REVIEW ARTICLE

Using Linked Markers to Infer the Age of a Mutation

Bruce Rannala[1]* and Giorgio Bertorelle[2]

[1]*Department of Medical Genetics, University of Alberta, Edmonton, Alberta, Canada*
[2]*Sezione di Biologia Evolutiva, Departimento di Biologia, Universita' di Ferrara, Ferrara, Italy*

Communicated by Richard G.H. Cotton

Advances in sequencing and genotyping technologies over the last decade have enabled geneticists to easily characterize genetic variation at the nucleotide level. Hundreds of genes harboring mutations associated with genetic disease have now been identified by positional cloning. Using variation at closely linked genetic markers, it is possible to predict the times in the past at which particular mutations arose. Such studies suggest that many of the rare mutations underlying human genetic disorders are relatively young. Studies of variation at genetic markers linked to particular mutations can provide insights into human geographic history, and historical patterns of natural selection and disease, that are not available from other sources. We review two approaches for estimating allele age using variation at linked genetic markers. A phylogenetic approach aims to reconstruct the gene tree underlying a sample of chromosomes carrying a particular mutation, obtaining a "direct" estimate of allele age from the age of the root of this tree. A population genetic approach relies on models of demography, mutation, and/or recombination to estimate allele age without explicitly reconstructing the gene tree. Phylogenetic methods are best suited for studies of ancient mutations, while population genetic methods are better suited for studies of recent mutations. Methods that rely on recombination to infer the ages of alleles can be fine-tuned by choosing linked markers at optimal map distances to maximize the information available about allele age. A limitation of methods that rely on recombination is the frequent lack of a fine-scale linkage map. Maximum likelihood and Bayesian methods for estimating allele age that rely on intensive numerical computation are described, as well as "composite" likelihood and moment-based methods that lead to simple estimators. The former provide more accurate estimates (particularly for large samples of chromosomes) and should be employed if computationally practical. Hum Mutat 18:87–100, 2001. © 2001 Wiley-Liss, Inc.

KEY WORDS: bioinformatics; DNA analysis; DMLE; mutation age; mutation detection; mutation rate; phylogenetic; population genetics

INTRODUCTION

In the last two decades, many rare mutations underlying simple genetic disorders have been identified by positional cloning. With the completion of the human genome project [Lander et al., 2001; Venter et al., 2001] there is the prospect that additional common mutations will be discovered that influence complex genetic disorders [Risch, 2000]. Virtually all genetic disorders for which genes have been identified, to date, display some degree of allelic heterogeneity. That is, the disease-associated chromosomes may carry different mutations of the same gene. In some cases, a single mutation may dominate in frequency, the ΔF508 mutation, for example, which accounts for roughly

70% of mutations that cause cystic fibrosis in Europeans [EWGCFG, 1990]. In other cases, the mutations underlying a disease in a population may all be quite rare, with no one mutation dominating. In the case of Wilson disease, for example, over 80 different mutations have been observed, all in relatively low frequency [Roberts and Cox, 1998]. In a recent study of 136 Wilson disease patients of Mediterranean descent, 50 different mutations were observed. The explanation for such patterns lies in the demographic histories of human populations and the historical influences of genetic drift, migration, and natural selection [see Cavalli-Sforza et al., 1994].

A first step in understanding the influence of historical events on human genetic variation is to consider the ages of particular mutations. Potential information about the ages of mutations is available from the variation observed at closely linked genetic markers. In this article, we review existing methods for estimating the age of a particular mutation using such information and provide some suggestions for future research. It will become clear that much remains to be done to develop robust statistical methods for estimating mutation ages. It will also become clear that, even in ideal situations, the resulting estimates of mutation age have wide margins of error. Despite these limitations, it is still worthwhile, in our opinion, to attempt to use the patterns of genetic variation at markers linked to significant mutations to seek insight into human demographic history, historical patterns of selection, and the ages of particular genetic disorders. It will often be the case that such insights are available from no other source.

Historical Background

Theoretical results bearing on the ages of mutations have roots at least as far back as the 1930s and in early work on the expected change in frequency of a mutation over time under the joint influences of selection, genetic drift, and migration [Fisher, 1930; Haldane, 1932; Wright, 1931]. Mutation age was not a primary focus of this work, however, and population geneticists did not explicitly consider the ages of mutations until the 1960s, when the molecular details of the process of DNA mutation became clearer and new mathematical approaches for studying population genetic structure using diffusion theory were developed [see Ewens, 1979].

Ohta and Kimura [1973] present a method for estimating the age of a mutation based on its population frequency; they derive the expected (average) age of a neutral mutation present at a given frequency in a population and obtain a simple method of moments estimator of mutation age by setting this expected frequency equal to the observed frequency and solving for the unknown mutation age. Maruyama [1974a, b] extended these results to allow estimation of the age of a mutation under overdominant (balancing) selection. His theory shows that overdominant mutations in populations tend to be older than neutral mutations at the same frequency. In fact, after sufficient time has elapsed the frequency of an overdominant mutation is effectively independent of its age, making it impossible to estimate age from frequency.

The emergence of allozyme electrophoresis techniques in the late 1960s [Harris, 1966; Lewontin and Hubby, 1966] presented new possibilities for estimating the ages of alleles at allozyme loci based on their population frequencies. However, the fact that many mutations do not affect electrophoretic mobility for allozymes, and some have identical effects on mobility, rendered ambiguous the meaning of the "age" of an electrophoretic variant.

Advances in DNA amplification and sequencing technologies, and microsatellite genotyping procedures, as well as the positional cloning of numerous disease mutations of humans in the late 1980s and early 1990s enabled geneticists to unambiguously characterize mutations for frequency-based studies of their ages. At about the same time, the focus of interest shifted from mutation frequency (and its relation to age) to the genealogy of a sample of chromosomes descended from a particular ancestral mutation and the information that this could provide about its age [e.g., Morral et al., 1994]. Information about the genealogy underlying a particular mutation can be obtained by examining genetic variation at closely linked markers.

In this article, we focus on the problem of estimating the ages of mutations using information from the genealogy underlying a sample of chromosomes carrying the mutation, rather than the frequency of the mutation in a population. The two approaches (frequency-based and genealogy-based) are complimentary, each providing a conditionally independent source of

information about the age of a mutation. Slatkin and Rannala [2000] provide a more general review that covers both types of estimators of mutation age.

INTRAALLELIC GENEALOGY AND THE AGE OF A MUTATION

It is important to unambiguously define what is meant by the age of a mutation. We define the age of a nonrecurrent mutation, M, to be the generation, t, in the past at which the mutation arose. This is not, in general, equal to the age of the most recent common ancestor (MRCA) of a sample of chromosomes bearing the mutation, as illustrated in Figure 1. Here, we use the terms "mutation" and "allele" synonymously. Mutation can also refer to the process by which alleles (mutations) arise, and we will occasionally use mutation to refer to the process as well, but the meaning of our usage should be clear from the context. The age of a mutation, M, will be estimated using a sample of chromosomes descended from the mutant, and possibly also a sample of normal chromosomes not bearing mutation M. We ignore complications arising from population subdivision, only briefly discussing its possible
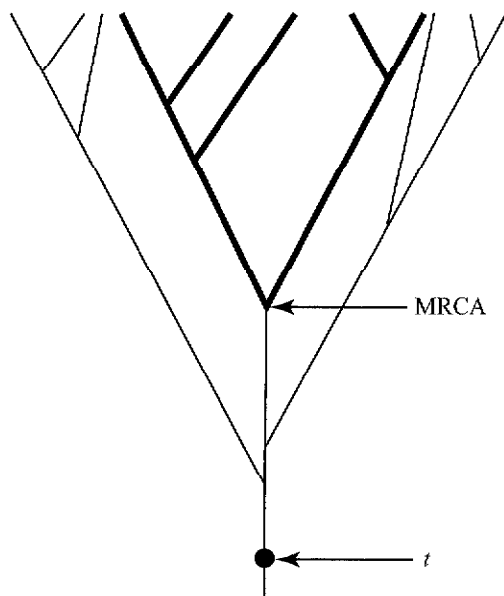

FIGURE 1.   The genealogy of a population of chromosomes bearing a mutation M that arose at time t in the past (mutation event is indicated as a filled circle). Lineages in bold are sampled, and lineages not in bold are not sampled. Note that the age of the most recent common ancestor (MRCA) of the sample differs from the age of the MRCA of the entire population of descendents of M and from the age, t, of the mutation itself.

influence on mutation age estimates. We will explicitly consider the effects of variable population size (and exponential population growth, in particular), as this is essential in studying mutations in expanding human populations.

Two distinct approaches can be used to estimate allele age using these kinds of data: 1) phylogenetic methods; and 2) population genetic methods. A phylogenetic approach reconstructs the gene tree of a sample of mutant chromosomes—and possibly also a sample of normal chromosomes—estimating the age of the mutation as the age of the most recent common ancestor (MRCA) in the gene tree. Population genetic approaches typically use maximum likelihood, Bayesian inference, or the method of moments, to derive estimators of the age of a mutation that do not rely on a particular reconstruction of the gene tree of mutant chromosomes. This difference can be quite important as such reconstructions may be very uncertain. We distinguish two classes of population genetic approaches: 1) numerical statistical methods for estimating allele age using maximum likelihood or Bayesian techniques [Slatkin and Rannala 1997; Rannala and Slatkin, 1998; Markovtsova et al., 2000]; and 2) composite likelihood or moment-based methods that allow simple parametric estimators to be obtained analytically [Risch et al., 1995; Neuhausen et al., 1996; Guo and Xiong, 1997; Reich and Goldstein, 1999].

Population genetic methods using maximum likelihood, or Bayesian, approaches require an explicit model of population demography. Both population genetic and phylogenetic approaches require models of the processes of mutation (and/ or recombination) at linked marker loci, although this may not always be explicitly stated. Several of the composite likelihood, or method of moments, estimators do not require an explicit model of population demography. However, these methods either invoke specific assumptions about the biological processes of mutation, recombination, or population structure to achieve this simplicity or, as is the case with moment estimators, provide only a point estimate of the allele age with no associated confidence interval.

Phylogenetic Methods for Estimating Mutation Age

The phylogenetic approach to estimating allele ages is straightforward. If one imagines that

a completely accurate phylogenetic tree can be reconstructed for a sample of disease chromosomes using variation at closely linked genetic markers, then a simple estimator of the age of the mutation is the age of the root (or MRCA) in this tree (see Fig. 2). Even if uncertainty about the age of the root is ignored, problems remain with this approach. The age of the MRCA depends on the number of disease chromosomes that are sampled and is always less than, or equal to, the age of the mutation. This is illustrated in Figure 2. In principle, one could use the age of the MRCA of the sample of mutation-bearing chromosomes to place a lower bound on the age of a mutation and then use as an upper bound the age of the MRCA with both mutation-bearing and non-mutation-bearing descendents (using a sample of both mutation-bearing and non-mutation-bearing chromosomes). This approach is illustrated in Figure 2. In practice, the difference between the upper and lower bounds obtained in this way may often be large.

A second problem with a phylogenetic approach is that the age of the MRCA in the gene tree inferred using information from linked markers will often be poorly known. Such inferences require a model specifying how mutation (or re-
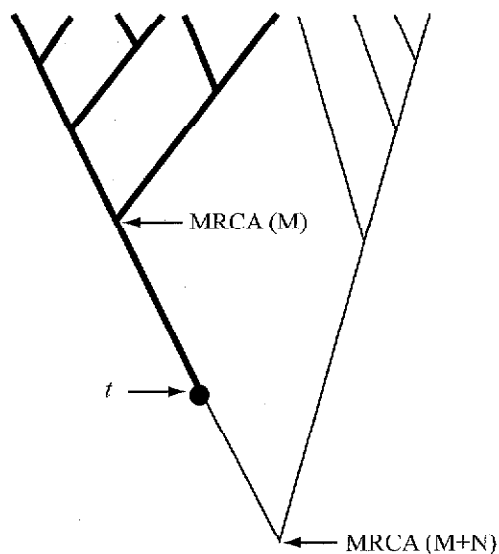

FIGURE 2. The genealogy of a sample of chromosomes bearing a mutation M (lineages in bold) that arose at time t (indicated by filled circle) and not bearing the mutation (lineages not bold). The age of the MRCA of the lineages descended from M places an upper bound on t, while the age of the first MRCA that has both non-mutation bearing, and mutation bearing, descendents places an upper bound on t.

combination) operates at the linked marker loci. In the case of microsatellite markers, most simplified models, such as the stepwise mutation model [Ohta and Kimura, 1973], do not adequately describe the actual mutation process. Realistic models will likely need to account for apparent heterogeneity in mutation rates among alleles at a locus [Valdes et al., 1993; Jin et al., 1996; Macaubas et al., 1997] and possible constraints on allele size. Although several recently developed models are a step in this direction [Di Rienzo et al., 1994; Kruglyak et al., 1998; Rose and Falush, 1998] their biological realism (and utility for phylogenetic analyses) have not been thoroughly investigated.

For mutations that are only a few hundred generations old there is the additional problem that even rapidly evolving linked markers, such as microsatellites, will have experienced few mutations during this time, providing limited information for inferring the genealogy, even when multiple tightly-linked markers are examined. As well, an accurate (preferably direct) estimate of mutation rates at the linked loci must be available. For young alleles, the phylogenetic approach to estimating allele age can therefore be expected to have low accuracy. A final troubling aspect of phylogeny-based methods for estimating the age of a mutation is that it is not currently possible to put confidence limits on the resulting estimates because phylogenetic uncertainty is not accounted for. Although, in principle, confidence limits for the estimated age of the MRCA in a gene tree could be obtained by bootstrap resampling [Felsenstein, 1985] we have not found any examples in the published literature in which this has been done.

In principle, any one of several approaches may be used to reconstruct the gene tree underlying a mutation from the patterns of genetic variation at tightly linked markers. However, applications to actual datasets have been mainly restricted to phylogenies obtained by distance methods, or maximum parsimony. In the case of mutations occurring in the mitochondrial genome, putatively neutral variation in the D-loop region has been used to reconstruct the underlying genealogy. This is possible because mtDNA experiences no recombination (or a very limited amount) and the ancestry of the D-loop region in a sample of individuals carrying a mutation is therefore an accurate reflection of the ancestry

of the mutation in the sample. Makino et al. [2000] used a phylogenetic analysis of the D-loop in normal Japanese subjects, and in Japanese individuals carrying two mutations in the ATPase 6 coding region (causing Leigh syndrome), to present evidence for multiple independent origins of each mutation. Although the ages of these mutations were not a focus of interest in this study, they could also have been inferred from the analysis.

If mutations are very ancient, DNA sequence variation in intronic regions can potentially be used to predict their ages in a phylogenetic analysis. For example, Bergstrom et al. [1999] carried out a phylogenetic analysis of intronic sequences to establish the relationships and ages of alleles at several MHC loci (see MIM# 142800). They concluded that alleles within major lineages were relatively young, having been generated within the last 250,000 years, although the major lineages appear to predate the separation of human and gorilla from a common ancestor. The MHC loci are unusual in many ways and are likely to be under the influence of overdominant selection [Doherty and Zinkernagel, 1975; Nei and Hughes, 1991], consequently harboring mutations that are much older than most neutral mutations with similar population frequency. Intronic variation is informative about the ages of mutations at MHC loci despite the low mutation rate (roughly 10^{-8} per base per generation) of most introns only because of their extreme ages.

For relatively young mutations, such as those underlying many rare genetic disorders, few mutations informative about age will have occurred at linked sites other than microsatellite loci. The majority of studies using phylogenies to infer mutation ages have therefore used variation at linked microsatellite loci. One of the early phylogenetic studies of the age of a specific mutation focused on the $\Delta F508$ mutation in the cystic fibrosis transmembrane regulator gene (CFTR; MIM# 602421), the most common cause of cystic fibrosis in Europeans [Morral et al., 1994]. These authors used the maximum parsimony (MP) algorithm to infer the gene tree underlying a sample of unrelated individuals carrying the $\Delta F508$ mutation (F508del) based on the variation observed at three closely linked microsatellite loci. The MP method infers the minimum number of mutations needed to account for the data on a particular gene tree; it

typically underestimates the actual number of mutations and incorporates assumptions very similar to those of an "infinite sites" model of DNA sequence mutation. The infinite sites model specifies that each nucleotide in a DNA sequence mutates at most once in the history of the sample [Watterson, 1975]. The inferred number of mutations on the genealogy was used in conjunction with a Poisson model of the mutation process and a direct estimate of the mutation rate [Weber and Wong, 1993] to predict the age of the root of the genealogy. This was estimated to be 2,625 generations, or 52,500 years.

Several studies aimed at reconstructing gene genealogies underlying specific mutations have used distance methods. A distance matrix among the haplotypes on which particular mutations are found is constructed using the pattern of mutations observed at linked markers. This is input into a tree-making method such as neighbor-joining [Saitou and Nei, 1987] to infer the gene tree. The relative lengths of the branches on this tree are used to predict the relative ages of alleles. An example is the study by Ajioka et al. [1997] that used a simple genetic distance based on the observed number of pairwise allelic differences between haplotypes at 24 marker loci, and the neighbor-joining algorithm, to construct a phylogeny relating the haplotypes of 85 unrelated hemochromatosis homozygous probands and 87 normal controls. This analysis suggested that the most common mutation associated with hemochromatosis arose quite recently. However, because the distance measure used by these authors does not increase linearly with time it is unlikely to be very accurate for determining the relative ages of all but the most recent mutations. A simple genetic distance has been proposed for use with microsatellite markers by Goldstein et al. [1995] that has an expected (average) value that is approximately linearly related to time. However, this distance is intended to be used for analyzing marker allele frequencies to infer the relationships of populations or species and cannot be readily applied to calculate distances between multilocus haplotypes. Slatkin [1995] provides an estimator of coalescent times between sampled chromosomes (as opposed to populations) based on microsatellite variation.

Simulation studies are needed to evaluate the relative performance of different phylogenetic

approaches for inferring gene trees underlying particular mutations; studies examining the accuracy of phylogenetic trees constructed from microsatellite markers using different genetic distances provide some guidance, suggesting that the Goldstein et al. [1995] distance can provide accurate reconstructions of the topology of a tree of closely related species when microsatellite loci mutate according to the stepwise mutation model [Takezaki and Nei, 1996]. However, new methods are needed for use with multiple linked loci. Most methods currently being used to reconstruct gene trees and infer ages of mutations derive their information from mutations occurring at completely linked markers. If recombination is considered, it is usually through approximate reconstructions of ancestral recombinant haplotypes. New methods are needed that incorporate both recombination and mutation in estimating ages of mutations by phylogenetic analysis.

Population Genetic Methods for Estimating Mutation Age

The processes of both recombination and mutation at linked markers provide potential information about the age of a specific mutation. This is illustrated in Figure 3. Population genetic approaches for estimating t require that statistical models of the processes of mutation and/or recombination be employed. Parametric likelihood and Bayesian methods require, in addition, that the population demography be explicitly modeled. The demographic model specifies the distribution of potential gene trees underlying the sampled sequences. Information about the age of a mutation arises from the decay of disequilibrium with nearby markers on disease chromosomes, which may be due to either recombination or mutation, and from the distribution of mutations among descendents. We illustrate this point for the simple case of a single marker locus linked to a disease mutation. We can define the coefficient of disequilibrium between mutation M and allele 0 at a linked marker (the allele present on the chromosome on which M arose) to be

$$D = p_{M0} - p_0,$$

where p_{M0} is the frequency of marker allele 0 on chromosomes bearing mutation M and p_0 is the
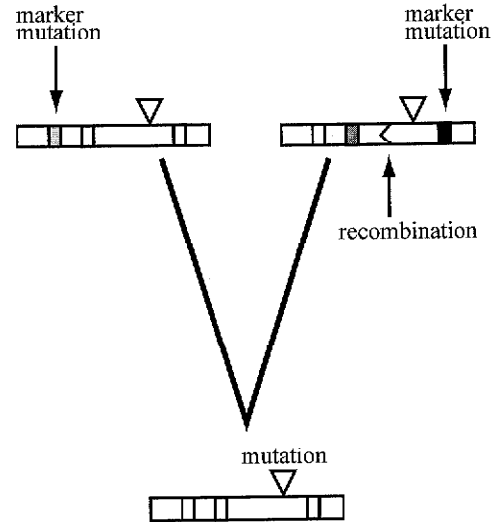

FIGURE 3. A simplified example of the several possible ways in which a pair of descendent haplotypes may be modified by comparison with the shared ancestral haplotype. The two branches linking the pair of chromosomes at the top of the figure with the ancestral chromosome at the bottom represent lines of descent through many ancestral chromosomes and may include thousands of meioses. Each chromosome is typed for three genetic markers represented as rectangles overlayed on the chromosomes. All chromosomes carry a particular mutation, M, denoted as an inverted triangle. The genealogy is that of mutation M and the three markers flanking this mutation may be altered by either recombination, mutation, or both. Different alleles at the linked markers are indicated by different shadings in the figure. The descendent haplotype at the upper left of the figure has experienced a mutation at a marker (the second marker to the left of M), while the descendent haplotype at the upper right of the figure has experienced both a mutation (to the right of M) and a recombination event (to the left of M). The region of the chromosome to the left of the recombination event is no longer inherited from the ancestral chromosome.

frequency of the allele on all chromosomes in the population. If recombination occurs with rate c per generation, then the expected (average) disequilibrium of marker allele on M-bearing chromosomes after t generations is

$$D_t = (1 - p_0)e^{-ct}. \qquad (1)$$

This result assumes that initially $p_{M0} = 0$ and that no genetic drift, migration, or selection is operating. The initial expected disequilibrium is $1 - p_0$, and this declines at an exponential rate as a function of both the number of generations (meiotic events) separating each chromosome from the ancestor on which M first arose and the recombination rate. For large t, the disequilibrium coefficient D approaches 0.

Recurrent mutation at linked markers can eliminate disequilibrium in a manner similar to recombination. To illustrate this, we consider a single nucleotide polymorphism (SNP) linked to M and assume a simple Jukes-Cantor model (J-C) [Jukes and Cantor, 1969] of DNA substitution. The J-C model specifies that changes to all four nucleotides are equally likely. If the mutation arises on a chromosome with nucleotide A at the linked site and all nucleotides have equal frequencies in the population (according to the stationary distribution for the Jukes-Cantor model) then (assuming no recombination) the expected disequilibrium of SNP nucleotide A on M-bearing chromosomes after t generations is

$$D_t = 3/4e^{-\mu t},$$

where μ is the mutation rate (per generation) at the SNP marker locus. More realistic models of DNA substitution are available [see Hillis and Moritz, 1996] and the J-C model was chosen merely to illustrate the concept. Different kinds of markers obviously require different models of mutation. Microsatellite loci, for example, which mutate by increasing or decreasing the number of short tandem repeat sequences, require more elaborate mutation models than do SNPs.

With multiple linked marker loci, the decay of LD under the processes of either recombination, mutation, or both, is quite complex. In addition to the simple decay of expected LD at linked markers under mutational or recombinational pressures, information about the age of a mutation (via its underlying gene genealogy) is also available from the full spectrum of haplotypes in the sample. In the case of recombination, information about the age of a mutation arises mainly from the decay of LD at linked markers given known rates of recombination (from a linkage map), whereas in the case of mutation, information arises both from the numbers of new alleles at the linked markers, given known rates of mutation, and the decay of LD. Both processes decrease the association between a disease mutation and an ancestral haplotype.

To develop parametric statistical methods for estimating mutation age, information is needed about a number of population demographic parameters, in addition to the sample of haplotypes on mutation-bearing and normal chromosomes, and the marker mutation rates and/or recom-

bination rates. We outline these additional parameters and their relation to the statistical likelihood of the data below. Here, we present the parameters outlined by Rannala and Slatkin [1998, 2000], who used the intraallelic coalescent for a rare disease mutation to develop an estimator of allele age. Other authors that instead use the population coalescent [e.g., Markovtsova et al., 2000] consider a slightly different set of parameters. Let $X = \{X_{ij}\}$ where X_{ij} denotes the allele observed at the jth marker of the ith M-bearing chromosome. Let $\Omega = \{\alpha, p, X_0, \mu\}$ be a vector of parameters of the models of mutation and recombination. This minimally includes the map distances between the L marker loci $\alpha = \{\alpha_1, \alpha_2, ..., \alpha_{L-1}\}$, the map position relative to marker locus 1, denoted as θ, of the mutation whose age is of interest, the matrix of allele frequencies on normal chromosomes $p = \{p_{ij}\}$ (estimated from a sample of normal chromosomes), where p_{ij} is the frequency of allele i at locus j, the ancestral haplotype, X_0, on which the mutation arose, and a vector $\mu = \{\mu_1, \mu_2, ..., \mu_L\}$ of locus-specific mutation rates where μ_i is the mutation rate at locus i. The parameter $\Theta = \{f, \xi\}$ is a vector of the demographic parameters affecting the likelihood and includes the population growth rate, ξ, and the fraction, f, of the total population of extant chromosomes bearing mutation M that are in the sample. The likelihood is

$$\Pr(X|t;\Omega,\Theta) = \int g(X \mid \tau;\Omega)f(\tau|t;\Theta)d\tau. \quad (2)$$

The parameter τ is the gene tree which includes both a tree topology and the set of $n - 1$ times, in the past, at which the n sampled chromosomes coalesce to shared ancestral chromosomes. This is illustrated in Figure 4. The gene tree is an unobserved random variable and it is therefore preferable to integrate over gene trees to obtain the marginal probability of the data (which is independent of the gene tree). This integration procedure is presented in Eq. (2), where the integral is multidimensional and involves summing over all possible gene trees and integrating over all possible coalescence times. The ancestral haplotype, X_0, on which the mutation first arose is unknown but may be estimated from the data (often it will be clear from the pattern of disequilibrium on disease chromosomes). The frequencies on normal chromosomes can be used
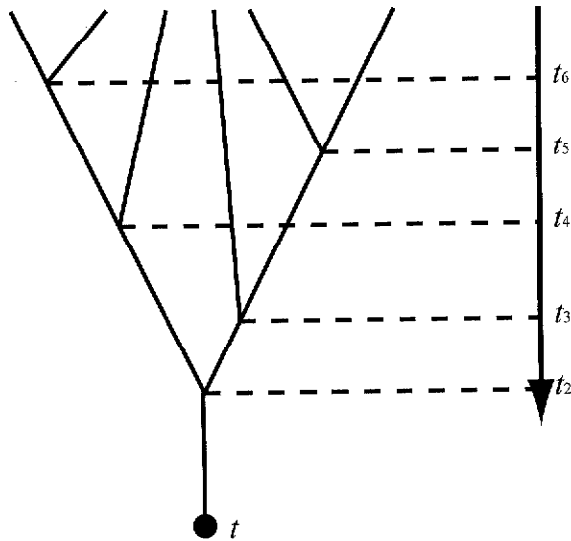
FIGURE 4. The gene tree and the coalescence times t_2 through t_6 of the genealogy underlying a sample of chromosomes descended from a mutation M. The coalescence times and the gene tree are unobserved random variables. The time at which the mutation arose is denoted as t.

to provide a prior probability distribution for X_0 [see e.g., Terwilliger, 1995].

Likelihood Methods

Slatkin and Rannala [1997] and Rannala and Slatkin [1998] developed Monte Carlo integration methods for evaluating Eq. (2) for either a model with recombination between a pair of linked markers, and no mutation [Rannala and Slatkin, 1998], or a model with no recombination between markers and only non-recurrent marker mutation [Slatkin and Rannala, 1997]. The model of marker mutation used by Slatkin and Rannala [1997] is appropriate when most mutation events result in new (previously unobserved) mutations and the method is best suited for rapidly evolving markers, such as microsatellites, located very near the mutation whose age is to be estimated (generally less than 100 kb). Slatkin and Rannala [1997] used the number of mutant alleles, S, at a set of completely linked marker loci as the observed data in their analysis, assuming no recombination.

A Bayesian Markov chain Monte Carlo (MCMC) method recently developed by Markovtsova et al. [2000] uses DNA sequence data, and a finite sites model of the mutation process (assuming no recombination), to infer the age of a sample of chromosomes carrying a

particular mutation. Apart from the fact that one method uses sequence data and the other uses linked markers, there are other differences between the method of Slatkin and Rannala [1997] and that of Markovtsova et al. [2000]. The main difference is that Slatkin and Rannala [1997] considered only the genealogy of a sample of chromosomes descended from a particular mutation (the intraallelic coalescent), whereas Markovtsova et al. [2000] considered a sample composed of both chromosomes that are descended from a particular mutation and chromosomes that are not (the population coalescent).

The methods of Slatkin and Rannala [1997] and Rannala and Slatkin [1998] are intended for use with rare mutations (population frequency less than about 1%). An implicit assumption of the method of Rannala and Slatkin [1998] is that recombination events in the history of the sampled chromosomes occurred exclusively in heterozygotes. The relative proportion of alleles transmitted through heterozygotes versus homozygotes in any generation (assuming Hardy-Weinberg equilibrium) is $2(1 - p)/p$. If $p < 0.01$, the relative proportion of mutation-bearing chromosomes transmitted through heterozygotes is more than 200 times the proportion transmitted through homozygotes, making the assumption reasonable. The method of Rannala and Slatkin [1998] also assumes that marker allele frequencies on normal chromosomes have remained constant during the time since the mutation arose. This is also most likely to be the case for relatively young alleles in low frequency in a population. The method of Markovtsova et al. [2000] could be used to estimate ages of common mutations as well as rare mutations, but assumes that the chromosomes are sampled at random from the population, which is typically not the case for chromosomes bearing a disease mutation.

The population demographic model employed by Markovtsova et al. [2000] provides a prior distribution for the range of plausible ages of a mutation before the intraallelic variability is examined (determined by the frequency of the mutation and a neutral model of population demography) and this is updated with information from the sequence data to obtain an estimate of the mutation age based on both the population genetic model and the patterns of

mutation among the DNA sequences. Slatkin and Rannala [1997] instead condition on the age of the mutation, t, to obtain a maximum likelihood estimate of this parameter without implementing prior information from a population genetic model.

If an estimate of the age of a mutation obtained using the Markovtsova et al. [2000] method differs greatly from an estimate obtained using the method of Slatkin and Rannala [1997] this may indicate that the population genetic model is having a large influence on the estimated age. Since the neutral population genetic model [based on the theory of Kingman, 1982] employed by Markovtsova et al. [2000] is an oversimplification, and will be violated in structured populations or for mutations that are under natural selection, such a discrepancy may be an indication that the estimate of mutation age is unreliable because one or more of these factors are operating. In general, the confidence interval for the estimated age of a mutation obtained using the method of Slatkin and Rannala [1997] should be wider than that obtained using a method such as that of Markovtsova et al. [2000], and analyzing the same data, because no prior constraints are placed on the possible age of t based on a population genetic model.

Rannala and Slatkin [1998] used a model of the process of recombination between a particular mutation and a linked marker (assuming no marker mutation) to estimate the age of the mutation. The method is best suited for markers with low rates of mutation, such as SNPs and RFLPs, that are located at map distances from a mutation ranging from 0.01 cM (roughly 10 kb) to 1 cM (roughly 1 Mb). Figure 5 shows the likelihood of the data of Hastbacka et al. [1992] as a function of the age of the mutation that causes diastrophic dysplasia (DTD; also SLC26A2; MIM# 222600). This likelihood function is based on a model of recombination only. Figure 5 was produced using the program DMLE (available from www.rannala.org). The data are observed marker haplotypes in samples of normal and disease-associated chromosomes [Hästbacka et al., 1992] having undergone recombination with a linked RFLP marker (EcoRI) that is roughly 70 kb telomeric to a particular DTD mutation in the Finnish population [Hästbacka et al., 1994]. Roughly 90% of the DTD mutations in this sample are descended from a single
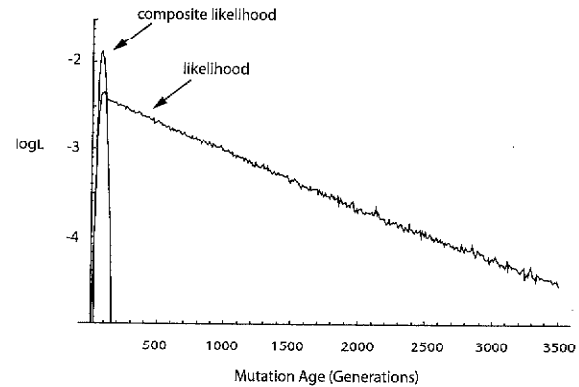

FIGURE 5. Log-likelihood as a function of the age of the diastrophic dysplasia (DTD) mutation (in units of generations) in the Finnish population. The data is from Hästbacka et al. [1992]. A single linked RFLP marker was used for this analysis (EcoRI). The approximate "composite" likelihood is shown, as well as the exact likelihood calculated numerically using the program DMLE. The composite likelihood is intended to approximate the true likelihood but in this case provides a confidence interval (determined by the width of the log-likelihood curve) that is too narrow.

GT->GC transition in a 5' untranslated exon of the SLC26A2 gene [Hästbacka et al., 1999]. It is the age of this mutation event that we are interested in estimating. The point estimate of the age of this mutation obtained by maximum likelihood is $t = 75$ generations, but the 95% confidence interval for the age estimate is quite large with a minimum age of 19 generations and a maximum age of 3,170 generations. Additional parameters used for this analysis are given in Rannala and Slatkin [1998]. As illustrated in this example, estimates of mutation age are often quite imprecise.

As an alternative to exact likelihood methods, several approximate methods have been developed which rely on a "composite" likelihood (CL). These methods treat joint probabilities as products of marginal probabilities, which assumes independence among variables that are actually dependent [see Rannala and Slatkin, 2000]. The methods therefore ignore one or more of the sources of uncertainty in estimates of mutation age and typically provide confidence intervals for estimates of t that are too narrow. A strength of these methods is that their simplicity can allow mutation and recombination to be modeled simultaneously [see Guo and Xiong, 1997]. A simple moment estimator obtained by setting the expected haplotype frequency equal to the observed frequency and

solving for the recombination rate provides estimates of allele age identical to those obtained using a particular type of CL [Rannala and Slatkin, 1998]. Below, we briefly discuss the approximate CL methods that have been developed, clarifying their assumptions.

Approximate Methods

Approximate statistical methods for estimating the age of a mutation, t, fall into two general categories. Rannala and Slatkin [2000] refer to these as type I and type II CLs. Estimators of mutation age based on a type I CL approximation [e.g., Risch et al., 1995; Guo and Xiong, 1997] ignore the dependence among chromosomes created by their shared genealogy. Those based on a type II approximation [e.g., Guo and Xiong, 1997] ignore the dependence of recombination events occurring among multiple linked markers, treating each pairwise recombination process as independent even though many recombination events affect multiple pairs of markers. Although the probability of multiple recombination events among closely linked markers in a single meiosis may be negligible, the probability of multiple recombination events among markers over the length of a branch on the gene tree is often large and cannot be neglected. Some authors employ both type I and II CLs [Neuhausen et al., 1996]. The relationship between the type I and type II CL approximations is further discussed in Rannala and Slatkin [2000].

If a type I CL approximation is used, and a single linked marker locus is considered, maximizing the CL results in the estimator [see Rannala and Slatkin, 2000 for a derivation]

$$\hat{t} = \frac{1}{\theta}\{\log(1-p_0) - \log(Y_0 - np_0) + \log(n)\},$$

where θ is the recombination rate per generation (map distance) between the marker and the mutation, p_0 is the frequency of the ancestral marker (the marker that was present on the chromosome on which the mutation first arose) on normal chromosomes (those not carrying the mutation), Y_0 is the number of chromosomes carrying both the mutation and the ancestral marker, and n is the total number of mutation bearing chromosomes that are sampled. An

equivalent estimator for use with mutational variation at tightly linked markers can be obtained by simply replacing θ with the mutation rate per generation, μ, where p_0 is then the probability of a reverse mutation to the ancestral allele, given that a mutation occurs. Assuming no reversals, this simplifies to [Risch et al., 1995]

$$\hat{t} = \frac{1}{\mu}\{\log(n) - \log(Y_0)\}.$$

Instead of explicitly modeling gene genealogy, Guo and Xiong [1997] used a diffusion theory model of population allele frequency evolution to derive a maximum likelihood method for estimating t. There is no analytical expression for their likelihood and they instead consider approximate likelihood estimators based on first and second order (linear and quadratic) Taylor series approximations of the likelihood. The first-order approximate likelihood of Guo and Xiong [1997] is equivalent to a type I CL. They also consider a type II CL approximation for treating multiple linked marker loci and use a stepwise mutation model to allow for mutations at linked microsatellite loci.

The log-likelihood surfaces as a function of the map position of the DTD mutation obtained relative to a single RFLP marker (EcoRI) linked to the DTD mutation [Hästbacka et al., 1992] using either a type I CL approximation, or an exact numerical likelihood calculation (using DMLE), are shown in Figure 5. The CL method generates a confidence interval (determined by the width of the log-likelihood surface) that is too narrow by comparison with the exact likelihood, as expected. The 95% confidence intervals (min t, max t) for the estimated age of the DTD mutation (in units of generations) are (32, 141) for the CL-based method and (19, 3170) for the ML-based method. Although the use of approximate CL methods greatly simplifies the derivation of estimators of mutation age, these approaches are only approximate and, at best, may be expected to perform only as well as ML estimators. In many cases, they may perform much worse. In general, CL methods produce confidence intervals that are too optimistic. The point estimates may also have larger mean square error than those obtained using likelihood, and

are not guaranteed to satisfy the large-sample statistical property of consistency. ML or Bayesian methods should be employed, in preference to CL methods, when this is feasible. The computer time required to carry out the calculations needed for the exact numerical methods using Monte Carlo integration or MCMC will depend on the number of chromosomes and markers sampled. For very large sample sizes, the computational demand of these methods may still be too great to make them practical; an approximate method may then be the only option available.

Relative Timescales of Mutation and Recombination

Often, a geneticist interested in estimating the age of a particular mutation (a disease mutation in an exon, for example) will have the choice of using either tightly linked (e.g., intragenic) markers with high mutation rates (e.g., microsatellite markers located in introns) or more distantly spaced markers with lower rates of mutation (e.g., SNPs located outside the gene). In the first case, information about genealogy arises from mutations occurring at the linked markers. In the second case, it arises from recombination events between the linked markers. These two sources of genealogical information operate on different timescales and this will affect the choice of marker in particular cases. The use of information arising from recombination among markers has the advantage that it allows one to potentially fine-tune the recombination rates to optimize the power of a study for inferring the age of a mutation by choosing markers at optimal map distances.

Typically, linked SNPs for which accurate map distances are available (from prior linkage analysis on pedigrees, for example) will have recombination rates ranging from 1 to 10 cM (e.g., θ = 0.01 to 0.1). Markers with these map distances will be most effective in resolving the ages of young mutations (ranging in age from 10 to 100 generations). For example, using equation 1 and assuming that p_0 = 0.01 (i.e., the ancestral marker is found on only 1% of normal chromosomes) we find that if θ = 0.01 and t − 25 then D = 0.77, and if t = 100 then D = 0.36, but if t = 500 then D = 0.08. In this last case, the expected LD is probably too low to be useful for inferring the age of a linked mutation. On the

other hand, if θ = 0.1 we find that if t = 10 then D = 0.36, and if t = 25 then D = 0.08. Thus, if we were interested in estimating the age of a mutation suspected to be very young, say 10 generations, then a marker at a distance of 10 cM would be optimal, while for a mutation 100 generations old, a marker at a distance of 1 cM would be optimal. In this example, the expected disequilibrium for a marker at a distance of 10 cM is much too low (D = 0.000045) to be detectable with realistic sample sizes.

At present, linkage maps available for most genomic regions have a resolution of less than 1 cM, limiting the usefulness of the methods for estimating mutation age (via recombination) to mutations younger than about 100 generations. Some authors have tried to overcome this limitation by assuming that a linear relationship exists between a radiation hybrid (RH) map and a linkage map [Stephens et al., 1998] and using linear regression to predict linkage map distances from RH map distances which are often available at a higher resolution. However, the relationship between RH and linkage maps is poorly understood at present limiting the accuracy of such techniques.

An alternative approach for creating a high resolution linkage map is to use the average relationship between recombination rate and physical map distance to predict the linkage map from the physical map of the markers. This is particularly attractive with the availability of a complete human genome sequence. Because the sex-averaged length of the human linkage map is roughly 3000 cM and the physical size of the human genome is roughly 3 Gb [Ott, 1999] the relationship 1 cM = 1 Mb provides a rough index for translating the physical map into a linkage map. By this approach, one could in principle obtain a linkage map at any resolution. However, large variation in rates of recombination across different regions of the genome should be accounted for in such a method; this could be done by implementing a specific model of rate variation based on the variance observed among regions in the low resolution linkage map. Another solution would be to carry out population LD studies to construct a high-resolution linkage map of humans. Such a map would be very valuable for studies of mutation ages and other demographic parameters in human populations.

DISCUSSION

The methods that have been developed for estimating the age of a mutation using the variation observed at closely linked genetic markers can be broadly categorized as either "direct" phylogenetic methods, which attempt to reconstruct the topology and branch lengths of the underlying gene tree, or "indirect" population genetic methods that model the processes of recombination, mutation, and population demography, to obtain estimators of mutation age that are not dependent on a specific gene tree. Each approach has its strengths and weaknesses. In general, statistical approaches should provide more accurate inferences of allele age when alleles are relatively young and no population subdivision exists. Phylogenetic approaches, on the other hand, should be more accurate for inferring ages of very ancient mutations, especially in the presence of population subdivision, or other demographic complications.

The influence of population subdivision on statistical estimators of the age of a mutation has not been adequately explored; it is critical that population subdivision be accounted for in future theoretical developments as it is an important aspect of human demographic history. Although population geneticists have long been interested in studying the influence of population structure on gene frequencies in populations at equilibrium [Wright, 1931; Kimura and Weiss, 1964; Maruyama, 1971], most human populations are far from equilibrium and, in any case, the existing theory of genetic structure in subdivided populations does not specifically address the distribution of the ages of specific mutations in either equilibrium, or non-equilibrium, populations. New models and statistical methods are needed.

A final factor that has received too little attention is the geographic distribution of mutations and the influence that this may have on estimates of their ages. Intuitively, it should be expected that more widespread mutations will tend to be older, and we have suggested in a recent study [Bertorelle and Rannala, 1998] that the relative frequencies of alleles within different populations may be used to infer the times at which the populations diverged. There is likely to be additional information about these divergence times contained in the pattern of mutations at linked loci associated with a particular mutation among populations. Little formal mathematical analysis has been carried out on this problem, however, and so far few data are available.

An interesting potential application of methods for estimating ages of mutations is to the study of patterns of selection in the human genome. Natural selection can potentially decouple the frequency of a mutation from its age. For example, two mutations with the same population frequency may have very different ages if one is neutral and the other is under the influence of overdominant selection. This decoupling can provide a signal for detecting genes and mutations that are under selection. If selection is operating on a mutation, this may potentially result in a large difference between estimates of the age of the mutation based on either variation at linked genetic markers, or the population frequency of the mutation. This is because the relative influence of selection on each type of estimator may be quite different. A large discrepancy between an estimate of mutation age based on linked genetic markers and one based on the population frequency of a mutation may therefore be an indication that selection is operating on the mutation, even if both estimates are obtained by assuming that the mutation is neutral.

The intraallelic coalescent model considered by Slatkin and Rannala [1997] is influenced by selection and population growth through a common parameter which is a sum of the two effects. Thus, for a single locus, one cannot separate the effects of selection versus population growth. However, although a common population growth rate applies across all genes (and mutations) selection coefficients will usually vary from one gene and/or mutation to another. This suggests that one could use the intraallelic marker variation associated with several different mutations to identify a subset that are under selection. To do this, one would need to jointly estimate the allele age and ξ (selection coefficient + population growth rate) for each mutation. In some cases, the population growth rate may be separately estimated from other sources (demographic records for a founder population, for example). Estimated values of ξ for particular mutations that differ greatly from the population growth rate might then indicate selection.

For many mutations involved in human disease the frequencies of particular mutations

among populations are now being catalogued and potential information on variation at linked markers might then also then be collected. The use of linked genetic markers to study the ages and histories of mutations in human populations is only beginning. We expect that with improved high-throughput genotyping methods and large-scale population screening of disease mutations in human populations many interesting patterns will emerge. We also expect that population geneticists will continue to develop better tools for analyzing these data; improving existing approaches, as well as developing new methods that are up to the tasks ahead.

ACKNOWLEDGMENTS

REFERENCES

Ajioka RS, Jorde LB, Gruen JR, Yu P, Dimitrova D, Barrow J, Radisky E, Edwards CQ, Griffen LM, Kushner JP. 1997. Haplotype analysis of hemochromatosis: evaluation of different linkage-disequilibrium approaches and evolution of disease chromosomes. Am J Hum Genet 60:1439–1447.

Bergstrom TF, Erlandsson R, Engkvist H, Josefsson A, Erlich HA, Gyllensten U. 1999. Phylogenetic history of hominoid DRB loci and alleles inferred from intron sequences. Immunol Rev 167:351–365.

Bertorelle G, Rannala B. 1998. Using rare mutations to estimate population divergence times: a maximum likelihood approach. Proc Natl Acad Sci USA 95:15452–15457.

Cavalli-Sforza LP, Menozzi P, Piazza A. 1994. History and geography of human genes. Princeton NJ: Princeton University Press.

Di Rienzo A, Peterson AC, Garza JC, Valdes AM, Slatkin M, Freimer NB. 1994. Mutational processes of simple-sequence repeat loci in human populations. Proc Natl Acad Sci USA 91:3166–3170.

Doherty PC, Zinkernagel RM. 1975. Enhanced immunological surveillance in mice heterozygous at the H-2 gene complex. Nature 256:50–52.

Ewens WJ. 1979. Mathematical population genetics. Berlin: Springer-Verlag.

European Working Group on CF Genetics. 1990. Gradient of distribution in Europe of the major CF mutation and of its associated haplotype. Hum Genet 85:436–441.

Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791.

Fisher RA. 1930. The genetical theory of natural selection. Oxford: Clarendon Press.

Goldstein DB, Ruiz Linares A, Cavalli-Sforza LL, Feldman MW. 1995. Genetic absolute dating based on microsatellites and the origin of modern humans. Proc Natl Acad Sci USA 92:6723–6727.

Guo SW, Xiong M. 1997. Estimating the age of mutant disease alleles based on linkage disequilibrium. Hum Hered 47:315–337.

Haldane JBS. 1932. The causes of evolution. London: Longmans and Green.

Harris H. 1966. Enzyme polymorphisms in man. Proc Roy Soc Lond B 164:298–310.

Hästbacka J, de la Chapelle A, Kaitila I, Sistonen P, Weaver A, Lander ES. 1992. Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. Nat Genet 2:204–211.

Hästbacka J, de la Chapelle A, Mahtani MM, et al. 1994. The diastrophic dysplasia gene encodes a novel sulphate transporter: positional cloning by fine-structure linkage diseqiulibrium mapping. Cell 78:1073–1087.

Hästbacka J, Kerrebrock A, Mokkala K, Clines G, Lovett M, Kaitila I, de la Chapelle A, Lander ES. 1999. Identification of the Finnish founder mutation for diastrophic dysplasia. Eur J Hum Genet 7:664–670.

Hillis DM, Moritz C. 1996. Molecular systematics. Sunderland, Mass: Sinauer Associates. p 430–445.

Jin LC, Macaubas J, Hallmayer A, Kimura A, Mignot E. 1996. Mutation rate varies among alleles at a microsatellite locus: phylogenetic evidence. Proc Natl Acad Sci USA 93:15285–15288.

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In: Munro HN, editor. Mammalian protein metabolism. New York: Academic Press. p 21–132.

Kimura M, Weiss GH. 1964. The stepping stone model of population structure and the decrease of genetic correlation with distance. Genetics 49:561–576.

Kingman JC. 1982. On the genealogy of large populations. J Appl Prob 19A:27–43.

Kruglyak S, Durrett RT, Schug MD, Aquadro CF. 1998. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. Proc Natl Acad Sci USA 95:10774–10778.

Lander ES, Linton RM, Birren B, Nusbaum C, Zody M, Baldwin J, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.

Lewontin RC, Hubby JL. 1966. A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of Drosophila pseudoobscura. Genetics 54:595–609.

Macaubas C, Jin L, Hallmayer J, Kimura A, Mignot E. 1997. The complex mutation pattern of a microsatellite. Genome Res 7:635–641.

Makino M, Horai S, Goto Y, Nonaka I. 2000. Mitochondrial DNA mutations in Leigh syndrome and their phylogenetic implications. J Hum Genet 45:69–75.

Markovtsova L, Marjoram P, Tavare S. 2000. The age of a unique event polymorphism. Genetics 156:401–409.

Maruyama T. 1971. Analysis of population structure II. Two-dimensional stepping stone models of finite length and other geographically structured populations. Ann Hum Genet 35:179–196.

Maruyama T. 1974a. The age of a rare mutant gene in a large population. Am J Hum Genet 26:669–673.

Maruyama T. 1974b. The age of an allele in a finite population. Genet Res 23:137–143.

Morral N, Bertranpetit J, Estivill X, Nunes V, Casala T, Gimenez J, Reis A, Varon-Mateeva R, Macek M, Kalaydjieva L, Angelicheva D, Dancheva R, Romeo G, Russo MP, Garnerone S, Restagno G, Ferrari M, Magnani C, Claustres M, Desgeorges M, Schwartz M, Schwarz M, Dallapiccola B, Novelli G, Ferec C, de Arce M, Nemeti M, Kere J, Anvret M, Dahl N, Kadasi L. 1994. The origin of the major cystic fibrosis mutation (ΔF508) in European populations. Nat Genet 7:169–175.

Nei M, Hughes AL. 1991. Polymorphism and evolution of the major histocompatibility complex loci in mammals. In: Selander RK, Clark AG, Whittam TS, editors. Evolution at the molecular level. Sunderland, MA: Sinauer Associates. p 222–247.

Neuhausen SL, Mazoyer S, Friedman L, Stratton M, Offit K, Caligo A, Tomlinson G, Cannon-Albright L, Bishop T, Kelsell D, Solomon E, Weber B, Couch F, Struewing J, Tonin P, Durocher F, Narod S, Skolnick MH, Lenoir G, Serova O, Ponder B, Stoppa-Lyonnet D, Easton D, King MC, Goldgar DE. 1996. Haplotype and phenotype analysis of six recurrent BRCA1 mutations in 61 families: results of an international study. Am J Hum Genet 58:271–280.

Ohta T, Kimura M. 1973. A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet Res 22:201–204.

Ott J. 1999. Analysis of human genetic linkage. Third edition. Baltimore: John Hopkins University Press.

Rannala B, Slatkin M. 1998. Likelihood analysis of disequilibrium mapping, and related problems. Am J Hum Genet 62:459–473.

Rannala B, Slatkin M. 2000. Methods for multipoint disease mapping using linkage disequilibrium. Gen Epidem 19(suppl 1):S71–S77.

Reich DE, Goldstein DB. 1999. Estimating the age of mutations using variation at linked markers. In: DB Goldstein, C Schlotterer, editors. Microsatellites: evolution and applications. Oxford: Oxford Univ Press. p 129–138.

Risch NJ, de Leon D, Ozelius L, Kramer P, Almasy L, Singer B, Fahn S, Breakefield X, Bressman S. 1995. Genetic analysis of idiopathic torsion dystonia in Ashkenazi Jews and their recent descent from a small founder population. Nat Genet 9:152–159.

Risch NJ. 2000. Searching for genetic determinants in the new millenium. Nature 405:847–856.

Roberts EA, Cox DJ. 1998. Wilson disease. Baillieres Clin Gastroenterol 12:237–256.

Rose O, Falush D. 1998. A threshold size for microsatellite expansion. Mol Biol 15:613–615.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425.

Slatkin M. 1995. A measure of population subdivision based on microsatellite allele frequencies. Genetics 139:457–462.

Slatkin M, Rannala B. 1997. Estimating the ages of alleles by use of intraallelic variability. Am J Hum Genet 60:447–458.

Slatkin M, Rannala B. 2000. Estimating allele age. Annu Rev Genomics Hum Genet 1:225–249.

Stephens JC, Reich DE, Goldstein DB, Shin HD, Smith MW, Carrington M, Winkler C, Huttley GA, Allikmets R, Schrim L, Gerrard B, Malasky M, Ramos, MD, Morlot S, Tzetis M, Oddoux C, di Giovine FS, Nasioulas G, Chandler D, Asev M, Hanson M, Kalaydjieva L, Glavac D, Gasparini P, Kanavakis E, Claustres M, Kambouris M, Ostrer H, Duff G, Baranov V, Sibul H, Metspalu A, Goldman D, Martin N, Duffy D, Schmidtke J, Estivill X, O'Brien SJ, Dean M. 1998. Dating the origin of the CCR5-Δ32 AIDS-resistance allele by the coalescence of haplotypes. Am J Hum Genet 62:1507–1515.

Takezaki N, Nei M. 1996. Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. Genetics 144:389–399.

Terwilliger JD. 1995. A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. Am J Hum Genet 56:777–787.

Valdes AM, Slatkin M, Freimer NB. 1993. Allele frequencies at microsatellite loci: the stepwise mutation model revisited. Genetics 133:737–749.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. 2001. The sequence of the human genome. Science 291:1304–1351.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. Theor Popul Biol 7:256–276.

Weber JL, Wong C. 1993. Mutation of human short tandem repeats. Hum Mol Genet 2:1123–1128.

Wright S. 1931. Evolution in Mendelian populations. Genetics 16:97–159.