Estimating Mutation Rate: How to Count Mutations?

Yun-Xin Fu[1] and Haying Huai

Human Genetics Center, University of Texas, Houston, Texas 77030

Manuscript received August 28, 2002
Accepted for publication March 11, 2003

ABSTRACT

Mutation rate is an essential parameter in genetic research. Counting the number of mutant individuals provides information for a direct estimate of mutation rate. However, mutant individuals in the same family can share the same mutations due to premeiotic mutation events, so that the number of mutant individuals can be significantly larger than the number of mutation events observed. Since mutation rate is more closely related to the number of mutation events, whether one should count only independent mutation events or the number of mutants remains controversial. We show in this article that counting mutant individuals is a correct approach for estimating mutation rate, while counting only mutation events will result in underestimation. We also derived the variance of the mutation-rate estimate, which allows us to examine a number of important issues about the design of such experiments. The general strategy of such an experiment should be to sample as many families as possible and not to sample much more offspring per family than the reciprocal of the pairwise correlation coefficient within each family. To obtain a reasonably accurate estimate of mutation rate, the number of sampled families needs to be in the same or higher order of magnitude as the reciprocal of the mutation rate.

A significant fraction of the genetic research of the last century has been to illuminate various aspects of mutation (De Vries 1901/1903; Luria and Delbrück 1943; McClintock 1950; Keightley and Eyre-Walker 1999). This is natural because mutations are the ultimate source of genetic variation upon which natural selection and other evolutionary forces can act (Kimura 1983; Lynch and Hill 1986; Johnson 1999). Early experiments on mutation rate include those by Castle (1905, 1929), Muller (1920, 1928), and Morgan (1950). To date, extensive mutation data, from either mutation experiments or surveys, are available for many species, particularly fruit flies (Schalet 1960; Crow and Simmons 1983), mice (Favor and Neuhauser-Klaus 1994; Russell and Russell 1996), and humans (Neel and Rothman 1978; Cooper and Krawczak 1993; Crow 1993, 1999). Due to the importance of mutation rate, such experiments will be likely to continue in the future, with more and more details being revealed by the advent of new molecular techniques (Kondrashov and Crow 1993; Fu 1994).

In a typical mutation experiment, some aspects of the progeny of well-characterized parents are examined. A mutant is identified if an offspring differs from its parents in a way that can be explained only by invoking a mutation (Auerbach 1959; Drake 1991, 1993). When a large number of offspring of mating pairs have been examined, the proportion of mutant progeny yields a direct estimate of the rate of mutation. These experiments may be time consuming, but the statistical method used for estimating mutation rate is straightforward and should not be controversial. It appeared indeed to be the case for early geneticists (Bridges 1919; Wright and Eaton 1926; Fisher 1930; Dobzhansky and Wright 1941). However, the increasing number of observations that some mutant offspring share the same mutation has prompted many contemporary geneticists to reconsider how mutation rate should be estimated (Engels 1979; Russell and Russell 1996; Neel 1998; Thompson *et al.* 1998).

A clustered mutation means that two or more progeny of a family inherit the same mutation (Purdom *et al.* 1968; Hartl and Green 1970; Favor and Neuhauser-Klaus 1994). Mutation clusters have been widely observed and are now considered as general rather than as the exception (Hall 1988; Drost and Lee 1995; Mohrenweiser and Zingg 1995; Huai and Woodruff 1997; Paashuis-Lew and Heddle 1998; Lewis 1999; for reference, see Woodruff *et al.* 1996). The most important issue created by mutation cluster is how to count the mutations for the purpose of estimating mutation rate. Several ways of counting have been proposed. One is to count each mutant offspring as one mutation, disregarding whether or not the mutation is shared (Haldane 1935; Spencer and Stern 1948; Auerbach 1962; Muller *et al.* 1963; Combes *et al.* 1989; Huai 1997). The second is to count each cluster as only one mutation (Russell 1977; Shukla *et al.* 1979; Heddle *et al.* 1996; Nishino *et al.* 1996). The third is to count only those mutations that are not clustered (Abrahamson and Wolff 1976; Russell and Russell 1992),

[1] *Corresponding author:* Human Genetics Center, SPH, University of Texas, 1200 Herman Pressler, Houston, TX 77030.
E-mail: yunxin.fu@uth.tmc.edu

but in many cases this third choice is made because only the induced mutations in limited stages of the life cycle are measured (Mason *et al.* 1987; Arrault *et al.* 2002). Various arguments have been put forward to support one method or the other (Russell and Russell 1992, 1996; Huai and Woodruff 1998a,b; Heddle 1999), but no resolution has been obtained to date (Thompson *et al.* 1998; Stuart and Glickman 2000).

In this article, we show for the first time that counting each mutant as one mutation regardless of cluster is the correct way to obtain an unbiased estimate of the mutation rate. Of equal significance is the formula for the sampling variance of the estimator, which does not require knowing all family sizes. We also discuss two important issues in designing a mutagenesis experiment, the sampling strategy and sample size requirement. Furthermore, we reanalyze several large data sets and show that some of the results in mutation rate estimates have an undesirably large variance.

THEORY

Counting mutations: Suppose a total of m haploid families are studied in an experiment. Let n_i be the number of offspring examined in the ith family and $n = n_1 + \ldots + n_m$ be total sample size. Considering the lth family, an individual is a mutant if it differs from its parent(s) by at least one mutation. It is important to realize that mutations in different mutant individuals are not necessarily distinct. Let K_l be the total number of mutations counted as follows: each mutation that is inherited by k individuals (more precisely k sequences) is counted k times. That is, suppose that there are in total I independent mutation events, and the ith mutation is inherited by h_i sequences. Then

$$K_l = h_1 + \ldots + h_I. \tag{1}$$

Note that if there is only one mutant for each mutation event, K is simply the number of mutant individuals.

The total number of mutations in the experiment is defined as

$$K = K_1 + \ldots + K_m. \tag{2}$$

A mutation is said to be size i if there are i mutants in the sample sharing that mutation. Let c_l be the number of clusters of mutations of size l. Then it is easy to see that

$$K = c_1 + 2c_2 + \ldots + Lc_L, \tag{3}$$

where L is the maximum cluster size.

Consider the lth family again. Although Equation 1 indicates that K is counted by considering each mutation at a time, it can also be counted by considering one sequence at a time. Let m_j be the number of mutations that occurred in individual (sequence) j. Then

$$K_l = m_1 + m_2 + \ldots + m_{n_l}. \tag{4}$$

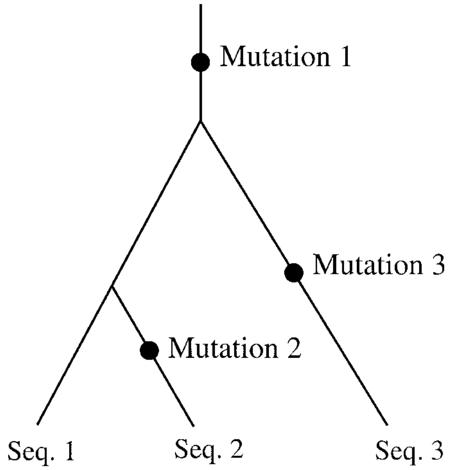

Figure 1.—An example of mutations and their relationships with h_i and m_j.

The equivalence of this counting to that of Equation 1 can be proved as follows. Let X_{ij} be the index variable that takes value 1 if the ith mutation is inherited by the jth sequence and value 0 otherwise. Figure 1 (see Table 1 also) gives an example, from which we can easily see the relationship between X_{ij} and h_i and m_j as

$$h_i = X_{i1} + \ldots + X_{in_l} \tag{5}$$

$$m_j = X_{1j} + \ldots + X_{Ij}. \tag{6}$$

It follows that

$$\begin{aligned} K_l &= h_1 + \ldots + h_I \\ &= \sum_{i=1}^{I} \sum_{j=1}^{n_l} x_{ij} \\ &= \sum_{j=1}^{n_l} \sum_{i=1}^{I} x_{ij} \\ &= m_1 + m_2 + \ldots + m_{n_l}. \end{aligned} \tag{7}$$

One special case deserves mentioning. When only one sequence is examined per family, each mutation event will be counted exactly once. Therefore, K in this situation is equal to the number of independent mutations in the experiment.

Representing K by Equation 4 provides a convenient basis for studying its statistical properties, and we discuss the mean and variance below.

The mean and variance of K: Let R represent the number of DNA replications between two successive generations and ν_i be the mutation rate for the ith cell replication. Let

$$\mu_i = \nu_1 + \ldots + \nu_i. \tag{8}$$

Then $\mu = \mu_R$ is the mutation rate per generation per

TABLE 1

An example of mutations and their relationships with h_i and m_j as shown in Figure 1

		Values of X_{ij}			
Sequences	Mutation:	1	2	3	m_j
1		1	0	0	1
2		1	1	0	2
3		1	0	1	2
h_i		3	1	1	$K_l = 5$

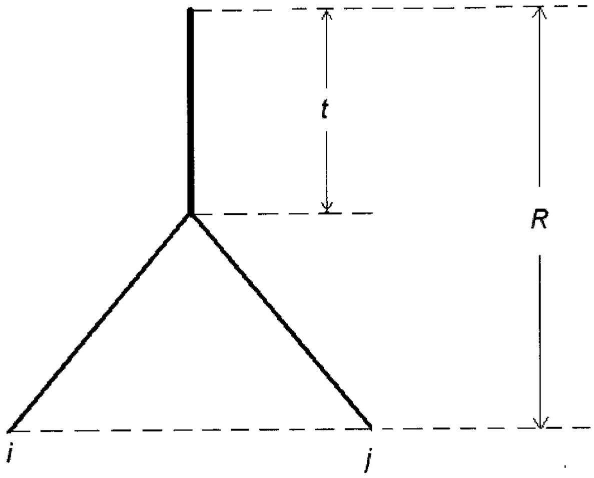

FIGURE 2.—Relationship between two sequences in a family.

sequence (Figure 2). We assume that R is a fixed number, which is appropriate for experiments in which individuals sampled are of the same sex and about the same age.

We assume that the number of mutations at replication i follows the Poisson distribution with both mean and variance of ν_i. Thus for any j sequence over R independent cell divisions,

$$E(m_j) = \mathrm{Var}(m_j) = \mu, \tag{9}$$

where $E(\ )$ and $\mathrm{Var}(\ )$ stand for expectation and variance, respectively. We also assume that mutations in different families are independent.

Since

$$\begin{aligned} E(K_l) &= \sum E(m_j) \\ &= n_l\mu, \end{aligned} \tag{10}$$

it follows that

$$\begin{aligned} E(K) &= \sum_l E(K_l) \\ &= \mu\sum_l n_l \\ &= n\mu. \end{aligned} \tag{11}$$

This equation suggests that an unbiased estimator of μ is

$$\hat{\mu} = K/n, \tag{12}$$

which is exactly the way the mutation rate has been estimated in some empirical studies (*e.g.*, HALDANE 1935; SPENCER and STERN 1948; AUERBACH 1962; MULLER *et al.* 1963; WOODRUFF and THOMPSON 1992; HUAI 1997; DRAKE *et al.* 1998; THOMPSON *et al.* 1998). Counting each cluster as one mutation or counting only independent mutations will result in underestimation of the true mutation rate; in some cases the underestimation may well be a few fold (PAASHUIS-LEW and HEDDLE 1998; SELBY 1998a,b; THOMPSON *et al.* 1998; HEDDLE 1999). Since the mutation rate μ in this article is defined as the sum of mutation rates over cell replications that are not necessarily equal, it does not matter if μ represents the rate of a single gene or multiple genes. Therefore, whether a single gene is considered or multiple genes are pooled, $\hat{\mu}$ is an unbiased estimator of μ.

We now consider the variance of $\hat{\mu}$. Consider first the case in which $n_l = 1$ for all i. As we mentioned earlier, K is equivalent to the number of mutation events. Since the mutations in different families are independent, we have

$$\mathrm{Var}(K) = \sum_l \mathrm{Var}(K_l) = n\mu. \tag{13}$$

Therefore

$$\mathrm{Var}(\hat{\mu}) = \mathrm{Var}(k)/n^2 = \mu/n. \tag{14}$$

In general, we have

$$\begin{aligned} \mathrm{Var}(K_i) &= \sum \mathrm{Var}(m_i) + \sum_{i<j} \mathrm{Cov}(m_i, m_j) \\ &= n\mu + \phi\sum_l n_l(n_l - 1), \end{aligned} \tag{15}$$

where

$$\phi = \mathrm{Cov}(m_i, m_j).$$

So ϕ is the covariance between m_i and m_j, that is, the covariance of the numbers of mutations in any pair of sequences from the same family. Therefore

$$\mathrm{Var}(K) = n\mu + \rho\mu\sum_l n_l(n_l - 1).$$

And then

$$\mathrm{Var}(\hat{\mu}) = \frac{\mu}{n} + \frac{\rho\mu\sum_l n_l(n_l - 1)}{n^2}, \tag{16}$$

where $\rho = \phi/\mu$ is the correlation coefficient between the numbers of mutations in two different sequences from the same family.

Two special cases are illuminating. The first is that an equal number of offspring in each family are examined, *i.e.*, $n_l = c$, $l = 1, \ldots, m$. It follows from Equation 16 that

$$K_l = K/(mc)$$

and

$$\mathrm{Var}(\hat{\mu}) = \frac{1}{m}\left(\frac{1}{c} + \rho\,\frac{c-1}{c}\right)\mu. \tag{17}$$

The second case is that all the offspring of each family are examined. Since offspring number of a family is typically not a predetermined quantity, we have

$$E(K) = \mu\sum E(n_l) \tag{18}$$

$$\mathrm{Var}(K) = \mu\sum E(n_l) + \phi\sum E[n_l(n_l - 1)]. \tag{19}$$

The most common practice is to assume that offspring number of a family follows a Poisson distribution with mean f. Then

$$E(K) = mE(K_l) = mf\mu \tag{20}$$

$$\mathrm{Var}(K) = m\mathrm{Var}(K_l) = mf\mu + m\phi f^2. \tag{21}$$

So we have

$$\mathrm{Var}(\hat{\mu}) = \frac{1}{m}\left(\frac{1}{f} + \rho\right)\mu. \tag{22}$$

Estimating the sampling variance: Since $\hat{\mu}$ is an unbiased estimator, the precision of estimation is thus determined by its variance. To compute the variance of $\hat{\mu}$, we need to know the value of the covariance between two mutations (see Figure 2).

Suppose that sequence i and sequence j shared a common ancestor t cell replications ago. Then we can express m_i and m_j as

$$m_i = r_{ij} + r_i \tag{23}$$

$$m_j = r_{ij} + r_j, \tag{24}$$

where r_{ij} represents the number of mutations in the common ancestor (shared mutations), and r_i and r_j are the numbers of mutations in sequences i and j, respectively, since the separation from their common ancestor (Figure 2). Conditional on the t value, r_{ij}, r_i, and r_j are independent Poisson variable with means equal to μ_t, $\mu - \mu_t$, and $\mu - \mu_t$, respectively. Therefore

$$\begin{aligned} E(m_i m_j) &= E_t[E(r_{ij}^2 + r_{ij}(r_i + r_j) + r_i r_j)] \\ &= E_t[\mu_t + \mu_t^2 + 2\mu_t(\mu - \mu_t) + (\mu - \mu_t)^2] \\ &= E_t[\mu_t + \mu^2] \\ &= \mu_{E(t)} + \mu^2. \end{aligned} \tag{25}$$

It follows that $\phi = E(m_i m_j) - E(m_i)E(m_j) = \mu_{E(t)}$. So ϕ is the expected number of shared mutations between two sequences from the same family. Since the correlation coefficient ρ is defined as

$$\rho = \frac{\mu_{E(t)}}{\mu}, \tag{26}$$

it is thus the proportion of mutations that are shared by two sequences from the same family.

Let $r_{l,ij}$ be the number of shared mutations for sequences i and j in the lth family. It thus follows that

$$E[r_{l,ij}] = \frac{\mu_{E(t)}}{\mu}. \tag{27}$$

The above equation suggests that an unbiased estimator of $\mu_{E(t)}$ is

$$\hat{\mu}_{E(t)} = \frac{2}{\sum_l n_l(n_l - 1)}\sum_l\sum_{i<j} r_{l,ij}, \tag{28}$$

where $n_l(n_l - 1)/2$ is the number of pairs of sequences for all the sample from the lth family. Suppose that there are in total L observed clusters and c_i is the size of cluster i. The above estimator can then be written as

$$\hat{\mu}_{E(t)} = \frac{\sum_{i=1}^{L} c_i(c_i - 1)}{\sum_l n_l(n_l - 1)}. \tag{29}$$

Substituting this into Equation 16 yields an unbiased estimate of the standard error of $\hat{\mu}$ as

$$\hat{\sigma} = \frac{1}{n}\sqrt{K + \sum_{i=1}^{L} c_i(c_i - 1)}\,. \tag{30}$$

Since many experiments on mutation rate were done over the many decades, detailed cluster sizes may not be available. So the above formula is difficult to apply in many situations. However, boundaries of the standard error of $\hat{\mu}$ can be obtained on the basis of partial information as follows.

If only the number S of singletons L and K is known, then the minimum value attainable by $\Sigma c_i(c_i - 1)$ is $(K - S)(K - S - L)/L$, corresponding to the situation in which all clusters are of equal size $(K - S)/L$. A lowerbound of the standard error is thus given by

$$\sigma_{\min} = \frac{1}{n}\sqrt{K + \frac{(K - S)(K - S - L)}{L}}\,. \tag{31}$$

Let $c_{\min}$ and $c_{\max}$ be the minimum and maximum cluster sizes, respectively. Then we have $2 \le c_{\min} \le c_{\max} \le K - S - c_{\min}(L - 1)$. The maximum value of $\Sigma c_i(c_i - 1)$ corresponds to the situation in which there are as many clusters of $c_{\min}$ size as possible. An upper bound of the standard error is therefore given by

$$\sigma_{\max} = \frac{1}{n}\sqrt{K + c_{\min}(L - b) + bc_{\max}(c_{\max} - 1)}\,, \tag{32}$$

where $b = (K - S - c_{\min}L)/(c_{\max} - c_{\min})$.

We can also construct a confidence interval for estimating $\hat{\mu}$. Note that $\hat{\mu}$ is the average of many variables with the same mean and variance and 0 or small covariance; therefore, by the central limit theorem of probability, $\hat{\mu}$ can be approximated by a normal distribution with mean μ and variance σ^2. The 95% confidence interval of $\hat{\mu}$ is estimated as $(\hat{\mu} - 1.96\hat{\sigma}, \hat{\mu} + 1.96\hat{\sigma})$.

TABLE 2

Examples of estimates of mutation rate and standard error

Species	K	S	L	n	$\hat{\mu}$ (10^{-4})	$\hat{\sigma}$ (10^{-4})
Drosophila melanogaster						
Sex-linked lethal[a]	3,616	2,852	69	1,955,989	18.487	0.695
Autosomal lethal[b]	194	118	18	10,166	190.83	25.69
Visible in males[c]	34	19	3	490,118	0.694	0.234
Visible in females[c]	17	4	2	340,306	0.500	0.283
Mice						
SLT in males[d]	558	69	8	1,487,177	3.752	1.885
SLT in females[e]	8	1	2	211,052	0.379	0.292

[a] Details in tables of Mason *et al.* (1985) and Woodruff and Thompson (1992).

[b] Details in Schalet (1960), Shukla *et al.* (1979), Woodruff *et al.* (1996), and also in Table 3 of Thompson *et al.* (1998).

[c] Schalet (1960) and Shukla *et al.* (1979); see also Table 3 of Thompson *et al.* (1998).

[d] Favor and Neuhauser-Klaus (1994), Russell and Russell (1996), Drost and Lee (1995, 1998), and Selby (1998a,b).

[e] Russell (1964), Russell and Russell (1992), and Selby (1998a,b).

Even though the mutation rate estimates of male and female mice have a not so small variance, we note that the mean ratio of the mutation rate estimates of mouse males over mouse females is 10 in Table 2, compared to the value of 2 in Chang *et al.* (1994). The sample size for female mice in the specific locus test (SLT) is far below that for male mice. Thus cluster mutations in female mice may well be underestimated or not recovered. More experiments are needed for a specific locus test of female mice.

DISCUSSION

Relation to others' work: Muller (1952, 1962) presented an estimator of the standard error of $\hat{\mu}$ as $\hat{\sigma} = q(1/n)\sqrt{\sum_{i=1}^{L} c_i(c_i - 1)}$ ($q = 1 - \hat{\mu}$, and singletons are considered as special cases of cluster mutations), which is close to but tends to underestimate $\hat{\sigma}$ compared with Equation 30 of this article. The derivation of this formula was never given, but Muller appeared to have obtained this formula with the assumption that the premeiotic mutations are not common; his brief explanation of the formulas suggested that it may be the result of a compound or generalized Poisson process, which is the sum of independent Poisson variables.

Engels (1979) also gave formulas for both mean and variance of mutation rates when cluster mutations are present. He focused mainly on a clustering model violating homogeneity of mutation probabilities among different parents. Engels visualized a conceptual two-step experiment. The first step consists of choosing from a pool of possible mutation rates for all different parents. The second step is the independent Bernoulli sampling within each family. So Engels did not address the more fundamental clustering model that violated assumption of Bernoulli (Poisson) distribution due to premeiotic mutation events. Even though his formulas appear to be more influential than Muller's, especially in Drosophila mutagenesis involving transposable elements (Margulies *et al.* 1986; Ehling and Neuhauser-Klaus 1988; Brown *et al.* 1989; Badge and Brookfield 1998), most of his conclusions and parameter estimations are not relevant to the premeiotic cluster mutations.

Sampling strategy: Since $\hat{\mu}$ is an unbiased estimator of μ, sampling strategy should aim at reducing the variance of the estimate. When to sample and how to sample both play important roles in determining the variance. Before we examine these two issues in turn, we note that the best possible strategy is to examine as many families as possible but only one sequence per family. This sampling strategy will minimize the sampling variance for a fixed total number of individuals examined (compare Equation 13 with Equation 16). Apparently this strategy is neither practical in most situations for multicellular organisms nor possible in somatic or bacteria mutagenesis (Luria and Delbrück 1943; Heddle 1999; H. Huai and Y.-X. Fu, unpublished results). There are good reasons for choosing large family sizes in certain situations. For example, it is important to know the value of ρ for the purpose of understanding the clustering phenomenon of mutations, and ρ can be estimated better with larger family sizes. Another reason favoring larger family size is that there may be a nearly fixed amount of effort or expense for each family screened regardless of its size.

It is clear from the variance equation (16) that the best sampling time is the one that minimizes the correlation coefficient ρ between any pair of progeny within a family. Apparently, this indicates that the best time to sample is right before somatic and germ-line differentia-

TABLE 3

The variation coefficient α ($\alpha = \hat{\sigma}/\hat{\mu}$) when all the m families are of the same size f and when $\rho = 0.10$

	Family size (f)						
$m \times u$	1	5	10	50	100	1,000	10,000
0.01	10.000	5.292	4.359	3.435	3.302	3.176	3.164
0.1	3.162	1.673	1.378	1.086	1.044	1.004	1.000
1	1.000	0.529	0.436	0.344	0.330	0.318	0.316
3	0.577	0.306	0.252	0.198	0.191	0.183	0.183
10	0.316	0.167	0.138	0.109	0.104	0.100	0.100

tion in multicellular species when germ cell number is small. Also note that males, especially older males in many species, usually have much more germ cell divisions than females have; hence it may be possible that the correlation coefficients ρ are much lower in males, especially in older males.

Given that the sampling time is determined, that is, ρ is fixed, how many individuals to sample from each family depends on the value of ρ. To demonstrate, consider the case where $n_l = c_{\max} = f$, $l = 1, \ldots, m$ or that family size follows Poisson distribution. We note from both Equations 17 and 22 that the variance of the estimate of mutation rate is approaching

$$\frac{1}{m}\,\hat{\mu}_{E(t)} \tag{33}$$

with an increasing and large enough family size. Once f (or $c_{\max}$) $\gg \rho^{-1}$, the further increase of mean family size does not help to reduce the variance much (see Table 3 and Figure 3). Therefore a reasonable strategy is to examine as many progeny per family as possible but not much more than ρ^{-1}.

Because of the important role of ρ in the estimate of mutation rate and genetic counseling (Hartl 1971; Wijsman 1991; Young 1991), it is useful to obtain an estimate of ρ from experimental data (Van Essen *et al.* 1992; Cooper and Krawczak 1993; Bridges 1994; Zlotogora 1998). From Equation 29, we can estimate $\hat{\mu}_{E(t)}$ by

$$\hat{\mu}_{E(t)} = \frac{\sum_{i=1}^{L} c_i(c_i - 1)}{(m-1)\sigma_f^2 + n(f-1)}, \tag{34}$$

where f is the mean family size and σ_f^2 is the sampling variance of the family size, *i.e.*,

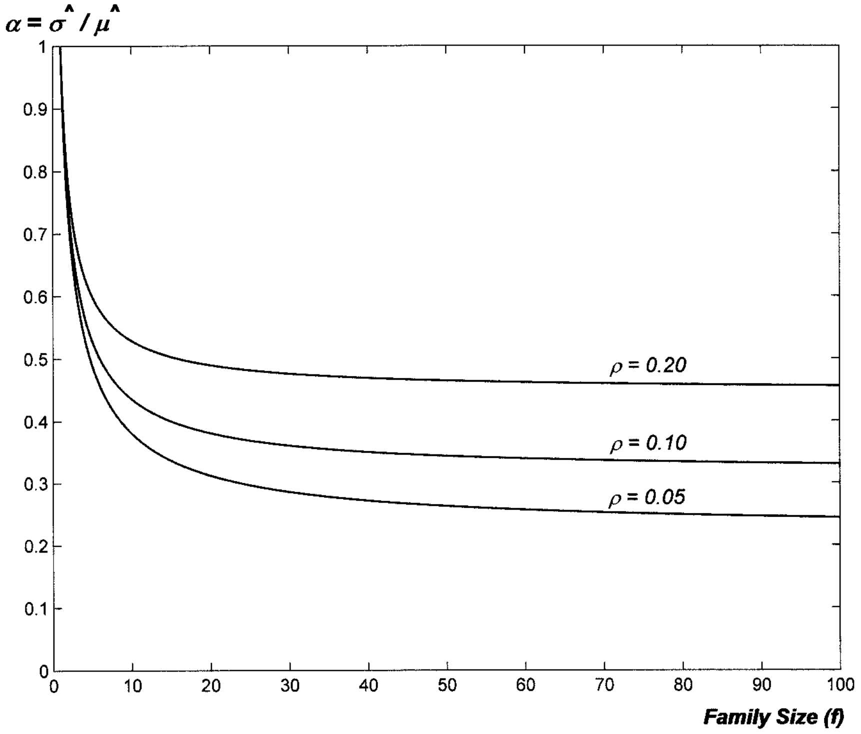

Figure 3.—The total sampled family number is fixed at $m = 1.00/\mu$, and all the m families are of the same size f. The controlled family sizes vary from 1.0 to 100 to show their effects on the variation coefficient of mutation rate estimation.

$$\sigma_f^2 = \frac{1}{m-1}\sum_i (n_i - f)^2 \qquad (35)$$

$$\approx \frac{1}{m}\sum_i n_i^2 - f^2. \qquad (36)$$

So ρ can be estimated by

$$\rho = \frac{\mu_{E(t)}}{\mu}$$

$$= \frac{n\sum_{i=1}^{L} c_i(c_i - 1)}{K[(m-1)\sigma_f^2 + n(f-1)]} \qquad (37)$$

$$\approx \frac{\sum_{i=1}^{L} c_i(c_i - 1)}{K(\sigma_f^2/f + f)}. \qquad (38)$$

K is the total number of mutants recovered in the experiment, counting all members of any clusters.

Sample size requirement: Besides sampling strategy to reduce variance in the estimate, it is important to determine the sample size required in achieving a given precision in estimation. It is obvious that the standard error of an estimate should not be larger than the estimate itself. For a good estimate, the standard error should probably be an order of magnitude smaller than the estimate.

For simplicity, consider the case that an equal number of progeny is examined for each family. Suppose we want to ensure that the standard error is as small as $\alpha\mu$. Then family number m and family size $f(c_{\max})$ need to satisfy

$$\frac{1}{m}\left(\frac{1}{f} + \rho\frac{f-1}{f}\right)\mu = \alpha^2\mu^2 \qquad (39)$$

or

$$m = \alpha^{-2}\mu^{-1}\left(\frac{1}{f} + \rho\frac{f-1}{f}\right). \qquad (40)$$

We can see from Table 3 and Figure 3 that it is obvious that once the uniformly sampled family size f is above the reciprocal of ρ, further increases of f will not help reduce the standard error of mutation rate estimation very much. In these cases there is an optimum family size (a little bit $>\rho^{-1}$) at which the experiment is most efficient. So it is not always true that the larger the sample size, the more precise the experiment.

On the other hand, given that the total sample size is fixed (shown in Figure 4), the best sampling strategy is to examine as many families as possible where the family size f is fixed at one, where each sample is independent from others. If this is not feasible, try to fix the family size as low as possible when there is no need for estimation of ρ.

Also family size f can affect the number of families that need to be studied, but the most important factor is the mutation rate itself. In general, we conclude that a useful experiment for estimating mutation rate should examine at least as many families as the reciprocal of the mutation rate (Table 3, Figures 3 and 4). This conclusion will reintroduce the problem of how to avoid preexisting mutations if the sampled family number is at least as large as the reciprocal of the mutation rate.

Can we find a way that *de novo* cluster mutations can be discriminated confidently from preexisting mutations? It is relatively easy to eliminate the preexisting mutations that are from grandparents' heterozygosity (Yang *et al.* 2001), but if one of the grandparents is genetic mosaic for a new mutation, then we need careful analysis of its timing and effects (H. Huai and Y.-X. Fu, unpublished results).

Estimating μ from the number of mutation events: Let I be the number of mutation events in an experiment that examined s offspring. As we have shown, when one sequence is examined per family, I is the same as K so using I/s yields an unbiased estimate of μ. When multiple offspring from a family are examined, I can be $<K$; the magnitude of difference depends on the number of offspring examined as well as on $\hat{\mu}_{E(t)}$.

Since K/n is an unbiased estimate of μ, it follows that I/n is an underestimate of mutation rate. It is tempting to suggest an estimate as

$$K_l = I/n', \qquad (41)$$

where n' is a constant satisfying $m \leq n' < n$. However, it is not obvious what value n' should be. Hence counting each mutant as one mutation regardless of the cluster's origin is the more straightforward way to obtain an unbiased estimate of the mutation rate. Nevertheless, it should be worth exploring efficient ways to use the frequencies of various mutation events.

Alternative ways to combine information: We note that K_l/n_l is an unbiased estimate of μ. So

$$\hat{\mu}' = \sum_i \beta_i K_l/n_l \qquad (42)$$

is an unbiased estimator, where $\beta_i \geq 0$ and $\Sigma_i\beta_i = 1$. Because

$$\mathrm{Var}(\hat{\mu}') = \sum_i (\beta_i^2/n_l)(\mu + \mu_{E(t)}(n_l - 1)), \qquad (43)$$

to obtain a best linear estimate of μ, β needs to satisfy

$$\frac{2\beta_i}{n_i(\mu + \mu_{E(t)}(n_i - 1))} = \frac{2\beta_m}{n_m(\mu + \mu_{E(t)}(n_m - 1))}, \qquad (44)$$

where $i = 1, \ldots, m-1$. That is,

$$\frac{\beta_i}{\beta_m} = \frac{n_i(\mu + \mu_{E(t)}(n_i - 1))}{n_m(\mu + \mu_{E(t)}(n_m - 1))}. \qquad (45)$$

So

$$\beta_i = \left(\sum\nolimits_i \frac{n_i}{1 + \rho(n_i - 1)}\right)^{-1} \frac{n_i}{1 + \rho(n_i - 1)}. \qquad (46)$$

It follows that when ρ is close to zero the best way to

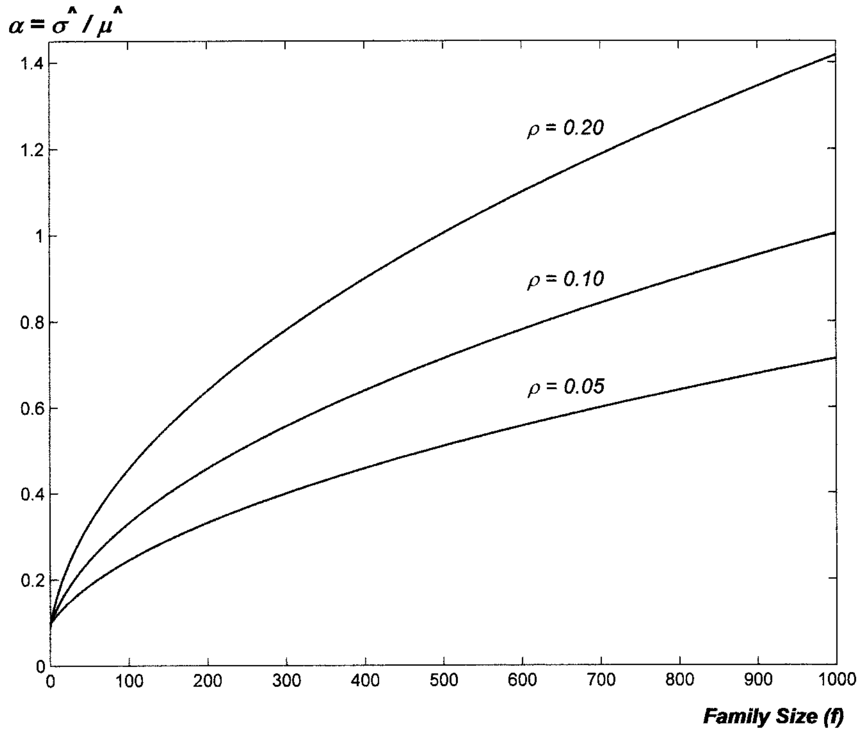

Figure 4.—The total sample size fixed at $m \times f = 100/\mu$, and all the m families are of the same size f. The designed family sizes vary from 1.0 to 1000 to visualize their effects on the variation coefficient of mutation rate estimation.

combine data from different families is to give each individual equal weight regardless of family affiliation, because each provides the same amount of independent information; when ρ is close to one, the best way is to give equal weight to each family, because each family provides the same amount of information regardless of size.

This work was supported in part by National Institutes of Health grants R29 GM50428 (Y.X.F.) and R01 HGO1708 (Y.X.F.).

LITERATURE CITED

Abrahamson, S., and S. Wolff, 1976 Re-analysis of radiation-induced specific locus mutations in mouse. Nature **264:** 715–719.

Arrault, X., V. Michel, P. Quillardet, M. Hofnung and E. Touati, 2002 Comparison of kinetics of induction of DNA adducts and gene mutations by a nitrofuran compound, 7-methoxy-2-nitronaphtho[2,1-b]furan (R7000), in the caecum and small intestine of Big BlueTM mice. Mutagenesis **17:** 353–359.

Auerbach, C., 1959 Spontaneous mutations in dry spores of Neurospora crassa. Z. Vererbungsl. **90:** 335–346.

Auerbach, C., 1962 *Mutation: An Introduction to Research in Mutagenesis. Part I. Methods.* Oliver & Boyd, Edinburgh.

Badge, R. M., and J. F. Y. Brookfield, 1998 A novel repressor of P-element transposition in *Drosophila melanogaster.* Genet. Res. **71:** 21–30.

Bridges, C. B., 1919 The stages at which mutations occur in the germ tract. Proc. Soc. Exp. Biol. Mod. **17:** 1–2.

Bridges, P. J., 1994 *The Calculation of Genetic Risks.* Johns Hopkins University Press, Baltimore.

Brown, A. J. L., L. S. Alphey, A. J. Flavell, T. I. Gerasimova and S. J. Ross, 1989 Instability in the Ctmr2 strain of *Drosophila melanogaster*: role of P-element functions and structure of revertants. Mol. Gen. Genet. **218:** 208–213.

Castle, W. E., 1905 The mutation theory of organic evolution: from the standpoint of animal breeding. Science **21:** 521–525.

Castle, W. E., 1929 A mosaic(intense-dilute) coat pattern in the rabbit. J. Exp. Zool. **52:** 471–480.

Combes, R. D., J. Bootman, M. G. Ford, J. Hepworth and D. W. Salt, 1989 Statistical method for the design and analysis of mutation experiments with the fruit fly *Drosophila melanogaster*, pp. 251–283 in *Statistical Evaluation of Mutagenicity Test Data*, edited by D. J. Kirkland. Cambridge University Press, Cambridge, UK.

Cooper, D. N., and M. Krawczak, 1993 *Human Gene Mutation.* BIOS Scientific Publishers, Oxford.

Crow, J. F., 1993 How much do we know about spontaneous human mutation rates? Environ. Mol. Mutagen. **21:** 122–129.

Crow, J. F., 1999 Spontaneous mutation in man. Mutat. Res. **437:** 5–9.

Crow, J. F., and M. J. Simmons, 1983 The mutation load in Drosophila, pp. 1–35 in *The Genetics and Biology of Drosophila*, Vol. 3c, edited by M. Ashburner, H. L. Carson and J. N. Thompson. Academic Press, London.

De Vries, H., 1901/1903 *Die Mutations Theorie*, Vols. 1 and 2, Veit, Leipzig, Germany (English translation, 1909/1910. Open Court, Chicago).

Dobzhansky, T. H., and S. Wright, 1941 Genetics of natural populations. V. Relations between mutation, rate and accumulation of lethals in populations of *Drosophila pseudoobscura.* Genetics **26:** 23–51.

Drake, J. W., 1991 Spontaneous mutation. Annu. Rev. Genet. **25:** 125–146.

Drake, J. W., 1993 Rates of spontaneous mutation among RNA viruses. Proc. Natl. Acad. Sci. USA **90:** 4171–4175.

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

Drost, J. B., and W. R. Lee, 1995 Biological basis of germline mutation: comparisons of spontaneous germline mutation rates among Drosophila, mouse, and human. Environ. Mol. Mutagen. **25** (Suppl. 26): 48–64.

Drost, J. B., and W. R. Lee, 1998 The developmental basis for the germline mosaicism in mouse and *Drosophila melanogaster.* Genetica **102/103:** 421–443.

Ehling, U. H., and A. Neuhauser-Klaus, 1988 Induction of specific-locus and dominant-lethal mutations by cyclophosphamide and combined cyclophosphamide-radiation treatment in male-mice. Mutat. Res. **199:** 21–30.

Engels, W. R., 1979 The estimation of mutation rates when premeiotic events are involved. Environ. Mutagen. **1:** 37–43.

Favor, J., and A. Neuhauser-Klaus, 1994 Genetic mosaicism in the house mouse. Annu. Rev. Genet. **28:** 27–47.

Fisher, R. A., 1930 Note on a tricolour (mosaic) mouse. J. Genet. **23:** 77–81.

Fu, Y. X., 1994 A phylogenetic estimator of population size or mutation rate. Genetics **136:** 685–692.

Haldane, J. B. S., 1935 The rate of spontaneous mutation in a human gene. J. Genet. **31:** 317–326.

Hall, J. G., 1988 Somatic mosaicism: observations related to clinical genetics. Am. J. Hum. Genet. **46:** 1187–1193.

Heddle, J. A., 1999 On clonal expansion and its effects on mutant frequencies, mutation spectra and statistics for somatic mutations in vivo. Mutagenesis **14:** 257–260.

Heddle, J. A., L. Cosentino, G. Dawood, R. R. Swiger and Y. Paashuis-Lew, 1996 Why do stem cells exist? Environ. Mol. Mutagen. **28:** 334–341.

Hartl, D. L., 1971 Recurrence risks for germinal mosaics. Am. J. Hum. Genet. **23:** 124–134.

Hartl, D. L., and M. M. Green, 1970 Genetic studies of germinal mosaicism in *Drosophila melanogaster* using the mutable w^c gene. Genetics **65:** 449–456.

Huai, H., 1997 The evolutionary implication of premeiotic clusters of mutation. Ph.D. Thesis, Bowling Green State University, Bowling Green, OH.

Huai, H., and R. C. Woodruff, 1997 Clusters of identical new mutations can account for the "overdispersed" molecular clock. Genetics **147:** 339–348.

Huai, H., and R. C. Woodruff, 1998a Clusters of new identical mutants and the fate of underdominant mutations. Genetica **102/103:** 489–505.

Huai, H., and R. C. Woodruff, 1998b With the correct concept of mutation rate, cluster mutations can explain the overdispersed molecular clock. Genetics **149:** 467–469.

Johnson, T., 1999 Beneficial mutations, hitchhiking and the evolution of mutation rates in sexual populations. Genetics **151:** 1621–1631.

Keightley, P. D., and A. Eyre-Walker, 1999 Terumi Mukai and the riddle of deleterious mutation rates. Genetics **153:** 515–523.

Kimura, M., 1983 *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge, UK.

Kondrashov, A. S., and J. F. Crow, 1993 A molecular approach to estimating the human deleterious mutation rate. Hum. Mutat. **2:** 229–234.

Lewis, S. E., 1999 Life cycle of the mammalian germ cell: implication for spontaneous mutation frequencies. Teratology **59:** 205–209.

Luria, S. E., and M. Delbrück, 1943 Mutations of bacteria from virus sensitivity to virus resistance. Genetics **28:** 491–511.

Lynch, M., and W. G. Hill, 1986 Phenotypic evolution by neutral mutation. Evolution **40:** 915–935.

Margulies, L., D. I. Briscoe and S. S. Wallace, 1986 The relationship between radiation-induced and transposon-induced genetic-damage during Drosophila oogenesis. Mutat. Res. **162:** 55–68.

Mason, J. M., R. Valencia, R. C. Woodruff and S. Zimmering, 1985 Genetic drift and seasonal variation in spontaneous mutation frequencies in Drosophila. Environ. Mutagen. **7:** 663–676.

Mason, J. M., R. Valencia, R. C. Woodruff and S. Zimmering, 1987 A guide for performing germ cell mutagenesis assays using Drosophila melanogaster. Mutat. Res. **189:** 93–102.

McClintock, B., 1950 The origin and behavior of mutable loci in maize. Proc. Natl. Acad. Sci. USA **36:** 344–355.

Morgan, W. C., 1950 A new tail-short mutation in the mouse. J. Hered. **41:** 208–215.

Mohrenweiser, H., and B. Zingg, 1995 Mosaicism: the embryo as a target for induction of mutations leading to cancer and genetic disease. Environ. Mol. Mutagen. **25** (Suppl. 26): 21–29.

Muller, H. J., 1920 Further changes in the white-eye series of Drosophila and their bearing on the manner of occurrence of mutation. J. Exp. Zool. **31:** 443–472.

Muller, H. J., 1928 The measurement of gene mutation rate in Drosophila, its high variability, and its dependence on temperature. Genetics **13:** 279–357.

Muller, H. J., 1952 The standard error of the frequency of mutants some of which are of common origin. Genetics **37:** 608.

Muller, H. J., 1962 *Studies in Genetics*, p. 301. University of Indiana Press, Bloomington, IN.

Muller, H. J., I. Oster and S. Zimmering, 1963 Are chronic and acute gamma irradiation equally mutagenic in Drosophila?, pp. 275–311 in *Repair from Genetic Radiation Damage and Differential Radiosensitivity in Germ Cells*, edited by F. Sobels. Pergamon Press, Oxford.

Neel, J. V., 1998 A reappraisal of studies concerning the genetic effects of the radiation of humans, mice, and Drosophila. Environ. Mol. Mutagen. **31:** 4–10.

Neel, J. V., and E. D. Rothman, 1978 Indirect estimates of mutation rates in tribal Amerindians. Proc. Natl. Acad. Sci. USA **75:** 5585–5588.

Nishino, H., D. J. Schaid, V. L. Buettner, J. Haavik and S. Sommer, 1996 Mutation frequencies but not mutant frequencies in big blue mice fit a Poisson distribution. Environ. Mol. Mutagen. **28:** 414–417.

Paashuis-Lew, Y. R. M., and J. A. Heddle, 1998 Rates of mutation during embryogenesis and growth. Mutagenesis **13:** 613–617.

Purdom, C. C., K. F. Dyer and D. G. Papworth, 1968 Allelic clusters among spontaneous mutations in Drosophila. Mutat. Res. **5:** 305–307.

Russell, L. B., 1964 Genetic and functional mosaicism in the mouse, pp. 153–181 in *The Role of Chromosomes in Development*, edited by M. Locke. Academic Press, New York.

Russell, L. B., and W. L. Russell, 1992 Frequency and nature of specific-locus mutations induced in female mice by radiation and chemical: a review. Mutat. Res. **296:** 107–127.

Russell, L. B., and W. L. Russell, 1996 Spontaneous mutations recovered as mosaics in the mouse specific-locus test. Proc. Natl. Acad. Sci. USA **93:** 13072–13077.

Russell, W. L., 1977 Mutation frequencies in female mice and the estimation of genetic hazards of radiation in women. Proc. Natl. Acad. Sci. USA **74:** 3523–3527.

Schalet, A., 1960 A study of spontaneous visible mutations in *Drosophila melanogaster*. Ph.D. Thesis, Indiana University, Bloomington, IN.

Selby, P. B., 1998a Major impacts of gonadal mosaicism on hereditary risk estimation, origin of hereditary diseases, and evolution. Genetica **102/103:** 445–462.

Selby, P. B., 1998b Discovery of numerous clusters of spontaneous mutations in the specific-locus test in mice necessitate major increases in estimates of doubling doses. Genetica **102/103:** 463–487.

Shukla, P. T., K. Sankaranarayanan and F. H. Sobels, 1979 Is there a proportionality between the spontaneous and the x-ray-induced rates of mutation? Mutat. Res. **61:** 229–248.

Spencer, W. P., and C. Stern, 1948 Experiments to test the validity of the linear r-dose/ mutation frequency relation in Drosophila at low dosage. Genetics **33:** 43–74.

Stuart, G. R., and B. W. Glickman, 2000 Through a glass, darkly: reflections of mutation from *lacI* transgenic mice. Genetics **155:** 1359–1367.

Thompson, J. N., R. C. Woodruff and H. Huai, 1998 Mutation rate: a simple concept has become complex. Environ. Mol. Mutagen. **32:** 292–300.

Van Essen, A. J., S. Abbs, M. Baiget, E. Bakker, C. Boileau *et al.*, 1992 Parental origin and germline mosaicism of deletions and duplications of the dystrophin gene: a European study. Hum. Genet. **88:** 249–257.

Wijsman, E. M., 1991 Recurrence risk of a new dominant mutation in children of unaffected parents. Am. J. Hum. Genet. **48:** 654–661.

Woodruff, R.C., and J. N. Thompson, 1992 Have premeiotic clusters of mutation been overlooked in evolutionary theory? J. Evol. Biol. **5:** 457–464.

Woodruff, R. C., H. Huai and J. N. Thompson, 1996 Clusters of new mutation in the evolutionary landscape. Genetica **98:** 149–160.

Wright, S., and O. N. Eaton, 1926 Mutational mosaic coat patterns of the guinea pig. Genetics **11:** 333–351.

Yang, H. P., A. Y. Tanikawa, W. A. Van Voorhies, J. C. Silva and A. S. Kondrashov, 2001 Whole-genome effects of ethyl methanesulfonate-induced mutation on nine quantitative traits in outbred *Drosophila melanogaster.* Genetics **157:** 1257–1265.

Young, I. D., 1991 *Introduction to Risk Calculations in Genetic Consulting.* Oxford University Press, Oxford.

Zlotogora, J., 1998 Germ line mosaicism. Hum. Genet. **102:** 381–386.

Communicating editor: M. K. Uyenoyama