Statistical Methods for Analyzing *Drosophila* Germline Mutation Rates

Yun-Xin Fu*,†,1

*Laboratory for Conservation and Utilization of Bio-Resources, Yunnan University, Kunming, 650091, China, and †Division of Biostatistics and Human Genetics Center, School of Public Health, University of Texas Health Science Center, Houston, Texas 77030

ABSTRACT Most studies of mutation rates implicitly assume that they remain constant throughout development of the germline. However, researchers recently used a novel statistical framework to reveal that mutation rates differ dramatically during sperm development in *Drosophila melanogaster*. Here a general framework is described for the inference of germline mutation patterns, generated from either mutation screening experiments or DNA sequence polymorphism data, that enables analysis of more than two mutations per family. The inference is made more rigorous and flexible by providing a better approximation of the probabilities of patterns of mutations and an improved coalescent algorithm within a single host with realistic assumptions. The properties of the inference framework, both the estimation and the hypothesis testing, were investigated by simulation. The refined inference framework is shown to provide (1) nearly unbiased maximum-likelihood estimates of mutation rates and (2) robust hypothesis testing using the standard asymptotic distribution of the likelihood-ratio tests. It is readily applicable to data sets in which multiple mutations in the same family are common.

SPERM and eggs experience many divisions after the fertilized egg, and mutations may occur each time a cell divides. Little is known about the patterns of mutations during development of the germline cell lineage. This is partly due to the scarcity of appropriate experimental data and to lack of proper statistical methods for analyzing such data. Recently, Gao *et al.* (2011) reported that mutation rates differ dramatically during germline development in *Drosophila*, with the rate for the first cell division the highest. But the method developed by Gao *et al.* (2011) is limited to handling only families with at most two mutations each. Also, their conclusions relied heavily on hypothesis testing and the statistical properties of the likelihood-ratio test they used are not known under such circumstances. Furthermore, their coalescent algorithm is too simplistic, not taking into consideration the details of spermatogenesis. Since the ability to make inferences about mutation rates at the level of single-cell division would be a significant step forward, it is desirable to make the inference rigorous and applicable for analysis of data in which more than two mutations per family are common.

Knowledge of development of the germline lineage is essential for inferring mutation rates. For *Drosophila melanogaster* males, each sperm from a young adult has experienced ≥36 divisions. The first 14 divisions occur in the cleavage stage characterized by fast cell divisions; the last 5 occur during spermatogenesis; those in between occur during gastrulation and organogenesis when the germline stem cells (GSC) divide asymmetrically. For *Drosophila*, it is well known (Drost and Lee 1995, 1998; Gilbert 2003) that (1) after the 8th cell division, ~4–6 cells become the primordial germ cells (PGC); (2) after the 12th division, the number of PGCs ranges from 23 to 52; (3) after the 14th division there are 5–6 PGCs in each gonad; and (4) from the 15th division to just before spermatogenesis the number of PGCs remains more or less constant. After the 31st division one of the two daughter cells of each stem cell remains as a stem cell; the other one differentiates into 64 sperm. Thus, if sperm are sampled after the 36th division, all have experienced exactly 36 divisions, but if they are sampled after, for example, the 38th division, some would have experienced 36 divisions, some 37, and some 38 divisions. The algorithm developed by Gao *et al.* (2011), using the principle of coalescence (Kingman 1982; Ewens 2004), does not account for these differences.

doi: 10.1534/genetics.113.151571
Manuscript received March 22, 2013; accepted for publication May 5, 2013
[1]Address for correspondence: Division of Biostatistics, School of Public Health, Box 20186, 1200 Herman Pressler, Houston, TX 77030. E-mail: yunxin.fu@uth.tmc.edu

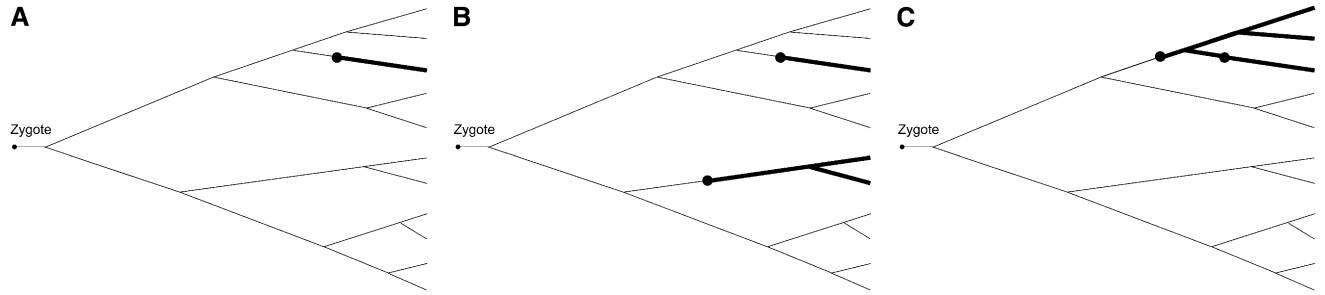

Figure 1 Examples of mutations and resulting mutation patterns. (A) Single mutation leading to mutation pattern ⟨1⟩; (B) two mutations leading to mutation pattern ⟨2, 1⟩; (C) two mutations leading to mutation pattern ⟨3⟩ due to the second mutation being masked by the first one.

To develop a thorough understanding of mutational patterns during germline development requires obtaining an estimate of mutation rate for each cell division and testing various hypotheses about mutation rates. This article describes further development of the inference framework, which overcomes previous shortcomings, investigates its statistical properties, improves the coalescent algorithm, and reanalyzes the published data. The improved inference framework has the advantage of being adaptable to analyzing mutation patterns in nucleotide polymorphism data, such as generated by next-generation DNA sequencing.

The Theory

Definitions and notations

Consider a sample of families, each consisting of a number of offspring (or sperm) from the same father. For each family with n sampled offspring, a *mutation pattern* is observed and represented by $\langle i_1, i_2, \ldots, i_k \rangle$ such that each element represents an identified mutation, its value equal to the number of mutants for the mutation, where k is the number of mutations. For example, $\langle 1 \rangle$ represents a mutation in a family that leads to only one mutant; $\langle 3, 2 \rangle$ represents two mutations, where one leads to three mutants and another to two mutants. Also we use $\langle\rangle$ to represent the case in which no mutation is observed.

Sequencing the same region of the genome among all sampled offspring in the same family will yield a mutation pattern. Alternatively, such information can be obtained from traditional experiments, particularly for some model organisms. For *Drosophila*, multigeneration mutation screening has been well developed (Muller 1928; Woodruff *et al.* 1984, 1996; Ashburner 1989; Greenspan 1997) and one such system was used by Gao *et al.* (2011) for the purpose of identifying recessive lethal or nearly lethal mutations. More experimental detail can be found in Gao *et al.* (2011) and for the main purpose of this article, it should suffice to outline the structure of information and the type of mutation being examined. The *recessive lethality d* of a recessive mutation is defined as one minus the maximal percentage of the homozygote (among all survival offspring) for that mutation. The data reported by Gao *et al.* (2011) correspond to those mutations with recessive lethality equal to 99%, that is, no more than 1% of survival offspring are z/z homozygote. Once a cell acquires a mutation with recessive lethality d, further mutation(s) is much more likely to increase lethality than to reverse it. Consequently recessive lethal mutations have a masking effect such that only the earliest one is identifiable. Figure 1 shows examples of how mutations in a genealogy lead to different mutation patterns.

Suppose branches of a sample genealogy are labeled by integers. For branch i, define $\Omega(i)$ as the set of branches consisting of the branch i and all its descendant branches, which are referred to as the subtree of the branch i. For the genealogy shown in Figure 2, for example, $\Omega(2) = \{2, 5, 6\}$ and $\Omega(3) = \{3, 4, 7, 9, 8, 9\}$. Therefore, two mutations, respectively on branches i and j, are both observable if and only if $\Omega(i) \cap \Omega(i) = \emptyset$.

Suppose the germ cell divisions from a fertilized egg to sperm are divided into I intervals. Let $[i, j]$ represent the interval from the ith to the jth cell divisions. Suppose the mutation rate per cell division for the lth interval is u_l and define $\mathbf{u} = (u_1, \ldots, u_I)^T$. For a given sample of sperm, there is a genealogy connecting them to the fertilized egg. Suppose each branch in the genealogy is identified by an unique integer (how branches are numbered is immaterial). Define for the ith branch, b_{ij} as the number of divisions it contains from the jth interval and $\mathbf{b}_i$ as a vector with elements $b_{ij}, j = 1, \ldots, I$. That is, $\mathbf{b}_i = (b_{i1}, \ldots, b_{iI})^T$. Define $\phi(i)$ as the *size* of the ith branch, *i.e.*, the number of descendants of the branch that are observed in the sample, and

$$\mathbf{a}_k = \sum_{i:\phi(i)=k} \mathbf{b}_i \quad \text{and} \quad \mathbf{t} = \sum_{k=1}^{n} \mathbf{a}_k. \tag{1}$$

Then a_{kj} is the total number of cell divisions from the jth interval that are of size k and the kth element, t_k, of $\mathbf{t}$ is the total number of cell divisions from the kth interval. For branch i, let $\mathbf{w}_i$ be the sum of lengths of all the branches in $\Omega(i)$, excluding the branch i itself. That is, $\mathbf{w}_i = -\mathbf{b}_i + \sum_{k:\Omega(i)} \mathbf{b}_k$. Figure 2 illustrates the aforementioned quantities in a genealogy of five alleles taken after the fifth division. It follows that $\mathbf{a}_1 = \sum_{k=5}^{9} \mathbf{b}_i = (0, 1, 5)^T$, $\mathbf{a}_2 = \mathbf{b}_2 + \mathbf{b}_4 = (1, 3, 0)^T$, $\mathbf{a}_3 = \mathbf{b}_3 = (1, 1, 0)^T$, $\mathbf{a}_4 = 0$, $\mathbf{a}_5 = \mathbf{b}_1 = (1, 0, 0)$,

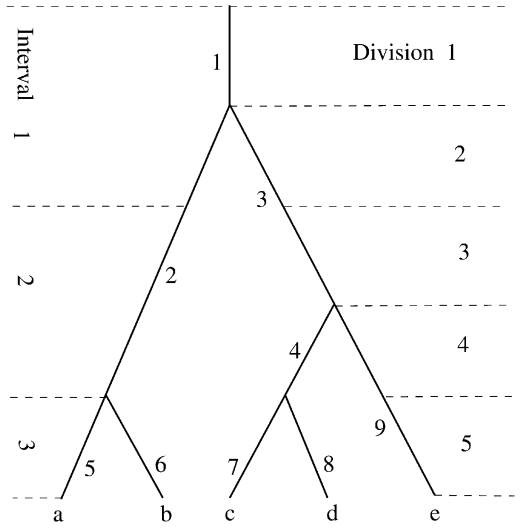

Figure 2 An example of the genealogy of five alleles ($a-e$) sampled from the cell population after the fifth division. Each branch is identified by a nearby integer. The sizes of the branches are $\varphi(1) = 5$, $\phi(2) = 2$, $\phi(3) = 3$, $\phi(4) = 2$, $\phi(5) = \phi(6) = \phi(7) = \phi(8) = \phi(9) = 1$; $\mathbf{b}_1^T = (1,0,0), \mathbf{b}_2^T = (1,2,0)$, $\mathbf{b}_3^T = (1,1,0), \mathbf{b}_4^T = (0,1,0)$, $\mathbf{b}_5^T = \mathbf{b}_6^T = \mathbf{b}_7^T = \mathbf{b}_8^T = (0,0,1)$, and $\mathbf{b}_9^T = (0,1,1)$.

and an example of $\mathbf{w}$ is that $\mathbf{w}_4 = \mathbf{b}_7 + \mathbf{b}_8$ and $\mathbf{w}_3 = \mathbf{b}_4 + \mathbf{b}_7 + \mathbf{b}_8 + \mathbf{b}_9$.

In addition to $\mathbf{a}$ and $\mathbf{t}$, we will encounter other quantities that are functions of $\mathbf{b}_i$, $i = 1, \ldots$, which will be defined as they are introduced. Each of these quantities has a value for a given genealogy, and often we need to evaluate its expectation (mean) over all genealogies. We use a bar over the variable to represent its expectation. For example,

$$\bar{\mathbf{t}} = \int_g \mathbf{t}\, dg, \quad \bar{\mathbf{a}}_k = \int_g \mathbf{a}_k\, dg. \tag{2}$$

Probability of a mutation pattern

Assume that the number of mutations in a branch of the sample genealogy g is a Poisson variable. Then the probability of no mutation in a family is equal to $e^{-\mathbf{t}^T\mathbf{u}}$. Since a single mutation leading to an observed pattern $\langle i \rangle$ must occur on a branch of size i, it follows that

$$\begin{aligned} Pr(\langle i \rangle | g) &= \sum_{k:\phi(k)=i} e^{-(\mathbf{t}-\mathbf{b}_k-\mathbf{w}_k)^T\mathbf{u}}\left(1 - e^{-\mathbf{b}_k^T\mathbf{u}}\right) \\ &= \sum_{k:\phi(k)=i} e^{-(\mathbf{t}-\mathbf{w}_k)^T\mathbf{u}}\left(e^{\mathbf{b}_k^T\mathbf{u}} - 1\right), \end{aligned} \tag{3}$$

where the summation is taken over all the branches of size i. In the summation, the first term $e^{-(\mathbf{t}-\mathbf{b}_k-\mathbf{w}_k)^T\mathbf{u}}$ is the probability that there is no mutation outside the subtree of branch k and the second term $(1 - e^{-\mathbf{b}_k^T\mathbf{u}})$ is the probability there is at least one mutation in branch k. This is because any mutation in the subtree will be masked by the mutation in branch k and thus not observable. In general, we have for a mutation pattern $\langle i_1, \ldots, i_l \rangle$ that

$$Pr(\langle i_1, \ldots, i_l \rangle | g) = \sum_{(k_1,\ldots,k_l) \in \mathcal{J}_g(i_1,\ldots,i_l)} \left[e^{-\left(\mathbf{t}-\mathbf{w}_{k_1,\ldots,k_l}\right)^T\mathbf{u}} \prod_{i=1}^{l} \left(e^{\mathbf{b}_{k_i}^T\mathbf{u}} - 1\right) \right], \tag{4}$$

where $\mathbf{w}_{k_1,\ldots,k_l} = \sum_i w_{k_i}$ and $\mathcal{J}_g(i_1, \ldots, i_l)$ is the collection of the branch sets of genealogy g on which mutations can lead to the observed mutational pattern. That is,

$$\begin{aligned} \mathcal{J}_g(i_1, \ldots, i_l) = \{ &(k_1, \ldots, k_l) : \phi(k_j) = i_j \text{ for } i = 1, \ldots, l \\ &\text{and } \Omega(k_i) \cap \Omega(k_j) = \emptyset \text{ for } i \neq j \}. \end{aligned} \tag{5}$$

Since sample genealogy is generally unobservable, one needs to consider all the possible sample genealogies from which the given mutational pattern can be generated, which leads to the general unconditional probability of the mutational pattern $\langle i_1, \ldots, i_l \rangle$ as

$$Pr(\langle i_1, \ldots, i_l \rangle) = \int_g \sum_{(k_1,\ldots,k_l) \in \mathcal{J}_g(i_1,\ldots,i_l)} \left[e^{-\left(\mathbf{t}-\mathbf{w}_{k_1,\ldots,k_l}\right)^T\mathbf{u}} \prod_{i=1}^{l} \left(e^{\mathbf{b}_{k_i}^T\mathbf{u}} - 1\right) \right] dg. \tag{6}$$

This formula provides the basis for the proposed inferences and detailed analysis of *Drosophila* data. For any given mutation pattern $\langle i_1, \ldots, i_l \rangle$ and $\mathbf{u}$, the probability can be evaluated as the average of $Pr(\langle i_1, \ldots, i_l \rangle | g)$ [which is given by (4)] over a reasonably large set of simulated sample genealogies. However, it is generally not efficient and often impractical to use the above formula directly if hundreds or even thousands of different $\mathbf{u}$ need to be evaluated.

Approximation to the Probability of a Mutation Pattern

Since Equation 6 is computationally expansive to use in general, accurate and yet-fast approximations to the probabilities of various mutation patterns are important and often necessary for large-scale data analysis. Gao *et al.* (2011) found approximations to the probabilities for up to two mutations in a family, using the Taylor expansion. For example, $e^{-\mathbf{t}^T\mathbf{u}} \approx 1 - \mathbf{t}^T\mathbf{u} + \frac{1}{2}\mathbf{u}^T(\mathbf{t}\mathbf{t}^T)\mathbf{u}$. Define $\mathbf{A}_{ij} = \mathbf{a}_i\mathbf{a}_j^T$, $\mathbf{A}_i = \mathbf{a}_i\mathbf{t}^T$, and $\mathbf{A}_0 = \mathbf{t}\mathbf{t}^T$. Then the probabilities $p_0 = Pr(\langle\rangle)$, $p_i = Pr(\langle i \rangle)$, and $p_{ij} = Pr(\langle i, j \rangle)$ can be approximated (Gao *et al.* 2011) by

$$p_0 \approx 1 - \bar{\mathbf{t}}^T\mathbf{u} + \frac{1}{2}\mathbf{u}^T\bar{\mathbf{A}}_0\mathbf{u}, \tag{7}$$

$$p_i \approx \bar{\mathbf{a}}_i^T\mathbf{u} - \mathbf{u}^T\bar{\mathbf{A}}_i\mathbf{u}, \tag{8}$$

$$p_{ij} \approx \frac{2 - \delta_{i-j}}{2}\mathbf{u}^T\bar{\mathbf{A}}_{ij}\mathbf{u}, \tag{9}$$

where $\delta_{i-j} = 1$ if $i = j$ and 0 otherwise. This method of approximation is referred to as the approximation by Taylor expansion (ATE). Although these approximations cover up

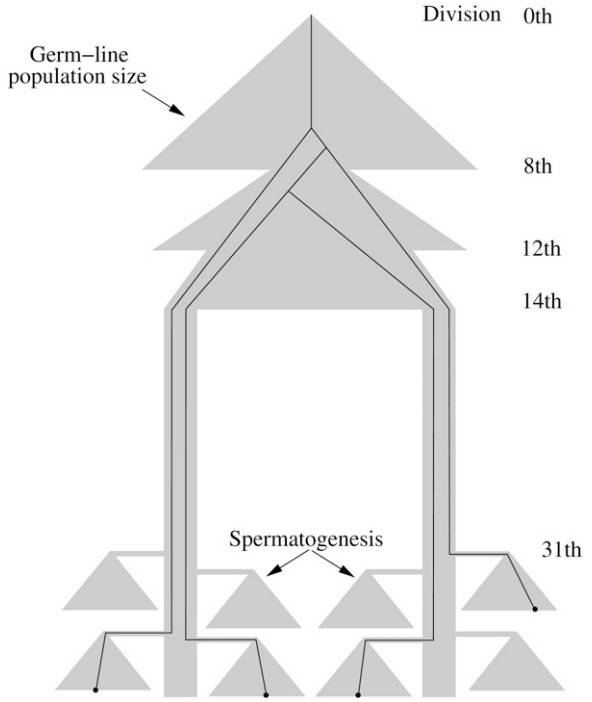

Figure 3 Population dynamics and an example of the genealogy of four sperm.

to two mutations per family, in principle a reasonably accurate approximation to the probability of any given mutation pattern can be obtained if a sufficient number of Taylor expansion terms are included. With increasing mutation rate, the number of required terms for each case will also increase, and due to the need to estimate a large number of coefficients in higher-order terms, their computations make the ATE inefficient.

Since typically $\mathbf{b}_k^T\mathbf{u} \ll 1$ in Equation 6, $e^{\mathbf{b}_{k_i}^T\mathbf{u}} - 1 \approx \mathbf{b}_k^T\mathbf{u}$. Furthermore one can simplify the expression by replacing w for each combination of branches by its average value and arrive at

$$Pr(\langle i_1,\ldots,i_l\rangle|g) \approx \sum_{(k_1,\ldots,k_l)\ \in\ \mathcal{J}_g(i_1,\ldots,i_l)} \left[e^{-\left[\mathbf{t}-\mathbf{w}_{k_1,\ldots,k_l}\right]^T\mathbf{u}} \prod_{i=1}^{l} \mathbf{b}_{k_i}^T\mathbf{u} \right] \tag{10}$$

$$\approx e^{-[\mathbf{t}-\mathbf{w}(\langle i_1,\ldots,i_l\rangle)]^T\mathbf{u}} S(\langle i_1,\ldots,i_l\rangle,\mathbf{u}), \tag{11}$$

where $w(\langle i_1,\ldots,i_l\rangle)$ is defined as the average value of $w_{k_1,\ldots,k_l}$ over the set $\mathcal{J}_g(i_1,\ldots,i_l)$ and

$$\begin{aligned} S(\langle i_1,\ldots,i_l\rangle,\mathbf{u}) &= \sum_{(k_1,\ldots,k_l)\in\mathcal{J}_g(i_1,\ldots,i_l)} \left(\prod_{i=1}^{l} \mathbf{b}_{k_i}^T\mathbf{u} \right) \\ &= \sum_{l_1,\ldots,l_m} a_{l_1\ldots l_m} u_{l_1}\ldots u_{l_m}, \end{aligned} \tag{12}$$

where $a_{l_1\ldots l_m} = \sum_{(k_1,\ldots,k_l)\ \in\ \mathcal{J}_g(i_1,\ldots,i_l)} b_{k_1 l_1}\ldots b_{k_m l_m}$, which leads to an approximation of Equation 6 as

Table 1 Constraints during the germline development of a male *Drosophila melanogaster*

Constraint no.	Detail
1	$N(8) \in [4, 6]$
2	$N(12) \in [23, 52]$
3	Population split into two with each $N \in [5, 9]$
4	Stem-cell stage starts from the 15th division onward with $p_2 = 0.001$
5	Differentiated cells after the 31st division starts spermatogenesis

$$Pr(\langle i_1,\ldots,i_l\rangle) \approx \int_g e^{-[\mathbf{t}-\mathbf{w}(\langle i_1,\ldots,i_l\rangle)]^T\mathbf{u}} S(\langle i_1,\ldots,i_l\rangle,\mathbf{u})dg. \tag{13}$$

A further simplification and approximation can be obtained by moving the integration inward and replacing each quantity by its integral (that is, its expectation). This leads to the approximation

$$Pr(\langle i_1,\ldots,i_l\rangle) \approx e^{-[\bar{\mathbf{t}}-\bar{\mathbf{w}}(\langle i_1,\ldots,i_l\rangle)]^T\mathbf{u}} \bar{S}(\langle i_1,\ldots,i_l\rangle,\mathbf{u}), \tag{14}$$

where $\bar{\mathbf{w}}(\langle i_1,\ldots,i_l\rangle)$ is the mean value of $\mathbf{w}(\langle i_1,\ldots,i_l\rangle)$ over all genealogies, and $\bar{S}$ is the mean of S over all genealogies, which can be computed as

$$\bar{S}(\langle i_1,\ldots,i_m\rangle,\mathbf{u}) = \sum_{l_1,\ldots,l_m} \bar{a}_{l_1\ldots l_m} u_{l_1}\ldots u_{l_m}, \tag{15}$$

where $\bar{a}_{l_1\ldots l_m}$ is the mean of $a_{l_1\ldots l_m}$ over all possible genealogies. This new approach is referred to as the approximation by inward integration (AII).

Let $S(\langle\rangle, u) = 1$ and $w(\{\}) = 0$. Then Equation 14 is applicable to any mutation pattern. The computation of Equation 14 is quite manageable now. In particular, for up to two mutations, we have

$$S(\langle i\rangle,\mathbf{u}) = \sum_{(k)\in\mathcal{J}_g(i)} \mathbf{b}_k^T\mathbf{u} = \mathbf{a}_k^T\mathbf{u}, \tag{16}$$

$$S(\langle i,j\rangle,\mathbf{u}) = \sum_{(k,l)\in\mathcal{J}_g(i,j)} \left(\mathbf{b}_k^T\mathbf{u}\right)\left(\mathbf{b}_l^T\mathbf{u}\right) = \mathbf{u}^T\mathbf{B}_{ij}\mathbf{u}, \tag{17}$$

where $\mathbf{B}_{ij} = \sum_{(k,l)\ \in\ \mathcal{J}_g(i,j)} \mathbf{b}_k\mathbf{b}_l^T$. Therefore, the corresponding new approximations up to two mutations are

$$p_0 \approx e^{-\bar{\mathbf{t}}^T\mathbf{u}}, \tag{18}$$

$$p_i \approx e^{-[\bar{\mathbf{t}}-\bar{\mathbf{w}}(\langle i\rangle)]^T\mathbf{u}} \left(\bar{\mathbf{a}}_i^T\mathbf{u}\right), \tag{19}$$

$$p_{ij} \approx e^{-[\bar{\mathbf{t}}-\bar{\mathbf{w}}(\langle i,j\rangle)]^T\mathbf{u}} \left(\mathbf{u}^T\bar{\mathbf{B}}_{ij}\mathbf{u}\right), \tag{20}$$

where $\bar{\mathbf{B}}_{ij}$ is the mean $\mathbf{B}_{ij}$ over all genealogies. Note that $\mathbf{B}_{ij}$ is not the same as $\mathbf{A}_{ij}$ due to constraints on the pair of branches that are compatible with the observed pattern. Gao *et al.* (2011) recognized the masking effect of mutations and

Table 2 The expected numbers of mutation patterns and quality of approximations in 8625 families, each having 20 offspring

		i =							
u(× 10^4)		0	1	2	3	4	5	6	7
1	NP_i	8,326.5	293.9	4.6	0.0	0.0	0.0	0.0	0.0
	D_i	−0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0
	D'_i	0.2	−0.1	−0.1					
5	NP_i	7,232.6	1,286.7	100.9	4.7	0.1	0.0	0.0	0.0
	D_i	−1.7	1.4	0.3	0.0	0.0	0.0	0.0	0.0
	D'_i	−1.7	14.1	−17.4					
10	NP_i	6,067.3	2,179.3	344.1	32.2	2.0	0.1	0.0	0.0
	D_i	−5.7	3.8	1.6	0.2	0.0	0.0	0.0	0.0
	D'_i	−38.6	131.9	−127.5					
50	NP_i	1,504.0	2,950.4	2,468.7	1,206.0	389.8	88.8	14.7	1.5
	D_i	−31.3	−25.5	12.7	23.8	13.6	4.5	1.0	0.1
	D'_i	−4,901.0	11,637.7	−8,438.5					

u, mutation rate; NP_i, expected number of occurrences based on exact probability; D_i, D values based on the AII; D'_i, D values based on the ATE.

estimated $\mathbf{A}_{ij}$ by $\mathbf{B}_{ij}$. We show in a later section that when three or more mutations in a family are rare, then both the ATE and the AII give excellent approximations to the true probabilities. With an increasing number of families with more than two mutations, we find that the new approach provides a more accurate approximation to Equation 6 than those by Gao *et al.* (2011).

The Likelihood Inference

Suppose there are in total m different mutation patterns in the data set, $c_1, \ldots, c_n$, and n_i is the occurrence of pattern c_i. Then, the likelihood of the data is

$$L = \prod_{i=1}^{m} Pr(c_i)^{n_i}, \tag{21}$$

where $Pr(c_i)$ is the probability of pattern c_i. Based on the new scheme for estimating the probabilities of each pattern, the maximum-likelihood estimates, $\hat{\mathbf{u}}$, of $\mathbf{u}$ can be derived from $\ln(L)$, which is

$$\ln(L) = -\sum_{i=1}^{m} n_i \left[\bar{\mathbf{t}} - \bar{w}(c_i)\right]^T \mathbf{u} + \sum_{i=1}^{m} n_i \ln\left[\bar{S}(c_i, \mathbf{u})\right]. \tag{22}$$

The asymptotic covariance of the estimates $\hat{\mathbf{u}}$ can also be obtained as the inverse of matrix $\mathbf{V} = -(\partial^2 \ln L / \partial u_k \partial u_l)|_{\mathbf{u}=\hat{\mathbf{u}}}$, whose computation is described in the *Appendix*. Let $\mathbf{r}^T = (r_1, \ldots, r_I)$, where r_k is the number of cell divisions in the kth interval. Then per generation mutation rate u can be estimated as

$$\hat{u} = r_1\hat{u}_1 + r_2\hat{u}_2 + \ldots + r_I\hat{u}_I. \tag{23}$$

The variance of this estimate is $\mathrm{Var}(\tilde{u}) = \mathbf{r}^T\mathbf{V}^{-1}\mathbf{r}$. Suppose the total number of mutant lines in the experiment is M and the total number of lines screened is N. Then an alternative estimate of u is $\tilde{u} = M/N$, which is unbiased regardless of whether mutation rates during development are the same (Fu and Huai 2003). A hypothesis can be tested through the likelihood-ratio test. For example, for testing the null hypothesis (H_0) that mutation rates at different cell divisions are all equal, against the alternative hypothesis H_1 that rates have no constraint, the test statistic

$$\mathrm{Lr} = -2(\ln(L_0) - \ln(L_1)) \tag{24}$$

follows asymptotically the χ^2-distribution with $I - 1$ d.f.

Table 3 Maximum-likelihood estimates of u_1, u_2, u_3, and u_4(× 10^4) with 20 offspring from each of n families

u_1, u_2, u_3, u_4	n = 1,000	n = 10,000
4, 4, 4, 4	4.21, 3.86, 4.04, 3.98[a]	3.95, 3.94, 4.02, 3.97
	4.80, 2.57, 1.37, 1.51[b]	1.80, 0.99, 0.48, 0.62
8, 4, 4, 4	7.83, 4.11, 3.93, 4.04	7.95, 3.94, 4.02, 3.97
	6.50, 3.02, 1.51, 1.53	2.35, 1.11, 0.53, 0.50
4, 4, 4, 8	4.19, 3.83, 4.07, 7.91	3.97, 3.94, 4.02, 7.96
	4.77, 2.62, 1.44, 1.72	1.84, 0.97, 0.50, 0.56
6, 4, 4, 6	5.93, 3.97, 4.01, 5.97	5.99, 3.92, 4.04, 5.95
	5.68, 2.82, 1.47, 1.62	2.11, 1.04, 0.52, 0.53
3, 6, 6, 3	3.67, 5.58, 6.13, 2.96	3.03, 5.83, 6.06, 2.95
	4.73, 3.00, 1.62, 1.67	1.75, 1.07, 0.56, 0.55

Result for each case is based on 1000 simulated data sets.
[a] Mean estimates.
[b] Standard deviations.

Cell Propagation and Simulation of Cell Genealogy

A discrete generation model is used for the propagation of cells in the germline lineage. We introduce two alternative modes of cell propagation.

Let $N(i)$ be the size of the ith population, which can be divided into two groups, one [size $N_0(i)$] without a sister cell and one [size $N_1(i)$] with one sister cell [$N(i) = N_0(i) + N_1(i)$]. The first mode of propagation assumes that for each cell in the ith population, the probability of having k(k = 0, 1, 2) daughter cell(s) in the ith population is p_k.

This mode of cell propagation is fully determined when the values of p_i are specified. $p_2 = 1$ corresponds to the case in which each cell yields two daughter cells, which is considered to be the default situation. Another special case is that every cell produces at least one daughter cell, which corresponds to $p_0 = 0$. For two randomly selected cells from the ith population, the probability that they will coalesce in the $i - 1$th population is

$$\frac{N_1(i)}{N(i)(N(i) - 1)}. \tag{25}$$

That is, two cells will coalesce if and only if the first cell selected has a sister cell [with probability $N_1(i)/N(i)$] and the second cell selected is its sister cell [with probability $1/(N(i - 1) - 1]$. When there are multiple pairs of cells being considered, multiple coalescence can occur, which is usually not allowed in the conventional coalescent theory. The exact probabilities of any particular pattern of coalescence (for example, two pairs of coalescence, five pairs of coalescence,

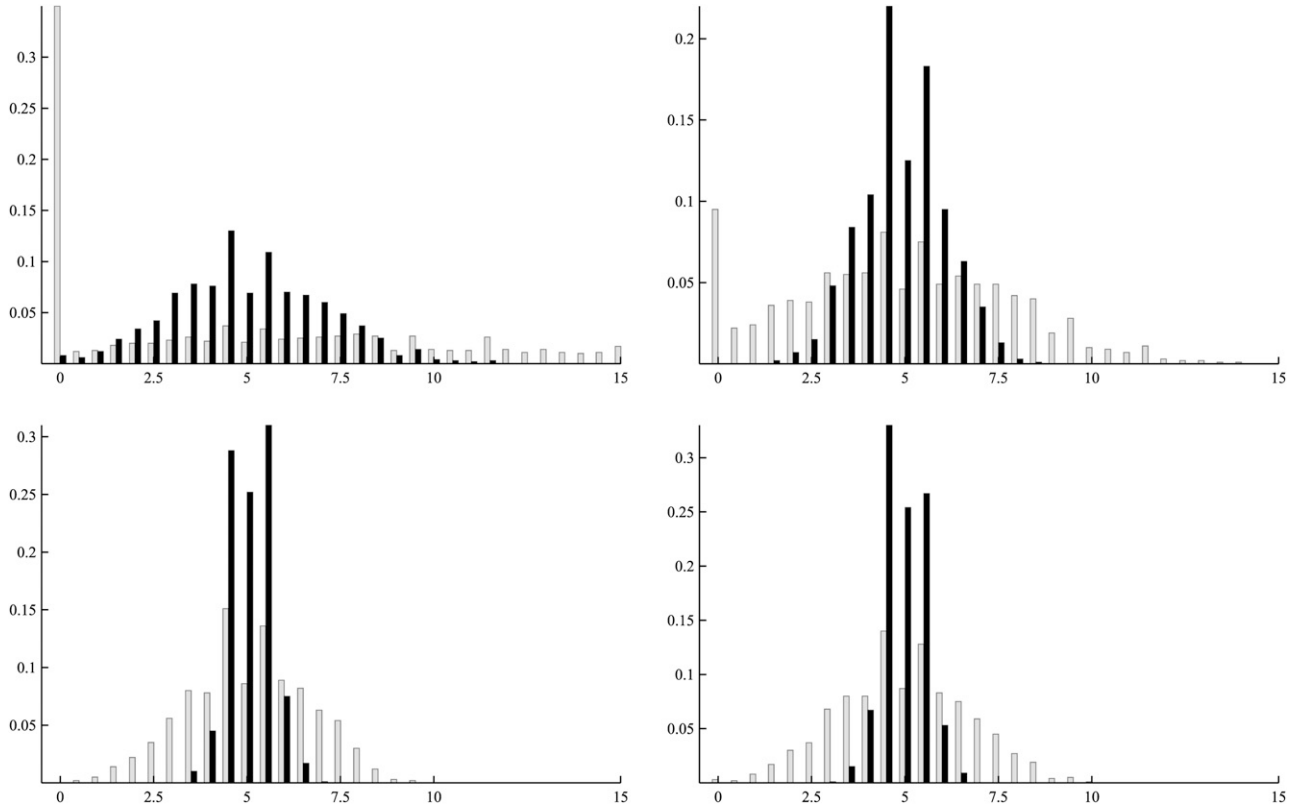

Figure 4 Distributions of the estimates for u_1, u_2, u_3, and u_4 (from top down). In each of the plots, shaded bars correspond to 1000 families and solid bars to 10,000 families (labels for the *x*-axis are multiplied by 10^4).

etc.) can be given analytically although they are not necessary for our purpose. What is critical is a proper algorithm to simulate this process as is discussed later in this section.

An alternative mode of cell propagation is as follows. Assume that each cell in the $(i-1)$th population divides to yield two daughter cells and the cells in the ith population are a random sample (without replacement) from these $2N(i-1)$ daughter cells. This mode of cell propagation is recognized when a range condition, such as $N(i) \in [a, b]$, is specified. In such a case, $N(i)$ is assumed to be a random integer in the given range $[a, b]$. Then the probability that two randomly selected cells from the ith population will coalesce in the $(i-1)$th population is

$$\frac{1}{2N(i-1)-1}, \tag{26}$$

which occurs only if the second cell selected is the sister cell of the first one. Again the probability of multiple coalescence can be derived. However, the sampling process will also yield $N(i)$ and $N_1(i)$; thus the coalescent probability is also given by Equation 25.

Simulation algorithms

A forward–backward two-step algorithm was used in Gao *et al.* (2011) and will continue to be used here. The first (forward) step is to simulate a history of the population sizes and the second (backward) step is to simulate the genealogy given the history of the populations sizes as follows.

Forward algorithm: Simulation of cell population dynamics: Given the value of $N(i-1)$, the value of $N(i)$ is simulated according to the transition mode. Meanwhile, the values of $N_k(i)$, $(k = 1, 2)$ are recorded.

Backward algorithm: Simulation of cell genealogy: Given a collection of n cells from the ith population:

1. Create an array of $N(i)$ integers as follows: $N_1(i)$ integers from 1 to $N_1(1)$ and two copies of each integer from $N_1(i) + 1$ to $N_1(i) + N_2(i)$.
2. Take a random sample of size n from the above array. If two integers in the sample are the same, a coalescent event occurs.
3. Update the collection of cells and repeat steps 1 and 2 until the 0th population (the zygote) is reached.

In the forward step, the cell lineage splits into two subpopulations after the 14th division and enters the stem cell lineage, which is specified by mode 1 with $p_1 = 1 - e$, $p_2 = 2$ for small values of e. After the 31st division, each of the differentiated cells from stem cells goes into spermatogenesis, resulting in 64 sperm. After the 37th cell division, the sperm population consists of sperm that are derived

from differentiated cells that have experienced different numbers of divisions. Therefore, the number of cell divisions for each of the sperm in a random sample can be different.

Numerical Results

To investigate statistical properties of the inference framework, the 36 cell divisions are divided into four intervals: [1, 3], [4, 14], [15, 31], and [32, 36], representing, respectively, the early cleavage, late cleavage, the stem-cell stage, and the spermatogenesis stage. Situations with maximal cell divisions >36 are also considered, so that different sperm in a sample might have experienced different numbers of divisions (see Figure 3). In such situations, the meaning of the last two intervals needs to be modified. For example, if a maximum of 38 divisions is allowed, then the last interval corresponds to the last 5 cell divisions, which for some lineages are from the 34th to the 38th division, while for some they are from the 32nd to the 36th division; and the second interval thus includes the 15th division in whatever is not included in the last interval.

Simulation of sample genealogy

As pointed out earlier, the process of simulating a sample genealogy is similar to that in Gao *et al.* (2011), with the exception of the gametogenesis stage. The process consists of forward and backward steps. The former is guided by a number of constraints about the population sizes mentioned in the Introduction. For example, after the 8th division, there are 256 cells from which only 4–7 cells become PGCs. Table 1 lists all the used constraints for population sizes during germline development. After the 31st division in the forward process, each of the differentiated cells will go into gametogenesis, which progresses through 5 additional divisions to produce 64 sperm. This aspect of the development is now explicitly modeled.

The backward process of the simulation is the same as that in Gao *et al.* (2011) except that the efficiency of the program has been improved. The resulting population dynamics with relation to the sample genealogy are illustrated in Figure 3. It is important to simulate a large number of genealogies from which the values of various coefficients in the inference framework can be obtained. To deal with up to four mutations in a family, we found that in general 250,000 genealogies are sufficient.

Accuracy of the approximations to the probabilities of mutation patterns

Since the expected numbers of occurrences and the difference in the expected numbers of occurrences are critical to the statistical inference, we use the following index to measure the accuracy of the approximations,

$$D_i = N\left(P_i - \hat{P}_i\right), \tag{27}$$

where P_i is the exact probability for a mutation pattern with i mutations, $\hat{P}_i$ is its approximation, and N is the number of families, which is set to 8625. The probabilities were estimated with 2 million simulated genealogies. Table 2 gives the results for several mutation rates. In all cases the AII is better than the ATE; the AII performs well for a wide range of mutation rates, including a very large mutation rate. The ATE appears to be sufficient for a mutation rate up to $\sim u \times 10^{-4}$, but with a rate closer to 10×10^{-4}, its errors becomes too large, in addition to not being able to handle more than two mutations. We focus on further studying the statistical properties of the AII because of its obvious superiority.

Table 4 Critical values for likelihood-ratio tests with μ = 0.0004

	Asmpt		n = 1,000		n = 5,000		n = 10,000	
$L_{i,j}$	c_5	c_1	c_5	c_1	c_5	c_1	c_5	c_1
$L_{0,1}$	5.99	9.21	6.18	9.56	6.17	9.35	6.23	9.44
$L_{0,2}$	3.84	6.64	3.47	5.26	3.96	6.88	3.93	6.66
$L_{0,3}$	7.82	11.35	7.51	10.55	8.01	11.42	8.21	11.98

c_5 and c_1 are, respectively, the upper 5% and 1% critical values. Simulation results for each case are based on 10,000 replicates.

Maximum-likelihood estimate of u

One major outcome of the inference is the maximum-likelihood estimate of the mutation rate **u**; thus it is important to understand the properties of the estimates, which were carried out using simulations. Table 3 shows the means and standard deviations of the maximum-likelihood estimates of u_1, u_2, u_3, and u_4 for several cases. The results show that the maximum-likelihood estimates are slightly biased but the bias decreases with increase in family number, which is expected from the well-known properties of the maximum-likelihood method. The standard errors of estimating $u_1, \ldots, u_4$ differ from each other, with those for u_3 and u_4 being the smallest and that for u_1 the largest. This pattern agrees with the fact that there are many more mutations that result in a smaller mutant size, most of which likely occurred during the third and fourth time intervals. As a result, there are more observations from these two intervals that lead to more accurate estimates.

Figure 4 shows the distributions of estimates of mutation rates corresponding to the first row of Table 3. Two obvious features from these distributions are as follows. The first is that with increased family number, each distribution becomes more concentrated around the true mutation rate. The second is that judging from the spread of the distributions, the quality of estimations for u_3 and u_4 is better than that for u_1 and u_2. Among the four, the quality of estimating u_1 is the poorest. These features agree well with the patterns of standard deviations in Table 3.

Likelihood-ratio test

Being able to obtain maximum-likelihood estimates under different assumptions also allows us to examine the distribution of the likelihood-ratio test. Various hypotheses about the pattern of mutation rates can be tested, as reported in Gao *et al.* (2011); however, the following four are

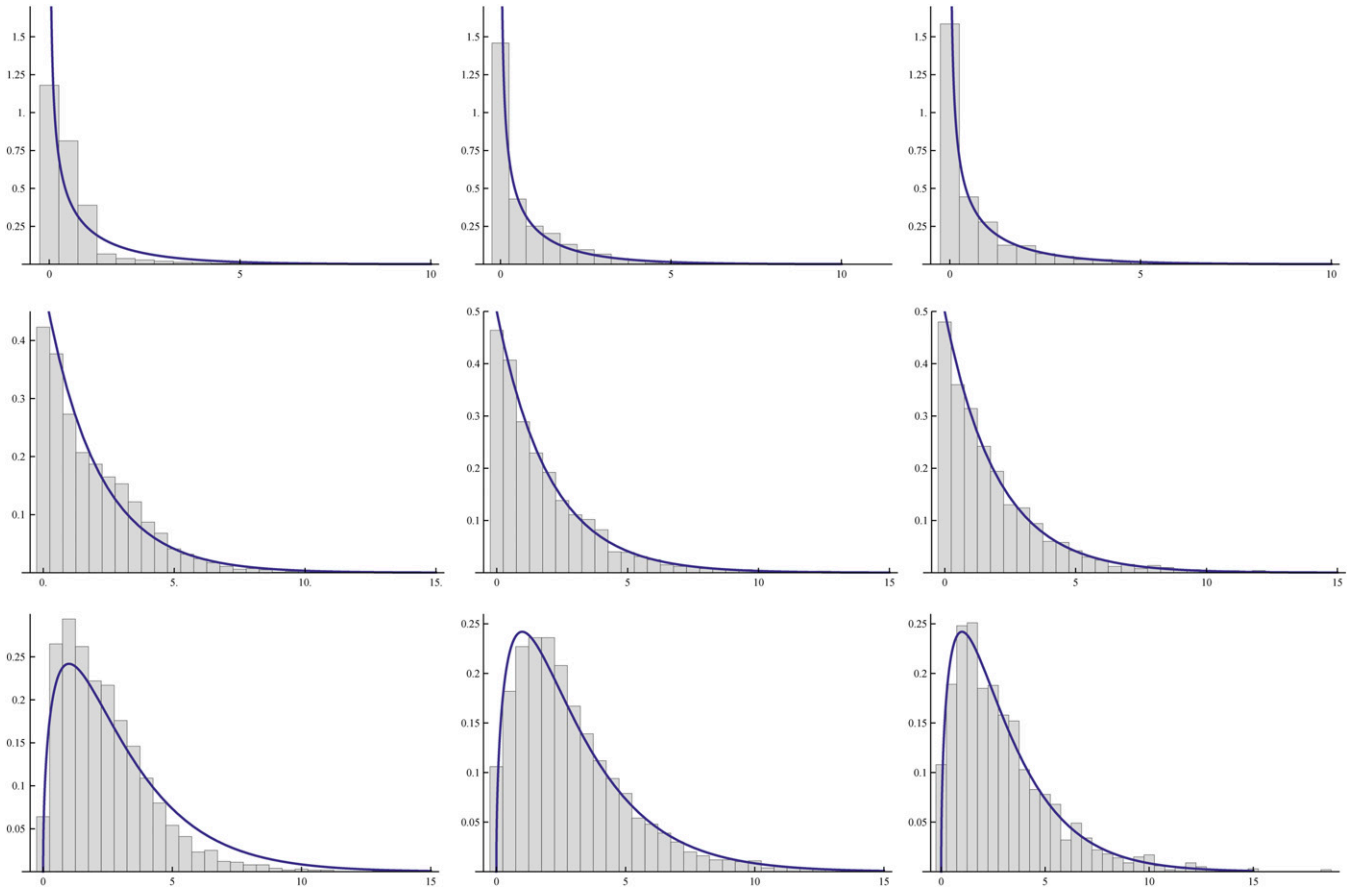

Figure 5 Distributions of likelihood-ratios $L_{0,1}$ (top three), $L_{0,2}$ (middle three), and $L_{0,3}$ (bottom three) for 200 (left), 1000 (center), and 10,000 (right) families (bottom), with smooth curves being the χ^2 densities.

representative, one for each value of the degrees of freedom: H_0, rates are constant; H_1, the last three are the same; H_2, the first two are the same; and H_3, no constraint.

Let L_{ij} be the log-likelihood ratio statistics between the ith and jth hypotheses. When H_0 is true, it is expected that $L_{0,1}$, $L_{0,2}$, and $L_{0,3}$ follow asymptotically χ^2-distributions with 1, 2, and 3 d.f., respectively. In the simulations, constant mutation rates are used and for each simulated sample, the maximum likelihood under each hypothesis is found, which leads to the likelihood-ratio statistics. Table 4 shows the upper-tail critical values for these three statistics. Comparing these critical values with the critical values of χ with 1, 2, and 3 d.f., respectively, indicates that the these critical values agree reasonably well with the asymptotic values with sample sizes as small as 500. The distributions of these statistics are given in Figure 5 for two different sample sizes, which shows the overall excellent agreement of the empirical density with asymptotic ones.

Reanalysis of the data

The data being reanalyzed here consist of those presented in Table 1 of Gao *et al.* (2011) and 7 additional families, 3 of which have three mutations and 1 of which has four mutations, giving thus a total of 8,625 families. For convenience of comparison, we used the same division of intervals: [1, 1], [2, 2], [3, 14], [15, 31], and [32, 36]. Table 5 shows the maximum-likelihood estimates using both the ATE and the AII (for the sake of space, only the results for four of the eight hypotheses considered in Gao *et al.* 2011 are given), while Table 6 gives the results of the likelihood-ratio tests. Comparing the entries of the ATE in these tables to those in Tables 4 and 5 of Gao *et al.* (2011), one can see that the differences are minimal. Furthermore, comparing the estimates by the ATE to those by the AII shows that the differences are also minor in almost all cases. Therefore, the improved method does not change the conclusions made previously. These analyses also included the case in which 38 cell divisions were assumed. In such situations, the patterns of the likelihood-ratio tests (Table 6) suggest that the mutation rates for the second, third, and fourth intervals may also be different, although the evidence is only marginal.

While it is comforting that the reanalysis reinforces the conclusions made earlier, this should not be regarded as the AII lacking importance. When the number of families with more than two mutations increases, one can expect to see increasing differences and a more rigorous new method than the ATE will become necessary. To illustrate, we simulated sets of 8625 families with four intervals of cells

Table 5 Maximum-likelihood estimates of $u \times 10^3$ under several hypotheses

Hypothesis	u_1	u_2	u_3	u_4	u_5	$-\ln(L)$
H_1	0.347[a]	0.347	0.347	0.347	0.347	4519.0
	0.343[b]	0.343	0.343	0.343	0.343	4494.8
	0.321[c]	0.321	0.321	0.321	0.321	4441.9
H_3	2.284	2.284	0.001	0.001	1.217	4153.2
	2.284	2.284	0.001	0.001	1.217	4129.0
	2.249	2.249	0.001	0.046	1.050	4126.4
H_5	4.864	0.001	0.007	0.007	1.217	4139.6
	4.864	0.001	0.007	0.007	1.217	4115.4
	4.655	0.001	0.037	0.037	1.050	4113.1
H_8	5.072	0.001	0.002	0.007	1.217	4139.5
	4.864	0.001	0.002	0.007	1.217	4115.4
	4.815	0.001	0.001	0.058	1.032	4110.8

H_1, $u_1 = \ldots = u_5$; H_2, $u_2 = u_3 = u_4$; H_3, $u_1 = u_2$; H_4, $u_2 = u_3$; H_5, $u_3 = u_4$; H_6, $u_4 = u_5$; H_7, $u_1 = u_5$; and H_8, no constraint.
[a] Estimates based on the ATE.
[b] Estimates based on the AII.
[c] Estimates based on the AII with 38 divisions.

Table 6 The values of the log-likelihood ratio test of various hypotheses listed in Table 5

	i =						
Contrast	2	3	4	5	6	7	8
H_1 *vs.* H_i	758.8	731.6	758.9	758.8	293.2	714.8	758.9
	758.7	731.7	758.8	758.7	292.7	714.6	758.8
	657.1	630.9	662.1	657.5	293.6	609.9	662.1
H_i *vs.* H_8	0.1	27.3	0.0	0.1	465.7	44.2	
	0.1	27.1	0.0	0.1	466.1	44.2	
	5.0	31.2	0.0	4.6	368.6	52.3	

[1, 1], [2, 14], [15, 33], and [34, 38], using two sets of mutation rates, one being equal rates for all the intervals and the other being one that produces mutational patterns resembling those from the experiment, which will be subjected to detailed analysis elsewhere. Table 7 shows the comparison of the two methods from which it is obvious that the ATE leads to underestimation of mutations rates. In the first case (equal mutations rates), the bias in estimating u_i increases with i and u_4 is about two-thirds of the true value. MSEs of the estimates also suggest that the AII performs considerably better (except for u_1 for which there is little difference between the two methods). For the second case, the downward bias in the estimates by the ATE is also obvious in all u_i and the MSEs by ATE are appreciably larger than those by the AII. Another shortcoming of the ATE is that due to differential degrees of underestimation of u_i, it can lead to rejection of certain hypotheses more often than specified by the given nominal level of significance. For example, for testing the hypothesis $u_1 = u_4$, the ATE in the first case leads to nearly 12% rejection while the AII has <5% rejection at the 5% significance level. These results agree well with an earlier conclusion made from Table 2, which is that when the mean mutation rate is $>10^{-4}$, the ATE starts to lose accuracy.

Discussion

This article presents a significantly improved framework for statistical inference of germline mutation rates, with specific reference to *D. melanogaster*. This framework includes coalescent theory and an improved algorithm for simulating sample genealogies to obtain various coefficients, a method for computing the probabilities of mutation patterns, and a likelihood method for estimating mutation rates and testing hypotheses about the pattern of mutation rates. Statistical properties of the inference framework were investigated through simulation. The new approximation method for computing the probabilities of mutation patterns is more accurate than the previous method by Gao *et al.* (2011), particularly when mutation rates are high. Nevertheless, the previous method is sufficiently accurate for the data reported by Gao *et al.* (2011), and thus all major conclusions remain intact. The new likelihood-based inference exhibits desirable and expected properties, including reduced bias and smaller standard deviation with increasing number of families. The asymptotic χ^2-distribution for the likelihood-ratio test is sufficiently accurate when the number of families is reasonably large and for the sample size reported in Gao *et al.* (2011).

This theoretical study paves the way for analysis of data from families with three and four mutations. Furthermore, the theoretical framework reported here can be adapted for studying germline mutational distribution in other organisms and for analyzing data generated through DNA typing or sequencing sperm samples. To apply the framework to other organisms the nature and mode of cell propagation of their germline populations would need to be determined. Application to data generated by DNA typing or sequencing will likely have its own issues, such as data accuracy, but since it is more economical to sequence larger regions with fewer samples than shorter regions with larger samples, observing multiple mutations will likely be the norm. Therefore, the statistical framework of inference described here will be relevant.

Table 7 Comparison of estimates of *u* based on the ATE and the AII when mutation rates are relatively high

Rates	Values ($\times 10^4$)	AII	MSE	ATE	MSE
u_1	2.000	2.098	0.648	1.923	0.627
u_2	2.000	1.898	0.031	1.602	0.186
u_3	2.000	2.015	0.007	1.587	0.177
u_4	2.000	1.934	0.016	1.417	0.348
u_1	20.000	19.581	2.4772	18.904	3.4325
u_2	0.200	0.203	0.0086	0.139	0.0111
u_3	0.167	0.165	0.0020	0.152	0.0021
u_4	5.000	4.943	0.0112	4.315	0.4742

MSE, mean square error. For each parameter set, 500 sets of 8625 families were generated.

Acknowledgments

I thank Sara Barton for her editorial assistance. This work was partly supported by grants from the Chinese National Science Foundation (30570248 and 91231120 YF) and by the Betty Wheless Trotter Endowment Fund from The University of Texas Health Science Center.

Literature Cited

Ashburner, M., 1989 *Drosophila: A Laboratory Handbook*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

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.

Ewens, W. J., 2004 *Mathematical Population Genetics*. Springer-Verlag, New York.

Fu, Y. X., and H. Huai, 2003 Estimating mutation rate: How to count mutations? Genetics 164: 797–805.

Gao, J. J., X. R. Pan, J. Hu, L. Ma, J. M. Wu *et al.*, 2011 Highly variable recessive lethal or nearly lethal mutation rates during germline development of male *drosophila melanogaster*. Proc. Natl. Acad. Sci. USA 108(38): 15914–15919.

Gilbert, S. F., 2003 *Developmental Biology*, Ed. 7. Sinauer Associates, Sunderland, MA.

Greenspan, S. F., 1997 *Fly Pushing: The Theory and Practice of Drosophila Genetics*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Kingman, J. F. C., 1982 On the genealogy of large populations. J. Appl. Probab. 19A: 27–43.

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

Woodruff, R. C., J. J. N. Thompson, M. A. Seeger, and W. E. Spivey, 1984 Variation in spontaneous mutation and repair in natural population lines of *Drosophila melanogaster*. Heredity 58: 223–234.

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

Communicating editor: Y. S. Song

Appendix

Asymptotic Covariance of the Maximum-Likelihood Estimates

The asymptotic covariance matrix of the maximum-likelihood estimates $\hat{\mathbf{u}}$ is the inverse of the following matrix:

$$\mathbf{V} = -\left(\frac{\partial^2 \ln L}{\partial u_k \partial u_l}\right)\Bigg|_{\mathbf{u}=\hat{\mathbf{u}}}.$$

Since

$$\ln(L) = -\sum_{k=0}^{m} n_k(\mathbf{t}-s_k)^T\mathbf{u} + \sum_{k=0}^{m} n_k \ln(S_k),$$

it follows that

$$\frac{\partial \ln(L)}{\partial u_i} = \sum_{k=0} n_k \left[S_k^{-1}\frac{\partial S_k}{\partial u_i} - (\mathbf{t}-s_k)_i\right] \tag{A1}$$

$$\frac{\partial^2 \ln(L)}{\partial u_i \partial u_j} = \sum_{k=1} n_k S_k^{-2}\left[S_k\frac{\partial^2 S_k}{\partial u_i \partial u_j} - \frac{\partial S_k}{\partial u_i}\frac{\partial S_k}{\partial u_j}\right]. \tag{A2}$$

Since S is of the form

$$S = \sum_{i_1, i_2, \ldots, i_k} t(i_1, i_2, \ldots, i_k) u_{i_1} \ldots u_{i_k},$$

where $t(i_1, i_2, \ldots, i_k)$ is constant, it follows that

$$\frac{\partial S}{\partial u_i} = \sum_{i_1, i_2, \ldots, i_k} t(i_1, i_2, \ldots, i_k)\ \frac{\partial u_{i_1} \ldots u_{i_k}}{\partial u_i} \tag{A3}$$

$$\frac{\partial^2 S}{\partial u_i \partial u_j} = \sum_{i_1, i_2, \ldots, i_k} t(i_1, i_2, \ldots, i_k)\ \frac{\partial^2 u_{i_1} \ldots u_{i_k}}{\partial u_i \partial u_j}. \tag{A4}$$

Furthermore, let n_i be the number of i in $i_1, \ldots, i_k$; then

$$\frac{\partial u_{i_1} \ldots u_{i_k}}{\partial u_i} = \frac{n_i u_{i_1} \ldots u_{i_k}}{u_i} \tag{A5}$$

$$\frac{\partial^2 u_{i_1} \ldots u_{i_k}}{\partial u_i \partial u_j} = \frac{n_i n_j u_{i_1} \ldots u_{i_k}}{(u_i u_j)} \tag{A6}$$

$$\frac{\partial^2 u_{i_1} \ldots u_{i_k}}{\partial u_i \partial u_i} = \frac{n_i(n_i-1) u_{i_1} \ldots u_{i_k}}{(u_i^2)}. \tag{A7}$$

Putting the results of Equations A3–A7 into Equation A2, together with $\mathbf{u}$ replaced by $\hat{\mathbf{u}}$, will lead to the numerical value of $\partial^2 \ln L/\partial u_k \partial u_l$.