Statistical methods for analyzing Drosophila germline mutation rates

Yun-Xin Fu\(^1,2,\ast\)

\(^1\) Laboratory for Conservation and Utilization of Bio-Resources, Yunnan University, Kunming, China
\(^2\) Division of Biostatistics and Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, USA

\(^\ast\) To whom correspondence should be addressed (yunxin.fu@uth.tmc.edu)

Copyright 2013.
Abstract

Most studies of mutation rates implicitly assume that they remain constant throughout development of the germline. However, Gao et al. (2011) 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 datasets in which multiple mutations in the same family is common.
Introduction

Sperm and eggs experience many divisions since 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 is 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 or more 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; Drost and Lee, 1998; Gilbert, 2003) that 1) after the 8th cell division, about 4-7 cells become the primordial germ cells (PGC); 2) after the 12th division, the number of PGCs ranges from 23-52; 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.
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 \( \langle i_1, i_2, ..., i_k \rangle \) such that each element represents an identified mutation, its value equals 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, one leads to three mutants and another to two mutants. Also we use \( \langle > \) 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 \textit{Drosophila}, multi-generation mutation screening have been well developed (Muller 1928, Woodruff et al. 1984, Ashbumer 1989, Woodruff et al. 1996 and 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 paper, it should suffice to outline the structure of information and the type of mutation being examined. The recessive lethality \( d \) is defined as one minus the maximal survival rate of descendants who are homozygote for the mutation. The data reported by Gao et al. (2011) corresponds to those mutations with recessive lethality equal to 99 percent, that is, at least 99% of the homozygote flies of such a mutation will die at or soon after birth.
Once a cell acquires a mutation with recessive lethality $d$, further mutation(s) are 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. Fig. 1 shows examples of how mutations in a genealogy lead to different mutation patterns.

(Figure 1 about here)

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 will be referred to as the subtree of the branch $i$. For the genealogy shown in Fig. 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 $\Omega(i) \cap \Omega(j) = \emptyset$.

Suppose the germ cell divisions from a fertilized egg to sperm are divided into $I$ intervals. Let $[i, j]$ represent the interval from the $i$-th to the $j$-th cell divisions. Suppose the mutation rate per cell division for the $l$-th interval is $u_l$ and define $u = (u_1, ..., u_I)^T$. 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 a unique integer (how branches are numbered is immaterial). Define for the $i$-th branch, $b_{ij}$ as the number of divisions it contains from the $j$-th interval and $b_i$ is a vector with elements $b_{ij}, j = 1, ..., I$. That is, $b_i = (b_{i1}, ..., b_{ii})^T$. Define $\phi(i)$ as the size of the $i$-th branch, i.e., the number of descendants of the branch that are observed in the sample, and

$$a_k = \sum_{i: \phi(i) = k} b_i \quad \text{and} \quad t = \sum_{k=1}^{n} a_k . \tag{1}$$

Then $a_{kj}$ is the total number of cell divisions from the $j$th interval that are of size $k$ and the $k$-th element, $t_k$, of $t$ is the total number of cell divisions from the $k$-th interval. For branch $i$, let $w_i$ be the sum of lengths of all the branches in $\Omega(i)$, excluding the branch $i$ itself. That is, $w_i = -b_i + \sum_{k: \Omega(i)} b_k$. Fig. 2 illustrates aforementioned quantities in a genealogy of five alleles taken after the 5th division. It follows that $a_1 = \sum_{k=5}^{9} b_i = (0, 1, 5)^T$, $a_2 = b_2 + b_4 = (1, 3, 0)^T$, $a_3 = b_3 = (1, 1, 0)^T$, $a_4 = 0$, $a_5 = b_1 = (1, 0, 0)$ and an example of $w$ is that $w_4 = b_7 + b_8$ and $w_3 = b_4 + b_7 + b_8 + b_9$.

(Figure 2 about here)

In addition to $a$ and $t$, we will encounter other quantities that are functions of $b_i, i = 1, ..., n$, 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 shall use a bar over the variable to represent its expectation. For example

\[ \bar{t} = \int_g t \, dg, \quad \bar{a}_k = \int_g a_k \, dg. \] (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 \( e^{-t^T u} \). Since a single mutation leading to an observed pattern \(<i>\) must occur on a branch of size \( i \), it follows that

\[
Pr(<i> | g) = \sum_{k : \phi(k) = i} e^{-(t - b_k - w_k)^T u} \left( 1 - e^{-b_k^T u} \right) = \sum_{k : \phi(k) = i} e^{-(t - w_k)^T u} \left( e^{b_k^T u} - 1 \right)
\] (3)

where summation is taken over all the branches of size \( i \). In the summation, the first term \( e^{-(t - b_k - w_k)^T u} \) is the probability that there is no mutation outside the subtree of branch \( k \) and the second term \( (1 - e^{-b_k^T u}) \) 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 \(<i_1, ..., i_l>\) that

\[
Pr(<i_1, ..., i_l> | g) = \sum_{(k_1, ..., k_l) \in J_g(i_1, ..., i_l)} \left[ e^{-(t - w_{k_1} - ... - w_{k_l})^T u} \prod_{i=1}^l \left( e^{b_{k_i}^T u} - 1 \right) \right]
\] (4)

where \( w_{k_1, ..., k_l} = \sum_i w_{k_i} \) and \( J_g(i_1, ..., i_l) \) is the collection of the branch sets of genealogy \( g \) on which mutations can lead to the observed mutational pattern. That is,

\[
J_g(i_1, ..., i_l) = \{ (k_1, ..., k_l) : \phi(k_j) = i_j \text{ for } i = 1, ..., l \text{ and } \Omega(k_i) \cap \Omega(k_j) = \emptyset \text{ for } i \neq j \}. \] (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 \(<i_1, ..., i_l>\) as

\[
Pr(<i_1, ..., i_l>) = \int_g \sum_{(k_1, ..., k_l) \in J_g(i_1, ..., i_l)} \left[ e^{-(t - w_{k_1} - ... - w_{k_l})^T u} \prod_{i=1}^l \left( e^{b_{k_i}^T u} - 1 \right) \right] dg.
\] (6)

This formula provides the basis for the proposed inferences and detailed analysis of Drosophila data. For any given mutation pattern \(<i_1, ..., i_l>\) and \( u \), the probability can be evaluated as the average of \( Pr(<i_1, ..., i_l> | g) \) (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 \( u \) 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, $e^{-t^T u} \approx 1 - t^T u + \frac{1}{2} u^T (tt^T) u$. Define $A_{ij} = a_i a_j^T$, $A_i = a_i t^T$ and $A_0 = tt^T$. Then the probabilities $p_0 = Pr(<>)$, $p_i = Pr(<i>)$ and $p_{ij} = Pr(<i,j>)$ can be approximated (Gao et al. 2011) by

$$p_0 \approx 1 - t^T u + \frac{1}{2} u^T A_0 u,$$  \hspace{1cm} (7)  

$$p_i \approx a_i^T u - u^T A_i u,$$  \hspace{1cm} (8)  

$$p_{ij} \approx \frac{2 - \delta_{i,j}}{2} u^T A_{ij} u$$  \hspace{1cm} (9)

where $\delta_{i,j} = 1$ if $i = j$ and 0 otherwise. This method of approximation will be referred to as the Approximation by Taylor Expansion (ATE). Although these approximations cover up 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 large number of coefficients in higher order terms, their computations make the ATE inefficient.

Since typically $b_k^T u << 1$ in Equation (6), $e^{b_k^T u} - 1 \approx b_k^T u$. Furthermore one can simplify the expression by replacing $w$ for each combination of branches by its average value, and arrive at

$$Pr(<i_1, ..., i_l|g) \approx \sum_{(k_1, ..., k_l) \in \mathcal{J}_g(i_1, ..., i_l)} \left[ e^{u^T b_{k_1} \ldots b_{k_l} u} \right]$$  \hspace{1cm} (10)  

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

$$S(<i_1, ..., i_l>, u) = \sum_{(k_1, ..., k_l) \in \mathcal{J}_g(i_1, ..., i_l)} \left[ \prod_{i=1}^l b_{k_i}^T u \right] = \sum_{l_1, ..., l_m} a_{l_1} \ldots a_{l_m} u_{l_1} \ldots u_{l_m}$$  \hspace{1cm} (12)

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

$$Pr(<i_1, ..., i_l>) \approx \int_g e^{-u^T w(<i_1, ..., i_l>) u} S(<i_1, ..., i_l>, u) d\mathcal{g}.$$  \hspace{1cm} (13)

7
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

$$
Pr(<i_1, \ldots, i_l>) \approx e^{-[\bar{t} - \bar{w}(<i_1, \ldots, i_l>)]^T u \bar{S}(<i_1, \ldots, i_l>, u)}
$$

(14)

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

$$
\bar{S}(<i_1, \ldots, i_m>, u) = \sum_{l_1, \ldots, l_m} \bar{a}_{l_1 \ldots l_m} u_{l_1} \ldots u_{l_m}
$$

(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 will be referred to as the Approximation by Inward Integration (AII).

Let \( S(<>, u) = 1 \) 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

$$
S(<i>, u) = \sum_{(k) \in \mathcal{J}_g(i)} b_k^T u = a_k^T u,
$$

(16)

$$
S(<i, j>, u) = \sum_{(k,l) \in \mathcal{J}_g(i,j)} (b_k^T u)(b_l^T u) = u^T B_{ij} u
$$

(17)

where \( B_{ij} = \sum_{(k,l) \in \mathcal{J}_g(i,j)} b_k b_l^T \). Therefore, the corresponding new approximations up to two mutations are

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

(18)

$$
p_i \approx e^{-[\bar{t} - \bar{w}(<i>)]^T u (\bar{a}_i^T u)},
$$

(19)

$$
p_{ij} \approx e^{-[\bar{t} - \bar{w}(<i,j>)]^T u (u^T B_{ij} u)}
$$

(20)

where \( B_{ij} \) is the mean \( B_{ij} \) over all genealogies. Note that \( B_{ij} \) is not the same as \( A_{ij} \) 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 estimated \( A_{ij} \) by \( B_{ij} \). We shall show in a later section that when three or more mutations in a family are rare, then both the ATE and AII give excellent approximations to the true probabilities. With 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 dataset, $c_1, ..., c_n$, and $n_i$ is the occurrence of pattern $c_i$. Then, the likelihood of the data is

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

(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, $\hat{u}$, of $u$ can be derived from $\ln(L)$ which is

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

(22)

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

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

(23)

The variance of this estimate is $Var(\hat{u}) = r^T V^{-1} r$. 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 $\bar{u} = M/N$ which is unbiased regardless if 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

$$Lr = -2(ln(L_0) - ln(L_1))$$

(24)

follows asymptotically the $\chi^2$ distribution with $I - 1$ degree of freedom.

Cell propagation and simulation of cell genealogy

A discrete generation model will be used for the propagation of cells in the germ-line lineage. We will introduce two alternative modes of cell propagation.
Let $N(i)$ be the size of the $i$-th 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 $i$-th population, the probability of having $k (k = 0, 1, 2)$ daughter cell(s) in the $i$-th 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 $i$-th population, the probability that they will coalesce in the $i - 1$-th population is

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

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 probability of any particular pattern of coalescence (for example, two pairs of coalescence, five pairs of coalescence, 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 will be 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 $i$-th 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 $i$-th population will coalesce in the $(i - 1)$-th population is

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

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 $i$-th 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 occurs,

3. Update the collection of cells and repeat steps 1 and 2 until the 0-th population (the zygote) is reached.

In the forward step, the cell lineage split into two sub-populations 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 31th division, each of the differentiated cells from stem-cells goes into spermatogenesis resulting in 64 sperm. After the 37th cell divisions, 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, stem-cell and spermatogenesis stage. Situations with maximal cell divisions larger than 36 is also considered, so that different sperm in a sample might have experienced different number of divisions (see Fig. 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 five cell divisions, which for some lineages are from the 34th to the 38th division, while for some it is the 32th to the 36th division; and the second interval thus includes the 15th division to 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 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 Introduction. For example, after the 8th divisions, there are 256 cells from which only 4 to 7 cells become primordial germ cells (PGC). Table 1 lists all the used constraints for population sizes during germline development. After the 31th 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 is illustrated by 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 4 mutations in a family, we found that in general a quarter million genealogies are sufficient.

(Table 1 and Figure 3 about here)

**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 will use the following index to measure the
accuracy of the approximations:

\[ D_i = N(P_i - \hat{P}_i) \]  \hfill (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 two 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 about \( u \times 10^{-4} \), but with a rate closer to \( 10 \times 10^{-4} \), its errors becomes too large, in addition to not being able to handle more than two mutations. We will focus on further studying the statistical properties of the AII because of its obvious superiority.

(Table 2 about here)

**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 \( u_1, ..., u_4 \) 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, and most of which likely occurred during the third and fourth time intervals. As a result, there are more observations from these two intervals which lead to more accurate estimates.

(Table 3 and Figure 4 about here)

Fig. 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 $u_4$ are 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 representative, one for each value of the degrees of freedom:

- $H_0$: Rates are constant
- $H_1$: Last three are the same
- $H_2$: First two are the same
- $H_3$: No constraint

Let $L_{ij}$ be the log-likelihood ratio statistics between the $i$th and $j$th hypothesis. When $H_0$ is true it is expected that $L_{0,1}, L_{0,2}$ and $L_{0,3}$ follow asymptotically $\chi^2$ distributions with 1, 2 and 3 degrees of freedom, 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 $\chi$ with 1, 2 and 3 degrees of freedom, respectively, indicate 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.

(Table 4 and Figure 5 about here)

**Re-analysis of the data**

The data being re-analyzed 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 has four mutations, 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 AII (for the sake of space, only the results for four out 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 Table 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 analysis also included the case in which 38 cell divisions was assumed. In such situations, the patterns of the likelihood ratio tests (Table 6) suggests that the mutation rates for the second, third and fourth interval may also be different, although the evidence is only marginal.

(Table 5 and 6 about here)

While it is comforting that the reanalysis reinforces the conclusions made earlier, this should not be regarded as lacking importance of the AII. 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 8,625 families with four intervals of cells [1, 1], [2, 14], [15, 33] and [34, 38] using two sets of mutation rates, one being equal rates for all the intervals and another 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 $2/3$ 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 less than 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 over $10^{-4}$, the ATE starts to loose accuracy.

(Table 7 about here)
Discussion

This paper presents a significantly improved framework for statistical inference of germline mutation rates, with specific reference to *Drosophila 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 chi-square 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 3 and 4 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 of larger samples, observing multiple mutations will likely be the norm. Therefore, the statistical framework of inference described here will be relevant.
Acknowledgment

This work was partly support by grants from the Chinese NSF (30570248 and 91231120 YF), and the Betty Wheless Trotter Endowment Fund from The University of Texas Health Science Center at Houston, USA. Editorial assistance by Sara Barton is appreciated.

Literature Cited


Appendix: Asymptotic covariance of the maximum likelihood estimates

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

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

Since

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

it follows that

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

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

Since $S$ is of the form

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

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

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

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

Furthermore, let $n_i$ be the number of i in $i_1,...,i_k$, then

$$\frac{\partial u_{i_1}...u_{i_k}}{\partial u_i} = n_i u_{i_1}...u_{i_k} / u_i$$

$$\frac{\partial^2 u_{i_1}...u_{i_k}}{\partial u_i \partial u_j} = n_i n_j u_{i_1}...u_{i_k} / (u_i u_j)$$

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

Putting the results of Equations (30) - (34) into (29), together with $u$ replaced by $\hat{u}$, will lead to the numerical value of $\frac{\partial^2 \ln L}{\partial u_k \partial u_l}$.
Figure legends:

Fig. 1. Examples of mutations and resulting mutation patterns. (a) single mutation leading to mutation pattern $<1>$, (b) two mutations leading to mutation pattern $<2,1>$ and (c) two mutations leading to mutation pattern $<3>$ due to the second mutation masked by the first one.

Fig. 2. An example of the genealogy of five alleles (a-e) sampled from the cell population after the 5th division. Each branch is identified by a nearby integer. The sizes of the branches are $\phi(1) = 5, \phi(2) = 2, \phi(3) = 3, \phi(4) = 2, \phi(5) = 1, \phi(6) = 1, \phi(7) = 1, \phi(8) = 1, \phi(9) = 1$; $b_{1T}^T = (1, 0, 0, 0), b_{2T}^T = (1, 2, 0), b_{3T}^T = (1, 1, 0), b_{4T}^T = (0, 1, 0), b_{5T}^T = b_{6T}^T = b_{7T}^T = b_{8T}^T = (0, 0, 1)$ and $b_{9T}^T = (0, 1, 1)$.

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

Fig. 4. Distributions of the estimates for $u_1, u_2, u_3$ and $u_4$ (top down). In each of the plots, white bars correspond to 1,000 families and black to 10,000 families (labels for the x-axis is multiplied by $10^4$).

Fig. 5. Distributions of likelihood ratio $L_{0,1}$ (top three), $L_{0,2}$ (center three) and $L_{0,3}$ (bottom three) for 200 left, 1,000 (center) and 10,000 (right) families (bottom), with smooth curves being the $\chi^2$ densities.
Figure 1
Figure 2
Figure 3
Figure 4
Figure 5
Table 1. Constraints during the germ-line development of a male *Drosophila melanogaster*

<table>
<thead>
<tr>
<th>Constraint #</th>
<th>Detail</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$N(8) \in [4, 6]$</td>
</tr>
<tr>
<td>2</td>
<td>$N(12) \in [23, 52]$</td>
</tr>
<tr>
<td>3</td>
<td>Population split into two with each $N \in [5, 9]$</td>
</tr>
<tr>
<td>4</td>
<td>Stem cell stage starts from the 15th division onward with $p_2 = 0.001$</td>
</tr>
<tr>
<td>5</td>
<td>Differentiated cells after the 31th division start spermatogenesis</td>
</tr>
</tbody>
</table>
Table 2. The expected numbers of mutation patterns and quality of approximations in 8,625 families, each having 20 offspring

<table>
<thead>
<tr>
<th>$u \times 10^4$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NP$_i$</td>
<td>8326.5</td>
<td>293.9</td>
<td>4.6</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>$D_i$</td>
<td>-0.1</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>$D'_i$</td>
<td>0.2</td>
<td>-0.1</td>
<td>-0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>5</td>
<td>NP$_i$</td>
<td>7232.6</td>
<td>1286.7</td>
<td>100.9</td>
<td>4.7</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>$D_i$</td>
<td>-1.7</td>
<td>1.4</td>
<td>0.3</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>$D'_i$</td>
<td>-1.7</td>
<td>14.1</td>
<td>-17.4</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>10</td>
<td>NP$_i$</td>
<td>6067.3</td>
<td>2179.3</td>
<td>344.1</td>
<td>32.2</td>
<td>2.0</td>
<td>0.1</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>$D_i$</td>
<td>-5.7</td>
<td>3.8</td>
<td>1.6</td>
<td>0.2</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>$D'_i$</td>
<td>-38.6</td>
<td>131.9</td>
<td>-127.5</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>50</td>
<td>NP$_i$</td>
<td>1504.0</td>
<td>2950.4</td>
<td>2468.7</td>
<td>1206.0</td>
<td>389.8</td>
<td>88.8</td>
<td>14.7</td>
</tr>
<tr>
<td></td>
<td>$D_i$</td>
<td>-31.3</td>
<td>-25.5</td>
<td>12.7</td>
<td>23.8</td>
<td>13.6</td>
<td>4.5</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>$D'_i$</td>
<td>-4901.0</td>
<td>11637.7</td>
<td>-8438.5</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

All entries are multiplied by 8,625. $u$: 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.
Table 3. Maximum likelihood estimates of $u_1, u_2, u_3$ and $u_4 \times 10^4$ with 20 offspring from each of $n$ families

<table>
<thead>
<tr>
<th>$u_1, ..., u_4$</th>
<th>$n=1000$</th>
<th>$n=10000$</th>
</tr>
</thead>
<tbody>
<tr>
<td>4,4,4,4</td>
<td>4.21, 3.86, 4.04, 3.98&lt;sup&gt;a&lt;/sup&gt;</td>
<td>3.95, 3.94, 4.02, 3.97</td>
</tr>
<tr>
<td></td>
<td>4.80, 2.57, 1.37, 1.51&lt;sup&gt;b&lt;/sup&gt;</td>
<td>1.80, 0.99, 0.48, 0.62</td>
</tr>
<tr>
<td>8,4,4,4</td>
<td>7.83, 4.11, 3.93, 4.04</td>
<td>7.95, 3.94, 4.02, 3.97</td>
</tr>
<tr>
<td></td>
<td>6.50, 3.02, 1.51, 1.53</td>
<td>2.35, 1.11, 0.53, 0.50</td>
</tr>
<tr>
<td>4,4,4,8</td>
<td>4.19, 3.83, 4.07, 7.91</td>
<td>3.97, 3.94, 4.02, 7.96</td>
</tr>
<tr>
<td></td>
<td>4.77, 2.62, 1.44, 1.72</td>
<td>1.84, 0.97, 0.50, 0.56</td>
</tr>
<tr>
<td>6,4,4,6</td>
<td>5.93, 3.97, 4.01, 5.97</td>
<td>5.99, 3.92, 4.04, 5.95</td>
</tr>
<tr>
<td></td>
<td>5.68, 2.82, 1.47, 1.62</td>
<td>2.11, 1.04, 0.52, 0.53</td>
</tr>
<tr>
<td>3,6,6,3</td>
<td>3.67, 5.58, 6.13, 2.96</td>
<td>3.03, 5.83, 6.06, 2.95</td>
</tr>
<tr>
<td></td>
<td>4.73, 3.00, 1.62, 1.67</td>
<td>1.75, 1.07, 0.56, 0.55</td>
</tr>
</tbody>
</table>

<sup>a</sup> mean estimates, <sup>b</sup> standard deviations. Result for each case is based on 1000 simulated data sets.
Table 4. Critical Values for likelihood ratio tests with $\mu = 0.0004$

<table>
<thead>
<tr>
<th>Asmpt</th>
<th>n=1000</th>
<th>n=5000</th>
<th>n=10000</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$c_5$</td>
<td>$c_1$</td>
<td>$c_5$</td>
</tr>
<tr>
<td>$L_{0,1}$</td>
<td>5.99</td>
<td>9.21</td>
<td>6.18</td>
</tr>
<tr>
<td>$L_{0,2}$</td>
<td>3.84</td>
<td>6.64</td>
<td>3.47</td>
</tr>
<tr>
<td>$L_{0,3}$</td>
<td>7.82</td>
<td>11.35</td>
<td>7.51</td>
</tr>
</tbody>
</table>

Notes: $c_5$ and $c_1$ are respectively the upper 5% and 1% critical values. Simulation results for each case are based on 10,000 replicates.
Table 5. Maximum likelihood estimates of $u \times 10^3$ under several hypotheses

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>$u_1$</th>
<th>$u_2$</th>
<th>$u_3$</th>
<th>$u_4$</th>
<th>$u_5$</th>
<th>$-\ln(L)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_1$</td>
<td>0.347</td>
<td>0.347</td>
<td>0.347</td>
<td>0.347</td>
<td>0.347</td>
<td>4519.0</td>
</tr>
<tr>
<td></td>
<td>0.343</td>
<td>0.343</td>
<td>0.343</td>
<td>0.343</td>
<td>0.343</td>
<td>4494.8</td>
</tr>
<tr>
<td></td>
<td>0.321</td>
<td>0.321</td>
<td>0.321</td>
<td>0.321</td>
<td>0.321</td>
<td>4441.9</td>
</tr>
<tr>
<td>$H_3$</td>
<td>2.284</td>
<td>2.284</td>
<td>0.001</td>
<td>0.001</td>
<td>1.217</td>
<td>4153.2</td>
</tr>
<tr>
<td></td>
<td>2.284</td>
<td>2.284</td>
<td>0.001</td>
<td>0.001</td>
<td>1.217</td>
<td>4129.0</td>
</tr>
<tr>
<td></td>
<td>2.249</td>
<td>2.249</td>
<td>0.001</td>
<td>0.046</td>
<td>1.050</td>
<td>4126.4</td>
</tr>
<tr>
<td>$H_5$</td>
<td>4.864</td>
<td>0.001</td>
<td>0.007</td>
<td>0.007</td>
<td>1.217</td>
<td>4139.6</td>
</tr>
<tr>
<td></td>
<td>4.864</td>
<td>0.001</td>
<td>0.007</td>
<td>0.007</td>
<td>1.217</td>
<td>4115.4</td>
</tr>
<tr>
<td></td>
<td>4.655</td>
<td>0.001</td>
<td>0.037</td>
<td>0.037</td>
<td>1.050</td>
<td>4113.1</td>
</tr>
<tr>
<td>$H_8$</td>
<td>5.072</td>
<td>0.001</td>
<td>0.002</td>
<td>0.007</td>
<td>1.217</td>
<td>4139.5</td>
</tr>
<tr>
<td></td>
<td>4.864</td>
<td>0.001</td>
<td>0.002</td>
<td>0.007</td>
<td>1.217</td>
<td>4115.4</td>
</tr>
<tr>
<td></td>
<td>4.815</td>
<td>0.001</td>
<td>0.001</td>
<td>0.058</td>
<td>1.032</td>
<td>4110.8</td>
</tr>
</tbody>
</table>

Notes: $H_1 : u_1 = ... = u_5$, $H_2 : u_2 = u_3 = u_4$, $H_3 : u_1 = u_2$, $H_4 : u_2 = u_3$, $H_5 : u_3 = u_4$, $H_6 : u_4 = u_5$, $H_7 : u_1 = u_5$ 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

\[
\begin{array}{cccccccc}
& & & & & & & \\
& i & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
\hline
H_1 vs H_i & 758.8 & 731.6 & 758.9 & 758.8 & 293.2 & 714.8 & 758.9 \\
& & 758.7 & 731.7 & 758.8 & 758.7 & 292.7 & 714.6 & 758.8 \\
& & 657.1 & 630.9 & 662.1 & 657.5 & 293.6 & 609.9 & 662.1 \\
H_i vs H_8 & 0.1 & 27.3 & 0.0 & 0.1 & 465.7 & 44.2 \\
& & 0.1 & 27.1 & 0.0 & 0.1 & 466.1 & 44.2 \\
& & 5.0 & 31.2 & 0.0 & 4.6 & 368.6 & 52.3 \\
\end{array}
\]
Table 7. Comparison of estimates of $u$ based on the ATE and AII when mutation rates are relatively high

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

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