10 Estimating the age of mutations using variation at linked markers

David E. Reich and David B. Goldstein

Chapter contents

Abstract

We present a general method for estimating the dates of mutations using variation at linked microsatellite markers. Risch *et al.* (1995) take a similar approach to estimating the age of the mutation causing idiopathic torsion dystonia among Ashkenazic Jews, but they do not describe how to produce a confidence interval for the date. Here, we not only obtain a confidence interval for the date by assessing the degree of correlation among samples, but also describe how to use a Markov transition matrix approach to take full account of the complexities of the recombination process. Finally, we show how the method has been applied to a specific example: estimation of a date for a mutation that confers resistance to HIV-1 infection (Stephens *et al.* 1998).

10.1 Introduction

It is possible to estimate the age of a mutation because of the non-random association of alleles (i.e. linkage disequilibrium) that is generated whenever a new mutation occurs. The immediate descendants of a mutant chromosome will be

monomorphic for a set of markers linked to the locus of interest. Over time, however, as recombination and mutation undo the linkage disequilibrium, the pattern of variation among mutant chromosomes will gradually reflect the pattern of variation in the population as a whole. By making a quantitative assessment of the extent to which the disequilibrium has been undone, and using known rates of mutation and recombination, we can estimate an age for the most recent common ancestor of mutant chromosomes.

10.2 Estimating the age of the mutation when almost all chromosomes have the ancestral haplotype

To estimate the date of the mutation when almost all mutant chromosomes are of a single type, we employ a two-pronged strategy. First, we assume that the common haplotype is the ancestral haplotype, a questionable assumption if the genealogical tree of relationships among individuals includes only a few ancient lineages, and in particular, if an early mutation or recombination event occurred on a lineage that was ancestral to the majority of current chromosomes. To determine the ancestral haplotype unequivocally, we use markers that are relatively close to the gene locus of interest. We then use the frequency (r) of mutation and recombination events that have the potential to unlink some chromosomes from the ancestral haplotype to find the most likely number of generations that have passed since the ancestral mutant chromosome.

To obtain the maximum likelihood estimate for the date of the mutation, we begin by considering a particular lineage of the genealogy, the chain of ancestors linking a present-day haplotype to the haplotype at coalescence. The probability that a haplotype remains ancestral during the time tracing back to the most recent common ancestor is given by the depth of the genealogy in generations, G, and the frequency r of mutation and recombination:

$$p = \mathrm{e}^{-Gr}. \tag{10.1}$$

Here, p is just the zero term in a Poisson series with parameter Gr.

To find p, we note that for a dramatically expanded population, one for which all lineages are essentially independent, an unbiased estimate of p is the proportion of observed haplotypes that are ancestral (Stephens *et al.* 1998). A surprising fact is that this statement is true even for constant-sized populations in which many lineages are highly correlated in the sense that pairs of alleles share extensive periods of co-ancestry during the time tracing back to the most recent common ancestor of the sample. The reason why the age estimate is independent of topology is that as long as mutations at the marker loci have no selective effect, the correlations in the tree amount to a process of pseudo-replication of lineages. This process will affect the variance of our estimate of p (see below); however, because the lineages that are replicated are selected independent of allelic state, the proportion of ancestral haplotypes will not be systematically affected.

Finally, to obtain G in terms of the estimate of p, we transform eqn (10.1):

$$G = -\ln(p)/r. \tag{10.2}$$

As discussed previously, this holds true whatever the shape of the genealogical tree.

10.3 A comprehensive approach for estimating the age of a mutation

The previous method produces an appropriate estimate for the age of the mutation when the large majority of observed chromosomes have become unlinked from the ancestral haplotype. However, when enough mutant chromosomes have become unlinked from the ancestral haplotype, the date estimate must account not only for the rate of loss of the ancestral haplotype by mutation or recombination, but also for regeneration of the ancestral haplotype among chromosomes that currently do not have it (Risch *et al.* 1995). When this process is included in our analysis, the estimated date of mutation becomes systematically older than that predicted by eqn (10.2).

To provide a complete description for a system in which a single locus is typed, we use a Markov transition matrix **K**. Note that Risch *et al.* (1995) have used an alternative approach to the same problem, involving differential equations. However, we have chosen to use the transition matrix approach instead because we find it to be very flexible, and because it allows us to easily incorporate mutation and recombination events into the same evolutionary process. Specifically, the entries in the Markov matrix give the probabilities, per generation, that any one haplotype will transform into any other. To calculate **K**, we take a weighted sum of matrices corresponding to recombination (**R**), mutation (**M**), and no event occurring (**I**):

$$\mathbf{K} = c\mathbf{R} + \mu\mathbf{M} + (1 - c - \mu)\mathbf{I}, \tag{10.3}$$

where c is the frequency of recombination, μ is the frequency of mutation, and $1 - c - \mu$ is the frequency of no event occurring. We now consider a single lineage tracing its ancestry back to the original mutation, and by multiplying **K** by the state vector generation by generation, evaluate the probability that after n generations, the mutation will have lost its linkage to the ancestral haplotype. This is exactly analogous to the method described in Section 10.2, except here we take into account regeneration of the ancestral haplotype as well as the rate of loss of that haplotype.

Consider the case in which only a single microsatellite marker has been typed. For this case, the state vector is represented as $(q, 1 - q)$, with the first entry the probability that the allele is of the ancestral type and the second the probability that it is not. The matrices **R** and **M**, and hence the Markov transition matrix, can then be derived straightforwardly from the distribution of alleles in non-mutant

chromosomes. We begin with the recombination matrix (**R**). After a recombination event, the probability that the allele will end up ancestral, regardless of the initial state, can be estimated as the proportion of alleles in the population that have the ancestral haplotype (a). The probability that the allele will be non-ancestral type is then $1 - a$:

$$\mathbf{R} = \begin{bmatrix} a & a \\ 1-a & 1-a \end{bmatrix}. \tag{10.4}$$

We now calculate the mutation matrix (**M**). According to the stepwise mutation model for microsatellites (Goldstein and Pollock 1997), mutations change the length of an allele by a single unit, with an equal chance of increasing or decreasing the length of the allele. Using this model, we estimate the probability that a mutation will transform a non-ancestral allele into an ancestral one as $b/2$, where b is the proportion of alleles that are one mutation step away from the ancestral haplotype, and the division by 2 occurs because only half of mutations at these alleles produce the ancestral type. Note that in the case of a mutation that occurs on an ancestral allele, the outcome is even simpler: the probability that an allele will remain ancestral is 0.

$$\mathbf{R} = \begin{bmatrix} 0 & b/2 \\ 1 & 1-b/2 \end{bmatrix}. \tag{10.5}$$

To find b in any generation, we require information that is not contained in the two-dimensional state vector: specifically, the frequencies of alleles that are one mutation step away from the ancestral chromosome. Thus, to describe the frequencies of all k possible alleles in the system, we require a k-dimensional state vector—a complicated circumstance because the **R** and **M** matrices would now have to be $k \times k$ rather than 2×2. Nevertheless, it is often possible to simplify the analysis when recombination occurs much more frequently than mutation. In this case, the distribution of non-ancestral alleles among mutant chromosomes is expected to be the same as in the control population, and b can be estimated directly from the proportions of alleles in the control population.

We now use eqn (10.3), and the matrices **R** and **M**, to obtain the Markov transition matrix **K**. Errors in **K** could arise either from misestimation of c and μ (since information about these parameters is often inaccurate), or from errors in a and b that might occur due to inappropriate selection of control populations or failure to type a sufficient number of chromosomes in the control population, or changes in the proportions of alleles in the population over the course of recent history. Since none of these sources of error is taken into account in our method for estimating a date of mutation, experimenters should consider a range of possible values of c, μ, a, and b, as a way of assessing how much variability in the estimate of the age of the mutation could arise from misestimation of parameters.

Under the assumption that **K** is correct, we can now consider a particular lineage of the genealogy—the chain of ancestors linking a present-day haplotype to the haplotype at coalescence—and use **K** to determine the probability that the

lineage remains ancestral at any given generation. We begin with the state vector representing the ancestral mutant chromosome, which has the form (1, 0) where the first entry is the probability that the lineage has the ancestral type. To evaluate the fate of the lineage in every subsequent generation, we multiply **K** by the state vector until we obtain a probability of observing an ancestral haplotype that is closest to the observed proportion, p, of mutant chromosomes. The number of times that **K** has been multiplied tells us the number of generations that have passed since the ancestral mutant chromosome.

10.4 Variance of the age estimate

The variance of the age estimate (unlike the age estimate itself) is systematically affected by a population's demographic history. The reason for this is that populations with different demographic histories have differently shaped genealogical trees. For example, in a population that has undergone a relatively recent and dramatic expansion, almost all lineages will trace their ancestry independently back to the time of the expansion, and the number of independent assessments of the age of the tree will be equal to the number of samples. For a constant-sized population, there will be high degree of shared ancestry among sampled chromosomes, as explained above, and the number of independent assessments of the age of the tree will therefore be much smaller than the number of sampled chromosomes. The relatively large number of age assessments in an expanding population means that the date estimate is more accurate.

To determine confidence intervals for the date, we use computer simulations based on a coalescent algorithm by R.R. Hudson (1990) to describe a wide variety of population histories from constant population size to fast growth (final population size and exponential growth rate are the variable parameters in our simulation). For each set of demographic parameters, the simulation generates a large number of genealogical trees and distributes mutation and recombination events along them according to a random (Poisson) process (we use the Markov transition matrix to determine which events turn an ancestral haplotype into a non-ancestral one and vice versa). Thus, the final distribution of haplotypes along a genealogical tree is affected by two sources of error: first, variability in the shapes of the genealogical tree, and second, variability in the mutation and recombination events that occur on those trees. The simulations allow us to take account of both these sources of error, generating a 95 per cent central confidence interval for the number of ancestral haplotypes that could be expected to be seen in such a sample. We can then reject certain combinations of demographic parameters if the confidence intervals do not contain the number of ancestral haplotypes that was actually observed.

To find allowed dates for the mutation, we consider each combination of demographic parameters separately, simulating many genealogical trees and considering only those simulations that result in the observed number of ancestral haplotypes (i.e. we condition the simulations on the observed results). From the

subset of trees obtained in this manner, we can then produce a 95 per cent central confidence interval for the date of the mutation. To obtain an allowed range of dates that is inclusive of all possible demographic histories, we then take the union of confidence intervals for each combination of parameters. The range of allowable dates can be constricted even further if we have additional information about the demographic history—for example, if the observed distribution pattern of non-ancestral haplotypes forbids particular combinations of demographic parameters, as explained in Section 10.6, below.

10.5 Age of the *CCR5-Δ32* AIDS resistance allele

The *CCR5* gene encodes a protein that serves as part of the primary entry port for HIV-1 in immune cells (Deng *et al.* 1996). Individuals homozygous for a particular 32 base-pair deletion mutation in the gene, which we designate as *CCR5-Δ32*, are resistant to HIV-1 infection (Dean *et al.* 1996). Indeed, as many as 26 per cent of northern Europeans carry at least one deleted copy of the gene, while the frequency of carriers drops to zero along a north–south gradient (no copies are observed among Africans). The pattern of distribution of the gene makes it seem likely that the mutation occurred recently, and it is therefore of interest to obtain a direct estimate for the date of origin of the mutation.

The data we use consist of 46 chromosomes carrying the *CCR5-Δ32* deletion, and 146 controls that do not carry the mutation. Each chromosome was typed at two microsatellite markers on the same side of the *CCR5* gene: GAAT12D11 (GAAT) and AFMB362wb9 (AFMB), with GAAT closest to the deletion locus. The ancestral haplotype is taken to be the one in which the GAAT marker carries the 197 base-pair allele and the AFMB marker carries the 215 base-pair allele. This haplotype occurs among 85 per cent of mutant chromosomes but only 36 per cent of the control population.

To calculate the Markov transition matrix for this system, we note that two polymorphic markers were typed, and that there are therefore four possible states in the system. Specifically, the states can be classified as follows: (1) both GAAT and AFMB are ancestral; (2) only GAAT is ancestral; (3) only AFMB is ancestral; and (4) neither GAAT nor AFMB is ancestral. The state vector can be represented as $(q_1, q_2, q_3, 1 - q_1 - q_2 - q_3)$, and the transition matrices, corresponding to mutation at GAAT, mutation at AFMB, recombination at GAAT or recombination at AFMB, will be four-dimensional (4×4) as well. The overall equation for the transition matrix $\mathbf{K}$ is then:

$$\mathbf{K} = \mu_{\mathrm{GAAT}}\mathbf{M}_{\mathbf{GAAT}} + \mu_{\mathrm{AFMB}}\mathbf{M}_{\mathbf{AFMB}} + c_{\mathrm{GAAT}}\mathbf{R}_{\mathbf{GAAT}} + c_{\mathrm{AFMB}}\mathbf{R}_{\mathbf{AFMB}} + (1 - c_{\mathrm{GAAT}} - c_{\mathrm{AFMB}} - \mu_{\mathrm{GAAT}} - \mu_{\mathrm{AFMB}})\mathbf{I}, \tag{10.6}$$

where μ_{GAAT}, μ_{AFMB}, c_{GAAT}, and c_{AFMB} are the rates of mutation and recombination for the GAAT and AFMB markers, and $\mathbf{M}_{\mathbf{GAAT}}$, $\mathbf{M}_{\mathbf{AFMB}}$, $\mathbf{R}_{\mathbf{GAAT}}$, and $\mathbf{R}_{\mathbf{AFMB}}$ are mutation and recombination matrices.

We must now estimate the parameters μ_{GAAT}, μ_{AFMB}, c_{GAAT}, and c_{AFMB}. To obtain the recombination rates c_{GAAT} and c_{AFMB}, we use physical distances that were determined from radiation hybrid mapping, and convert these to recombination distances using a linear regression that applies on average across the chromosome on which the mutation was found. To estimate the mutation rates μ_{GAAT} and μ_{AFMB}, we use the published value for dinucleotide microsatellites, $\mu = 0.00053$ (Weber and Wong 1993). In this analysis, error in estimation of the recombination rate was much more of a worry to us than error in the mutation rate, since the recombination rate is so much larger in absolute terms.

To obtain the mutation matrices, we use the frequencies of alleles in the control population that are one mutation step away from the ancestral GAAT (b_1) and ancestral AFMB (b_2) alleles (see eqn (10.4)). It follows that for mutation at the GAAT marker, the matrix is $\mathbf{M_{GAAT}}$, while for mutation at the AFMB marker, the matrix is $\mathbf{M_{AFMB}}$.

$$\mathbf{M_{GAAT}} = \begin{vmatrix} 0 & 0 & b_1/2 & 0 \\ 0 & 0 & 0 & b_1/2 \\ 1 & 0 & -b_1/2 & 0 \\ 0 & 1 & 0 & -b_1/2 \end{vmatrix}, \quad \mathbf{M_{AFMB}} = \begin{vmatrix} 0 & b_2/2 & 0 & 0 \\ 1 & -b_2/2 & 0 & 0 \\ 0 & 0 & 0 & b_2/2 \\ 0 & 0 & 1 & -b_2/2 \end{vmatrix}. \tag{10.7}$$

To obtain the recombination matrices, we follow eqn (10.5), dealing first with the case in which the recombination occurs between the gene locus of interest and GAAT, and then the case in which the recombination event occurs between GAAT and AFMB. In the first case, the situation is exactly analogous to eqn (10.4), and the frequencies of each possible outcome can be estimated as the proportion of alleles in the control population that are of each haplotypic state. We designate these frequencies, respectively, as a_1, a_2, a_3, and a_4, recalling that $a_4 = 1 - a_1 - a_2 - a_3$. The resulting matrix is designated $\mathbf{R_{GAAT}}$. In the second case, in which the recombination occurs between GAAT and AFMB, the alleles change at only a single locus (AFMB), and the only relevant parameters are the frequency of alleles for which the AFMB marker had the ancestral type ($a_1 + a_3$), and the frequency of alleles for which the AFMB marker was non-ancestral ($a_2 + a_4$). The overall 4×4 transition matrix, $\mathbf{R_{AFMB}}$, then becomes:

$$\mathbf{R_{GAAT}} = \begin{vmatrix} a_1 & a_1 & a_1 & a_1 \\ a_2 & a_2 & a_2 & a_2 \\ a_3 & a_3 & a_3 & a_3 \\ a_4 & a_4 & a_4 & a_4 \end{vmatrix}, \quad \mathbf{R_{AFMB}} = \begin{vmatrix} a_1 + a_3 & a_1 + a_3 & 0 & 0 \\ a_2 + a_4 & a_2 + a_4 & 0 & 0 \\ 0 & 0 & a_1 + a_3 & a_1 + a_3 \\ 0 & 0 & a_2 + a_4 & a_2 + a_4 \end{vmatrix}. \tag{10.8}$$

We now use eqn (10.6) to calculate $\mathbf{K}$. Ignoring any error in the Markov transition matrix (more likely to be due to errors in estimation of the recombination rate and recombination parameters rather than errors in the mutation rate), the most likely age for the *CCR5-Δ32* mutation is 29 generations, or 725 years assuming a generation time of 25 years. For comparison, if the calculation is done according

to the method of Section 10.2, the estimate is 28 generations, slightly younger because no Markov transition matrix is used to take into account regeneration of the ancestral haplotype. Note that the estimated date of the mutation is likely to be systematically lower than the date of first appearance of the mutation, since the estimation procedure only finds information about the age of the most recent common ancestor of the sampled chromosomes. Thus, our estimate of the date must be interpreted with caution: if a dramatic expansion occurred in the population of mutant chromosomes, it is likely that the most recent common ancestor of the mutant chromosomes dates to before the expansion (although it is difficult to say how much earlier). Note that Slatkin and Rannala (1997) provide an approach for dating mutations that takes this systematic bias into account.

10.6 Estimating a variance for the date by reconstructing the genealogy of *CCR5-Δ32*

To obtain a confidence interval for the date estimate, we use simulations that take into account all possible combinations of demographic parameters and genealogical trees, as described in Section 10.4. To place further restrictions on the allowed dates of the mutation, we forbid certain genealogical trees—in the simplest case by using prior knowledge of population history. For the *CCR5-Δ32* data, for example, we assume that during the past 10 000 years, northern European populations have had a certain minimum size. By specifying that the initial effective population size was at least 5000, we conclude that the date of the most recent common ancestor was between 11 and 75 generations in the past (275–1875 years, assuming 25 years per generation).

In a much more fundamental way, it is also possible to use the distribution of non-ancestral haplotypes among mutant chromosomes to put restrictions on the shape of the genealogical tree. For example, if the haplotypes all derive from separate mutation or recombination events, the lineages of the genealogical tree are uncorrelated, and consistent with a dramatically expanded population. If the non-ancestral haplotypes derive from relatively few mutation or recombination events (which have been recopied and amplified within the lower branches of the genealogical tree), then the history of the mutant chromosomes is more likely to be consistent with a constant-sized population. By focusing on the distribution of non-ancestral haplotypes among *CCR5-Δ32* chromosomes, we are then able to directly assess the degree of correlation in the tree, and from there to assess the variability of the date estimate.

To implement this approach, we consider the fact that of the seven non-ancestral *CCR5-Δ32* chromosomes that were observed, there were four distinct haplotypes. The number of mutation and recombination events that actually gave rise to the four haplotypes was probably larger than four, since the distribution of non-mutant *CCR5* chromosomes indicates that given six or seven chances, several haplotypes would be generated more than once (and, as expected from this hypothesis, the non-ancestral haplotypes we observe are the ones that are most frequent in the

control population). We surmise that the non-ancestral haplotypes derive for the most part from separate mutation and recombination events, and that in the present sample, we are observing the results of at least six and perhaps seven different events. Note that it would have been possible to determine the number of events with even more precision if more that two microsatellite markers had been typed.

To make explicit use of this information, we modify the simulation described in Section 10.4 to report not only the number of non-ancestral haplotypes but also the number of distinct mutation and recombination events that gave rise to these haplotypes. Thus, for each set of demographic parameters in the *CCR5-Δ32* data set, we simulate a large number of genealogical trees that gave rise to seven out of 46 non-ancestral haplotypes, and then determine the proportion of these replicates that were derived from seven distinct events. If we require that no fewer than 5 per cent of replicates have fewer than seven distinct haplotypes, we can restrict the date of the mutation to between nine and 214 generations in the past (225–5350 years, assuming 25 years per generation). While this restriction on the date of the mutation is less stringent than the one derived from a historical assumption about effective population sizes, it is valuable precisely because it is independent of such assumptions.

10.7 The analysis of new data sets

In applying the method to a new data set, it is always appropriate to begin by picking microsatellite markers that have the proper distance from the gene locus of interest. The markers should be chosen to be close enough to the locus of interest to define the ancestral haplotype, but far enough away to allow as many lineages as possible to have had a chance to become non-ancestral. A good strategy for identifying markers is to select a panel that are at varying distances from the gene locus of interest, and then to pick out ones that comply with the criteria described above.

The analysis of data from a single microsatellite locus can often extract most of the relevant information about the date of a mutation. However, the use of multiple markers (e.g. in the *CCR5-Δ32* experiment) may have a particular value in assessing the variance of the date estimate, allowing for a better assessment of the shape of a genealogical tree than would be possible with a single marker. The reason for this is that multiple markers allow us to reconstruct more accurately the history of mutation and recombination events. If even more markers are typed, it becomes possible to pinpoint the exact number of distinct mutation and recombination events that had led to the observed number of non-ancestral haplotypes, further restricting the allowed range of genealogical tree. On the other hand, multiple markers have a drawback because they can make an analysis more complicated, forcing the estimation of more matrix parameters, recombination distances and mutation rates.

Another factor to consider in designing future experiments is that some mutations will be sufficiently old that only markers close to the locus will display

disequilibrium. In this case, it will be difficult to determine the recombination distances of markers from the locus, and it is appropriate to use markers that are sufficiently close to the gene that mutation serves as the main molecular clock for estimating a date for the mutation. Errors in estimating the mutation rate (and not the recombination rate) then become the main source of systematic error in determining the age of the mutation, and to reduce this error, it is appropriate to use several markers that are close to the gene locus of interest, with an average mutation rate that in general will be more predictable than that of a single marker (Goldstein and Pollock 1997). In practice, however, it may be difficult to find enough markers that are sufficiently close to the gene locus of interest to make this possible, except perhaps on the Y chromosome, where a large number of microsatellites are completely linked.