be applied to microsatellite data. However, while the typical heterozygosity of such data essentially eliminates the problem of chance sharing (i.e. inferring inflated lengths), depending on the size of the gaps between marker loci, termination points of ancestral segments might only be obtained approximately, leading to approximate age estimates. Depending on the aim of the study, however, approximate results may be sufficient. It is the sheer density of modern SNP data which essentially eliminates the problem of chance sharing, despite the low heterozygosity of the loci, and allows ancestral segment lengths to be obtained with precision. Moreover, unlike other LD methods, marker allele frequencies are not necessarily required for the Gamma method: allele frequencies are only needed when correcting for subtle chance sharing effects in certain situations (see Appendix E), but this
correction is optional. This can be an advantage in situations where allele frequencies are not always available (e.g. genome sequencing data).

One large disadvantage of the Gamma method, like all other LD methods, is that it is not useful for dating ancient mutations where the LD surrounding the mutation has decayed so much as to make dating infeasible. In such cases, frequency or gene tree methods are the only viable options.

A simple R script is available\(^2\) which performs the calculations of the Gamma method given a sample of ancestral segment lengths as input, obtained using the straightforward procedure described above. No extensive computations are required. The Gamma method approach has been successfully applied in Corbett et al. (2011), Lomax et al. (2013), and Lim et al. (2014).

**ACKNOWLEDGEMENTS**

We thank Helen Lindsay for several helpful discussions concerning the simulation study.

**LITERATURE CITED**


McPeek, M. S. and A. Strahs, 1999 Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. The American Journal of Human Genetics 65: 858-875.


Appendices

Appendix A: Derivation of the Gamma model

Our derivation assumes the Haldane recombination model: the genetic length between adjacent recombination events on a single chromosome is Exponentially distributed with rate 1 (where this rate is set by definition). No further assumptions are required.

Let $a_l$ and $a_r$ denote the lengths of the left and right ancestral segment “arms” measured from the mutation. Suppose that a mutation has been created on the ancestral chromosome, and consider subsequent recombination events to the right of the mutation. Let $D_1, D_2, \ldots, D_\tau$ denote the distances from the mutation to the next recombination point (to the right) in each of the $\tau$ generations since the mutation’s creation. Since the Exponential distribution is memoryless, the Haldane model implies that $D_1, D_2, \ldots, D_\tau$ are Exponential(1) variables. These variables are independent, since the recombination process in each generation occurs independently.

Now observe that the right arm length $a_r$ is the minimum of the $D_1, D_2, \ldots, D_\tau$ variables; since the ancestral segment never gets longer, the recombination “bite” closest to the mutation occurring over the $\tau$ generations, i.e. the minimum of the $D_i$’s, marks the extent of the ancestral segment. That is, $a_r = \min\{D_1, D_2, \ldots, D_\tau\}$. But, the minimum of $k$ independent Exponential($\lambda$) variables has an Exponential($k\lambda$) distribution. Thus, $a_r \sim$ Exponential($\tau$). A similar argument shows that $a_l \sim$ Exponential($\tau$).

The $a_l$ and $a_r$ are independent, which follows from the memoryless property. Hence, since the sum of two independent Exponential($\lambda$) variables has a Gamma(2, $\lambda$) distribution, and since $l = a_l + a_r$, it follows that $l \sim$ Gamma(2, $\tau$).

Appendix B: MLE, bias correction factor, and confidence intervals

Under the assumption that a sample of lengths $l_1, l_2, \ldots, l_n$ is independent and identically distributed with $l_i \sim$ Gamma(2, $\tau$), it is straightforward to show that the MLE of $\tau$ is

$$\hat{\tau}_{\text{MLE}} = \frac{2n}{\sum_{i=1}^{n} l_i}.$$

To calculate the bias of this estimate and to calculate confidence intervals for $\tau$, we make use of the following fact: if $X_1, X_2, \ldots, X_n$ are independent and identically distributed

---

3The Haldane model does not take genetic interference into account, and consequently is not the most accurate model of recombination. Despite this, the model will be a good approximation for our purposes, since the genetic distances that will concern us will be short enough that multiple recombination events will be extremely unlikely within those distances.

4This amounts to assuming that each ancestral segment is the same age, i.e. that the $\tau$ associated with each segment is the same, an assumption which should hold, at least approximately, in practice.
Gamma($k, \lambda$) variables, then
\[ \sum_{i=1}^{n} \psi X_i \sim \text{Gamma}(nk, \lambda/\psi). \] (22)

To calculate the bias of $\hat{\tau}_{\text{MLE}}$, observe that fact (22) implies
\[ \hat{\tau}_{\text{MLE}}^{-1} \sim \text{Gamma}(2n, 2n\tau), \] (23)

and from this a routine calculation shows
\[ E(b(n)\hat{\tau}_{\text{MLE}}) = \tau, \]
where $b(n) = (2n - 1)/2n$ is the bias correction factor. Thus,
\[ \hat{\tau} = b(n)\hat{\tau}_{\text{MLE}} = \frac{b(n)2n}{\sum_{i=1}^{n} l_i} \]
is an unbiased estimate of $\tau$, and since $\hat{\tau}$ is a function of a sufficient statistic for $\tau$, the Rao-Blackwell theorem implies that it has the smallest variance amongst all unbiased estimates.

To obtain confidence intervals for $\tau$ we again make use of fact (22). Reasoning identical to that leading to (23), leads to
\[ \hat{\tau}^{-1} \sim \text{Gamma}(2n, 2n\tau b(n)). \]

Multiplying by $\tau$ and applying fact (22) once more, this time with $\psi = \tau$, we obtain
\[ \tau/\hat{\tau} \sim \text{Gamma}(2n, 2nb(n)), \]
showing that $\tau/\hat{\tau}$ is a pivotal quantity. Thus, letting $\gamma_{n;\alpha}$ denote the $\alpha$ percentile of the Gamma($2n, 2nb(n)$) distribution, we easily obtain $1 - \alpha$ confidence intervals:
\[ P(\hat{\tau}\gamma_{n;\alpha/2} \leq \tau \leq \hat{\tau}\gamma_{n;1-\alpha/2}) = 1 - \alpha. \]

This confidence interval is exact: it is not an asymptotic statement, i.e. it has coverage $1 - \alpha$ regardless of the sample size. Note that no penalty has been suffered by calculating the interval with the unbiased estimate (e.g. the interval is not wider). The scaling properties of the Gamma distribution make the interval invariant to the choice of $b(n)$.

Appendix C: Parameters for the Gamma approximation

We first calculate the mean and variance of the estimate $\hat{\tau}^{-1}$. The mean is
\[ E(\hat{\tau}^{-1}) = E\left( \frac{\sum_{i=1}^{n} l_i}{2n} \right) = \frac{1}{2n} \frac{2n}{\tau} = \tau^{-1}, \] (24)
since $E(l_i) = 2/\tau$. To calculate the variance, observe the equivalence

$$
\hat{\tau}^{-1} = \frac{\sum_{i=1}^{n} l_i}{2n} = \frac{1}{2n} 1_n^T L,
$$

(25)

where $1_n^T = [1, 1, \ldots, 1]$ and $L^T = [l_1, l_2, \ldots, l_n]$. Observe also that, given the assumed correlation structure, i.e. $\text{Corr}(l_i, l_j) = \rho$ for $i \neq j$, the covariance matrix for $L$ is

$$
\mathbb{V}(L) = \sigma^2 \begin{bmatrix}
1 & \rho & \ldots & \rho \\
\rho & \ddots & \ddots & \vdots \\
\vdots & \ddots & \ddots & \rho \\
\rho & \ldots & \rho & 1
\end{bmatrix}.
$$

(26)

Now recall the following standard result: if $X$ is a random vector and $a$ is vector of constants of the same length, then $\mathbb{V}(a^T X) = a^T \mathbb{V}(X) a$. Using this result with (25) and (26), it follows that

$$
\mathbb{V}(\hat{\tau}^{-1}) = \mathbb{V} \left( \frac{1}{2n} 1_n^T L \right)
= \frac{1}{4n^2} 1_n^T \mathbb{V}(L) 1_n
= \sigma^2 \frac{(1 + (n - 1)\rho)}{4n}
= \frac{1 + (n - 1)\rho}{2n\tau^2},
$$

(27)

since $\sigma^2 = \mathbb{V}(l_i) = 2/\tau^2$.

We now equate expressions (24) and (27) to the mean and variance of a Gamma($\alpha, \beta$) random variable, i.e. $\alpha/\beta$ and $\alpha/\beta^2$, respectively, to obtain

$$
\alpha = \frac{2n}{1 + (n - 1)\rho} \quad \text{and} \quad \beta = \frac{2n\tau}{1 + (n - 1)\rho},
$$

after some elementary algebra, which are the required parameters for the Gamma approximation.

**Appendix D: Correlation estimate**

We first obtain the following more general results:

$$
\mathbb{E} \left( n\overline{X}^2 \right) = n\mu^2 + \sigma^2 (1 + (n - 1)\rho),
$$

(28)

$$
\mathbb{E} \left( \sum_{i=1}^{n} (X_i - \overline{X})^2 \right) = (n - 1)\sigma^2 (1 - \rho),
$$

(29)
where \( X_1, X_2, \ldots, X_n \) are identically distributed random variables with common mean \( \mu \), variance \( \sigma^2 \), and common pairwise correlation \( \rho \), where \( \bar{X} \) is their average.

We begin by noting that \( X = [X_1, X_2, \ldots, X_n]^T \) has mean \( \mu \mathbf{1}_n \) and covariance matrix

\[
\mathbb{V}(X) = \begin{bmatrix}
\sigma^2 & \gamma & \cdots & \gamma \\
\gamma & \ddots & \ddots & \vdots \\
\vdots & \ddots & \ddots & \gamma \\
\gamma & \cdots & \gamma & \sigma^2
\end{bmatrix},
\]

where \( \gamma = \sigma^2 \rho \). Observe that \( \sigma^2 + (n-1)\gamma \) is an eigenvalue of \( \mathbb{V}(X) \) with eigenvector \( \mathbf{1}_n \), and so the symmetric and idempotent matrix \( \mathbf{G} = \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T \) is an orthogonal projector onto the space \( \text{Span}\{\mathbf{1}_n\} \). Also observe that \( \sigma^2 - \gamma \) is an eigenvalue, but with algebraic multiplicity \( n-1 \), and that the symmetric and idempotent matrix \( \mathbf{I} - \mathbf{G} \) is a projector onto the orthogonal complement of \( \text{Span}\{\mathbf{1}_n\} \), i.e. the corresponding eigenspace. Thus, given that \( \mathbb{V}(X) \) is symmetric, its spectral decomposition is

\[
\mathbb{V}(X) = (\sigma^2 + (n-1)\gamma)\mathbf{G} + (\sigma^2 - \gamma)(\mathbf{I} - \mathbf{G}).
\] (30)

Next, observe that \( \mathbf{G}X \) and \( (\mathbf{I} - \mathbf{G})X \) are orthogonal, and therefore uncorrelated:

\[
\text{Cov}(\mathbf{G}X, (\mathbf{I} - \mathbf{G})X) = \text{Cov}(\mathbf{G}X, \mathbf{G}(\mathbf{I} - \mathbf{G})X) = 0,
\]

since \( \mathbf{G} \) is symmetric and idempotent. Also observe that

\[
\|\mathbf{G}X\|^2 = n\bar{X}^2,
\] (31)

since \( \mathbf{G}X = [\bar{X}, \ldots, \bar{X}]^T \), and that

\[
\|(\mathbf{I} - \mathbf{G})X\|^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2.
\] (32)

We now use of the following fact: if \( \mathbf{A} \) is an \( n \times n \) matrix of coefficients then

\[
\mathbb{E}\|\mathbf{AX}\|^2 = \mathbb{E}[\|\mathbf{AX}\|^2] = \mathbb{E}[\mathbf{A}^T \mathbf{X}^T \mathbf{A} \mathbf{X}] = \mathbb{E}[\text{trace}(\mathbf{X}^T \mathbf{A}^T \mathbf{A} \mathbf{X} \mathbf{X}^T)] = \mathbb{E}[\text{trace}(\mathbf{A}^T \mathbf{A} \mathbb{E}(\mathbf{X} \mathbf{X}^T))] = \text{trace}(\mathbf{A}^T \mathbf{A} \mathbb{E}(\mathbf{X} \mathbf{X}^T) + \mathbb{V}(X))
\] (33)

where we have used standard properties for the trace, and the fact that \( \mathbb{E}(\mathbf{X} \mathbf{X}^T) = \mathbb{E}(\mathbf{X})\mathbb{E}(\mathbf{X}^T) + \mathbb{V}(X) \). Thus, using (33), (31), (30), and the fact that \( \text{trace}(\mathbf{G}) = 1 \), we
obtain

\[ E\left( n\mathbf{X}^2 \right) = E\|\mathbf{G}\mathbf{X}\|^2 = \text{trace}[\mathbf{G}^T \mathbf{G}(E(\mathbf{X})E(\mathbf{X}^T) + \mathbf{V}(\mathbf{X}))] \]
\[ = \text{trace}[\mathbf{G}(n\mu^2 \mathbf{G} + ((\sigma^2 + (n - 1)\gamma) \mathbf{G} + (\sigma^2 - \gamma)(\mathbf{I} - \mathbf{G})))] \]
\[ = n\mu^2 \text{trace}[\mathbf{G}] + (\sigma^2 + (n - 1)\gamma)\text{trace}[\mathbf{G}] \]
\[ = n\mu^2 + \sigma^2(1 + (n - 1)\rho), \]

which establishes result (28). Similarly, using (32) and the fact that \( \text{trace}(\mathbf{I} - \mathbf{G}) = n - 1 \), we obtain

\[ E\left( \sum_{i=1}^{n} (X_i - \bar{X})^2 \right) = E\|\mathbf{I} - \mathbf{G}\mathbf{X}\|^2 = \text{trace}[(\mathbf{I} - \mathbf{G})^T(\mathbf{I} - \mathbf{G})(E(\mathbf{X})E(\mathbf{X}^T) + \mathbf{V}(\mathbf{X}))] \]
\[ = \text{trace}[(\mathbf{I} - \mathbf{G})(n\mu^2 \mathbf{G} + (\sigma^2 + (n - 1)\gamma) \mathbf{G} + (\sigma^2 - \gamma)(\mathbf{I} - \mathbf{G})))] \]
\[ = (\sigma^2 - \gamma)\text{trace}(\mathbf{I} - \mathbf{G}) \]
\[ = (n - 1)\sigma^2(1 - \rho), \]

which establishes result (29). Importantly, we have established that the random quantities inside the expectation of results (28) and (29) are uncorrelated.

Applying results (28) and (29) to the situation where \( l_1, l_2, \ldots, l_n \) are Gamma(2, \( \tau \)) variables with common pairwise correlation \( \rho \), using the fact that \( \mu^2 = 2\sigma^2 \), and using the standard definition of the sample variance \( S^2 \), these equations become

\[ E\left( nl^2 \right) = \sigma^2(2n + 1 + (n - 1)\rho), \]
\[ E\left( S^2 \right) = \sigma^2(1 - \rho). \]

An unbiased estimate of the left hand side of each equation is simply given by the quantity inside the expectation, and since these quantities are uncorrelated, we can obtain a single estimate of \( \rho \) by eliminating \( \sigma^2 \). Thus, following some elementary algebra, we obtain

\[ \hat{\rho} = \frac{nl^2 - S^2(2n + 1)}{nl^2 + S^2(n - 1)}, \]

which is the correlation estimate.

Even though \( \hat{\rho} \in [-1, 1] \), some values of \( \hat{\rho} \) produce inappropriate values of \( n^* \). When this occurs we adjust the calculation of \( n^* \). Recall that there are two places where \( n^* \) is used: the bias correction factor \( b(n^*) \), and parametrizing the Gamma approximation, i.e. \( \hat{\tau}^{-1} \sim \text{Gamma}(2n^*, 2n^*\tau) \). Also recall that

\[ n^* = \frac{n}{1 + (n - 1)\rho}. \]
Observe that $n^*$, as a function of $\rho$, is a rectangular hyperbola on $[-1, 1]$ with a vertical asymptote at $\rho = -1/(n - 1)$. There are three sets of $\hat{\rho}$ values which are potentially problematic:

$$I_1 = \left[-1, -\frac{2}{n - 1}\right], \quad I_2 = \left(-\frac{2}{n - 1}, \frac{-1}{n - 1}\right), \quad I_3 = \left(-\frac{1}{n - 1}, 0\right).$$

When $\hat{\rho} \in I_2$ or $\hat{\rho} \in I_3$ then $|n^*| > n$. If we interpret the correlation as effectively reducing the sample size, then an estimated correlation falling within either of these intervals suggests that there is more information in the sample than a sample of size $n$. For our concerns, the effective sample size should be no greater than a sample size assuming independence, i.e. $n^* \leq n$. An added complication is that when $\hat{\rho} \in I_1$ or $\hat{\rho} \in I_2$ then $n^* < 0$. Such $n^*$ values, however, cannot be used to parametrize the Gamma approximation.

Thus, we adopt the following procedures. When calculating the bias correction factor $b(n^*)$, if $\hat{\rho} \in I_2$ or $\hat{\rho} \in I_3$ we bound $n^*$ so that $|n^*| \leq n$. This produces accurate bias correction. When parametrizing the Gamma approximation, if $\hat{\rho} \in I_2$ we set $n^* = n$, and if $\hat{\rho} \in I_1$ we recalculate $n^*$ using $|\hat{\rho}|$ as the correlation value. This produces reliable confidence intervals since the $|\hat{\rho}|$ value, used to recalculate $n^*$, reliably picks out the correct effective (reduced) sample size. Note that intervals $I_2$ and $I_3$ become narrower (approaching zero length) as the sample size increases.

Finally, note that in a correlated genealogy samples can sometimes have repeated observations, i.e. more than one individual has inherited the very same length of ancestral segment. While this is rare, repeated observations can bias $\hat{\rho}$, and so when calculating the estimate we remove any repeats of distinct observations.

Appendix E: Correction for chance sharing

The effect of chance sharing only becomes noticeable when a mutation is old (e.g. $\tau \geq 100$). This is because $\hat{\tau} \propto 1/\sum l_i$, i.e. the age estimate is a hyperbolic function of the sum of lengths. Thus, when the mutation is old and the sum of lengths small, a small change in the lengths due to chance sharing can be magnified into a noticeable (albeit small) change in the estimated age. On the other hand, when the mutation is younger and the sum of lengths larger, any change in the estimated age due to a small change in lengths is undetectable. While this problem is minor, we develop a simple correction.

We assume that the marker loci are bi-allelic, with alleles A and B. Consider one of the shorter inferred ancestral segments $S$ in relation to a longer segment $T$, as we head away from the mutation on one side (see Figure 4). Now consider marker locus $L_1$, the first locus before the inferred endpoint of $S$. The probability that $S$ and $T$ share the same alleles at $L_1$ by chance is

$$P(\text{match}) = f_A^2 + f_B^2 = f_A^2 + (1 - f_A)^2,$$

where $f_A$ and $f_B$ are the population frequencies of alleles A and B, respectively. The probability of matching by chance at loci $L_2, L_3, \ldots$, as we head toward the mutation, can
be calculated similarly. Let $p_i$ denote the probability of matching at locus $i$. Thus, ignoring LD, the probability $P_{\text{run}}$ of a “run” of matches up to some locus $k$ is then

$$P_{\text{run}} = P(\text{run of matches}) = p_1p_2\cdots p_k.$$  

To deal with chance sharing the idea is to find the number $k$ such that $P_{\text{run}} < \epsilon$, where $\epsilon$ is some threshold (we set $\epsilon = 0.01$), and then remove these $k$ loci from the length calculation. We now use the following approximate strategy. In the expression for $p_i$, replace the particular $f_A$ with the median of the $f_A$ frequencies across the observed loci on the relevant chromosome. So, letting $\langle f_A \rangle$ denote the median, we have

$$p = p_i = \langle f_A \rangle^2 + (1 - \langle f_A \rangle)^2,$$

for all $i$. To find the number $k$ of loci to remove from the length calculation, we find the $k$ such that

$$P_{\text{run}} = p_1p_2\cdots p_k = p^k < \epsilon,$$

which is $k = \log(\epsilon)/\log(p)$. For standard SNP arrays, with $\epsilon = 0.01$, $k$ is on the order of 10 loci. This is translated into genetic length by multiplying by the average genetic length between observed markers, which we denote by $\phi$. Thus, to correct for chance sharing, for each segment arm on either side of the mutation we subtract $\phi k$ from its inferred length.$^5$

$^5$This correction is not used for independent genealogies for sampled sizes $n \leq 10$, since the age estimate in this circumstance has a slight upward bias which is offset by the slight downward bias caused by chance sharing, thus obviating the need to correct for chance sharing.