# Highly variable recessive lethal or nearly lethal mutation rates during germ-line development of male *Drosophila melanogaster*

Jian-Jun Gao[a], Xue-Rong Pan[a], Jing Hu[a], Li Ma[a], Jian-Min Wu[a], Ye-Lin Shao[a], Sara A. Barton[b], Ronny C. Woodruff[c], Ya-Ping Zhang[a,d,1], and Yun-Xin Fu[a,b,1]

[a]Laboratory for Conservation and Utilization of Bioresources, Yunnan University, Kunming 650091, China; [b]Human Genetics Center and Division of Biostatistics, School of Public Health, University of Texas, Houston, TX 77030-3998; [c]Department of Biological Science, Bowling Green State University, Bowling Green, OH 43403-0208; and [d]State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China

Edited* by Wen-Hsiung Li, University of Chicago, Chicago, IL, and approved July 22, 2011 (received for review January 5, 2011)

**Each cell of higher organism adults is derived from a fertilized egg through a series of divisions, during which mutations can occur. Both the rate and timing of mutations can have profound impacts on both the individual and the population, because mutations that occur at early cell divisions will affect more tissues and are more likely to be transferred to the next generation. Using large-scale multigeneration screening experiments for recessive lethal or nearly lethal mutations of *Drosophila melanogaster* and recently developed statistical analysis, we show for male *D. melanogaster* that (*i*) mutation rates (for recessive lethal or nearly lethal) are highly variable during germ cell development; (*ii*) first cell cleavage has the highest mutation rate, which drops substantially in the second cleavage or the next few cleavages; (*iii*) the intermediate stages, after a few cleavages to right before spermatogenesis, have at least an order of magnitude smaller mutation rate; and (*iv*) spermatogenesis also harbors a fairly high mutation rate. Because germ-line lineage shares some (early) cell divisions with somatic cell lineage, the first conclusion is readily extended to a somatic cell lineage. It is conceivable that the first conclusion is true for most (if not all) higher organisms, whereas the other three conclusions are widely applicable, although the extent may differ from species to species. Therefore, conclusions or analyses that are based on equal mutation rates during development should be taken with caution. Furthermore, the statistical approach developed can be adopted for studying other organisms, including the human germ-line or somatic mutational patterns.**

within-host coalescent | mutation cluster | likelihood

Because mutations manifest their effect through cell descendants, it is essential to determine the timing and rate of mutations during individual development. However, little is known about this fundamental aspect of life even for well-studied model organisms. Germ-line mutations (i.e., mutations that occur in the lineage of germ cells) are of particular importance because only they are inherited, and thus may have a lasting effect on a population. The past few decades have witnessed tremendous advances in obtaining the rate of mutation per generation for genes for many organisms (1–3). There has been slow but steady progress in documenting and understanding the relationship between various human genetic disorders and parental ages (4–7), ever since nearly a century-old observation (8) that achondroplasia was more frequently found in children whose fathers were more advanced in age. In comparison, little is known about the details of mutation at different stages of germ cell development. Our knowledge from biochemistry, individual development, and observation of the frequencies of some human genetic disorders indicates that mutation rates at different stages may differ. Knowing the details of mutational distribution during germ cell development will not only improve the understanding of many genetic disorders but shed light on broader issues in mutation research, particularly in population/evolutionary biology. Dissecting the mutational distribution requires not only knowledge of the dynamics of germ cell lineage and high-resolution data but a proper integration of both. To date, available observations and experiments from humans have yet to lead to a breakthrough in this area, perhaps partly because of the complexity of human germ-line development, the difficulty in separating compounding factors in observations, and a lack of proper mathematical models to integrate the information.

Central to the dissection of the mutational pattern during germ-line development is to observe mutants in families that each has many offspring. Furthermore, different mutations leading to observable mutants in the same family need to be identified. *Drosophila* is one of the higher organisms that were first used to identify spontaneous and induced mutations (9, 10). The development of a germ cell lineage in *Drosophila melanogaster* has been continuously studied for the past 70 y. As a result, the dynamics of the germ cell population are well understood. This study takes advantage of well-established techniques from decades of *Drosophila* research to generate an unprecedented mutation dataset in a well-controlled environment. The mutation screening experiment we used led to cost-effective observations of the number of mutants and the frequency of each independent mutation (usually 1 or 2) in each of 8,618 families.

Also necessary to the understanding of mutational patterns is a proper statistical framework for inference. We developed a likelihood framework for analyzing such data, which can be described as follows. For each family, suppose that there are, at most, two mutations. Let $n_0$ be the number of families without any mutation; $n_i$ the number of families with one mutation of size $i$ ($i > 0$) (i.e., the number of mutants among offspring is $i$); and $n_{ij}$ the number of families with two mutations, one of size $i$ and one of size $j$. Then, the likelihood of the data is

$$L = \left(\prod_{i=0}^{n} p_i^{n_i}\right)\left(\prod_{ij} p_{ij}^{n_{ij}}\right), \quad [1]$$

where $p_0$ is the probability that there is no mutation in a family; $p_i$ is the probability that there is one mutation of size $i$; and $p_{ij}$ is the probability that there are two mutations, one of size $i$ and one of size $j$. To make inferences about mutation rates at various stages of germ-line development, it is necessary to express $p_i$ and $p_{ij}$ in terms of mutation rates at various stages. The germ cell divisions

Author contributions: J.-J.G., R.C.W., Y.-P.Z., and Y.-X.F. designed the experiment; R.C.W. contributed fly stocks; J.-J.G., X.-R.P., J.H., L.M., J.-M.W., Y.-L.S., and S.A.B. performed research; Y.-X.F. developed statistical methods and analyzed data; and Y.-X.F. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

Freely available online through the PNAS open access option.

[1]To whom correspondence may be addressed. E-mail: yunxin.fu@uth.tmc.edu or zhangyp@mail.kiz.ac.cn.

from a fertilized egg to sperm will be divided into $I$ intervals. Suppose the mutation rate per cell division for the $i$-th interval is $u_i$ and is defined as $\boldsymbol{u} = (u_1, \ldots, u_I)^T$. Ideally, $I$ is equal to the total number of cell divisions, such that the mutation rate at each cell division can be inferred; however, even with the large volume of data from our experiment, we still only have the resolution for a relatively small value of $I$. Nevertheless, tremendous insight into the rate variations can be learned. For the genealogy of a sample, each cell division corresponds to a segment of a branch. Let $t_k$ be the number of cell divisions from the $k$-th interval and $\boldsymbol{t} = (t_1, \ldots, t_I)^T$. Fig. 1 shows a hypothetical example of a sample genealogy of five cells with five cell divisions divided into three intervals (1: [1, 1], 2: [2, 4], and 3: [5, 5]), which results in $\boldsymbol{t} = (1, 9, 5)^T$. Assume that the number of mutations in a branch follows a Poisson distribution, with its parameter equal to the branch length times the mutation rate per cell division. Then, for the genealogy in Fig. 1, the probability of no mutation is $e^{-(1u_1+9u_2+5u_3)}$. In general, the number of mutations in a given genealogy is a Poisson variable with the parameter $\boldsymbol{t}^T\boldsymbol{u}$. Therefore, the probability that there is no mutation in a given genealogy is

$$e^{-\boldsymbol{t}^T\boldsymbol{u}} \approx 1 - \boldsymbol{t}^T\boldsymbol{u} + \frac{1}{2}\boldsymbol{u}^T\boldsymbol{A}_0\boldsymbol{u}, \qquad [2]$$

where $\boldsymbol{A}_0 = \boldsymbol{t}\boldsymbol{t}^T$. One does not generally know the sample genealogy; therefore, taking into consideration many possible genealogies for the sample, we have

$$p_0 \approx 1 - \bar{\boldsymbol{t}}^T\boldsymbol{u} + \frac{1}{2}\boldsymbol{u}^T\bar{\boldsymbol{A}}_0\boldsymbol{u}, \qquad [3]$$

where $\bar{\boldsymbol{t}}$ and $\bar{\boldsymbol{A}}_0$ are, respectively, the expected value of $\boldsymbol{t}$ and $\boldsymbol{A}_0$ over all possible sample genealogies, which can be estimated numerically (an example is given in *Materials and Methods*). Similar but more involved analysis leads to the expression of other required probabilities for maximum-likelihood analysis as ($i > 0$):

$$p_i \approx \bar{\boldsymbol{a}}_i^T\boldsymbol{u} - \boldsymbol{u}^T\bar{\boldsymbol{A}}_i\boldsymbol{u}, \qquad [4]$$

$$p_{ij} \approx \frac{2 - \delta_{i-j}}{2}\boldsymbol{u}^T\bar{\boldsymbol{A}}_{ij}\boldsymbol{u}, \qquad [5]$$

where $\bar{\boldsymbol{a}}_i, \bar{\boldsymbol{A}}_i$, and $\bar{\boldsymbol{A}}_{ij}$ are constant vectors and matrices that can be estimated similarly as $\bar{\boldsymbol{t}}$ and $\bar{\boldsymbol{A}}_0$. The likelihood function, together with these equations, allows for both the estimation and the hypothesis testing of $\boldsymbol{u}$ using the maximum-likelihood framework.

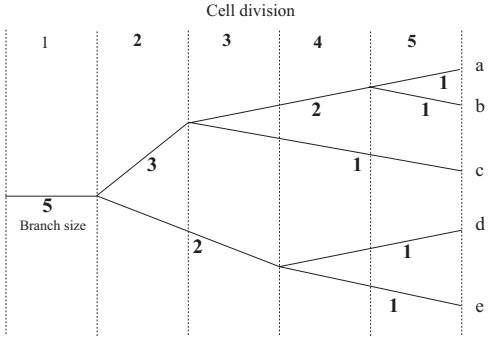


**Fig. 1.** Example genealogy of five alleles (*a*–*e*) sampled from the cell population after the fifth division. The value under a branch is the size of that branch (i.e., the number of descendant cells in the sample). Cell divisions are divided into three intervals, 1 [1, 1], 2 [2, 4], and 3 [5, 5], which lead to $\mathbf{t}^T = (1, 9, 5)$.

## Results

**Mutational Distribution.** A total of 8,618 families were successfully screened in our experiment over a 4-y period. Throughout the paper, a lethal or nearly lethal mutation is defined as one leading to no more than 1% of the surviving z/z offspring, which means that at least 100 offspring need to be examined for each claimed mutant. To minimize the chance that a mutant is not counted because of randomness, allelism tests were conducted for all lines with the percentage of z/z individuals up to 5%. Furthermore, to make the claim that two mutant lines share the same mutation, we required that among the offspring of the cross, the percentage of z/z individuals must also be no more than 1%. This stringent requirement will ensure a high quality for each identified cluster of mutants but has a slight tendency to lead to smaller cluster sizes than the true ones. Our plan was to screen 20 lines for each family; however, to ensure success, most families were screened for more than 20 lines. In our analyses, we randomly remove the extra lines in some families, such that each family has exactly 20 lines. We carried out analyses on several slightly different datasets derived as such. The results are virtually the same. Thus, we report one such analysis only. To make the framework of inference (Eq. **1**) applicable, we excluded several families with 3 or 4 mutations. Table 1 gives the frequencies of various mutation configurations. The distribution of families with various numbers of mutations can be derived from Table 1. From 8,618 families successfully screened, there were 954 harvested mutations, leading to a total of 1,036 different mutations. The number of families with 0, 1, and 2 mutations are, respectively, 7,664, 872, and 82. Among the 872 families with 1 mutation, 755 led to a singleton mutant. Roughly, the number of families with $i$ mutations is an order of magnitude smaller than that with $i - 1$ mutations. The number of mutants sharing the same mutation is said to be the size of that mutation or cluster size. Each of the mutations thus falls into a size between 1 and 20. The frequencies of various size mutations can also be derived from Table 1, and they are given in Table 2. Although a mutation predominantly leads to a singleton mutant, the mean size of the clusters is 2.03 (i.e., a mutation leads, on average, to 2.03 mutants in a family of 20 offspring).

**Statistical Inference.** The pattern of mutation rates along the germ cell lineage can be explored by dividing the germ cell development into intervals, such that estimates of the mutation rate, as well as the hypothesis test, can be made. For male *D. melanogaster*, each sperm from a young mature male is expected to have experienced 36 or more divisions, among which the first

**Table 1. Frequencies of mutation configurations among 8,618 families with 20 lines each**

| Mutation configuration | Frequency |
|---|---|
| (1) | 755 |
| (1, 1) | 48 |
| (2) | 50 |
| (3) | 13 |
| (18) | 11 |
| (2, 1), (16) | 8 |
| (17) | 6 |
| (15), (19) | 5 |
| (3, 1), (6), (20) | 4 |
| (4), (13) | 3 |
| (4, 1), (5), (12, 1), (13, 1), (17, 1) | 2 |
| (4, 3), (5, 2), (5, 4), (7), (7, 1), (9, 1), (9, 2), (10, 1), (11), (11, 1), (13, 3), (14), (14, 1), (14, 2), (14, 3), (15, 2), (16, 1) | 1 |
| Total | 954 |

Note: ($k$) denotes a family with one mutation only, which is of size $k$, and ($i$, $j$) denotes a family with two mutations, one of which is of size $i$ and one of which is size $j$.

**Table 2. Size distribution of 1,036 mutations**

| Size | Frequency | Size | Frequency | Size | Frequency | Size | Frequency |
|---|---|---|---|---|---|---|---|
| 1 | 877 | 6 | 4 | 11 | 2 | 16 | 9 |
| 2 | 62 | 7 | 2 | 12 | 2 | 17 | 8 |
| 3 | 20 | 8 | 0 | 13 | 6 | 18 | 11 |
| 4 | 7 | 9 | 2 | 14 | 4 | 19 | 5 |
| 5 | 4 | 10 | 1 | 15 | 6 | 20 | 4 |

14 divisions belong to the cleavage stage, the last 5 to spermatogenesis, and those between to gastrulation and organogenesis, in which the germ cells are known as germ-line stem cells that divide asymmetrically. In our analyses, we explored several ways to partition the divisions and found that the overall results are consistent. Therefore, we shall report the analysis based on one configuration that captures the essence of the results. The intervals are given in Table 3.

Before the likelihood analysis on mutation rates, coefficients in the expressions of $p_i$ and $p_{ij}$ need to be known, which can be accomplished by estimation using a Monte Carlo approach. Such estimation requires simulating genealogies that represent the sample from the sperm population in a male *D. melanogaster*, which can be accomplished through a two-step process. The first step is to simulate the dynamics of population size from a fertilized egg to the time spermatozoa are sampled. This step will take advantage of experimental evidence accumulated over decades of *Drosophila* research. The principle evidence (11–13) we used is as follows: (*i*) after the eighth cell division, about 4–6 cells become the primordial germ cells (PGCs); (*ii*) after the 12th division, the PGC number ranges from 23–52; (*iii*) after the 14th division, there are 5–9 PGCs in each gonad; and (*iv*) from the 15th division to right before spermatogenesis, the number of PGCs remains more or less constant. Once the first step is completed, we take a sample of 20 alleles from the population and use a coalescent-based approach (14) to simulate their ancestral process back to the fertilized egg. The process continues for many replicates from which the estimates of coefficients can be derived. Our experience indicates that 100,000 replicates are usually sufficient; however, to ensure a high accuracy, we obtained our estimates from 500,000 replicates.

Once the coefficients in the expression of $p$ are obtained, one can proceed to estimate $\boldsymbol{u}$. The maximum-likelihood estimates of $\boldsymbol{u}$ under eight different assumptions are given in Table 4. Under the assumption that mutation rates at different stages are equal, the common mutation rate per cell division is then equal to $0.345 \times 10^{-3}$ (corresponding to the row of $H_1$ in Table 4). On the other hand, when no constraint is imposed, $\boldsymbol{u}^T \times 10^3 = (5.043, 0.001, 0.001, 0.006, 1.225)$, which indicates the mutation rate at the first cell division is the highest, followed by the rate in the spermatogenesis stage. Note that the SEs associated with these two estimates are substantially smaller than the estimates themselves, indicating excellent quality of these estimates. Without any constraint on mutation rates, the per generation mutation rate is estimated to be $\hat{u} = 0.0125 \pm 0.001$ using Eq. **13**. In comparison, a prior method (15) gives $\tilde{u} = 0.0122 \pm 0.005$. Although the two estimates differ little, the maximum-likelihood estimate is superior because it carries a substantially smaller SE.

**Table 3. Intervals of cell divisions**

| Interval | Divisions | Developmental stage |
|---|---|---|
| 1 | [1, 1] | Cleavage |
| 2 | [2, 2] | Cleavage |
| 3 | [3, 14] | Cleavage |
| 4 | [15, 31] | Gastrulation and organogenesis |
| 5 | [32, 36] | Spermatogenesis |

**Table 4. Maximum-likelihood estimates of $u \times 10^3$**

| Hypothesis | $u_1$ | $u_2$ | $u_3$ | $u_4$ | $u_5$ | $-ln(L)$ |
|---|---|---|---|---|---|---|
| $H_1$ | 0.345 | 0.345 | 0.345 | 0.345 | 0.345 | 4,516.0 |
| $H_2$ | 4.919 | 0.006 | 0.006 | 0.006 | 1.225 | 4,140.4 |
| $H_3$ | 2.279 | 2.279 | 0.001 | 0.001 | 1.225 | 4,154.2 |
| $H_4$ | 5.043 | 0.001 | 0.001 | 0.008 | 1.225 | 4,140.3 |
| $H_5$ | 4.919 | 0.001 | 0.006 | 0.006 | 1.225 | 4,140.4 |
| $H_6$ | 4.345 | 0.001 | 0.001 | 0.411 | 0.411 | 4,370.4 |
| $H_7$ | 1.256 | 2.223 | 0.001 | 0.001 | 1.256 | 4,162.4 |
| $H_8$ | 5.043 | 0.001 | 0.001 | 0.008 | 1.225 | 4,140.3 |
| | (0.905) | (1.068) | (0.048) | (0.022) | (0.029) | |

$H_1$: $u_1 = \ldots = u_5$, $H_2$: $u_2 = u_3 = u_4$, $H_3$: $u_1 = u_2$, $H_4$: $u_2 = u_3$, $H_5$: $u_3 = u_4$, $H_6$: $u_4 = u_5$, $H_7$: $u_1 = u_5$, and $H_8$: no constraint. Values in parentheses are the SEs ($\times 10^3$) for estimates under $H_8$.

In addition to estimating the mutation rates, likelihood ratio tests of several hypotheses about $\boldsymbol{u}$ can be constructed from the values in Table 4, and their values are given in Table 5. For example, to test the null hypothesis that mutation rates over the development of the germ lines are all equal against the alternative that mutation rates at different intervals can all be different, we have the log-likelihood ratio statistic equal to $Lr = -2[ln(L_1) - ln(L_8)] = 751$; compared with the critical value of the $\chi^2$ distribution with 4 df, this result is highly significant. Table 5 shows that the hypothesis of constant mutation rates is overwhelmingly rejected with each of the seven alternative hypotheses. In comparison, hypotheses about two or more of $u_2$, $u_3$, and $u_4$ being equal cannot be rejected. We noted that the estimated mutation rate for the first cell division is considerably higher than that of spermatogenesis. Table 5 shows that the hypothesis of these two rates being equal is also rejected, as well as the hypothesis that the mutation rate in spermatogenesis is the same as that of the previous interval.

## Discussion

Our large-scale experiment was designed for detecting differences of mutation rates at various stages as small as half of the maximum rate. Taking advantage of the high resolution of the data, an inference framework that incorporates the knowledge of *Drosophila* development, and a rigorous statistical/computational method, we explored both the estimates of mutation rates at different stages of germ-line development as well as hypothesis testing. Our analyses show beyond a doubt that mutation rates vary significantly during the development of germ cells. Overall, the mutation rates in germ cell development exhibit a U shape, with the highest rate at the first cleavage, followed by a high rate at spermatogenesis, whereas most divisions in the middle have a mutation rate one or more orders of magnitude smaller. The likelihood ratio test of the null hypothesis of equal mutation rates for all cell divisions is overwhelmingly rejected with all the alternatives considered. Therefore, the notion of constant mutation rates during germ cell development should be abandoned.

Although it may not be too surprising to see a higher mutation rate at the first cleavage, it is unexpected in our analysis that the rate drops sharply from the second cleavage onward for the remaining divisions of cleavage. This is because the prior observation was that *Drosophila*-fertilized eggs divide about every 10 min in the cleavage stage and cellular activity is controlled by

**Table 5. Values of the log-likelihood ratio test of various hypotheses listed in Table 4**

| | i | | | | | | |
|---|---|---|---|---|---|---|---|
| | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| $H_1$ vs. $H_i$ | 751 | 724 | 751 | 751 | 291 | 707 | 751 |
| $H_i$ vs. $H_8$ | 0.1 | 27.7 | 0.0 | 0.1 | 460 | 44.2 | |

maternal proteins stored in the egg. Because the switch to zygotic control occurs at the end of the cleavage stage, one would probably expect that a significant change of mutation rate would occur around the end of the cleavage stage. To guard against incorrect assumption artifacts, we examined the consequences of alternative assumptions on the dynamics of germ cell lineage and on the outcome of the analysis, among which the assumption on the population after the eighth division appears to be most influential. It turns out that if one relaxes the range of the germ cells after the eighth division from 4–6 to 4–10, or restricts it to 2–4, and increases the total number of germ cell divisions from 36 to 40, the numerical results differ only slightly and all major conclusions remain the same. Our analysis also assumes that PGCs are formed by random sampling from the 256 cells after the eighth division. Although this is consistent with the *Drosophila* embryonic development literature (16), it is conceivable that some degree of nonrandomness leading to PGCs may exist because of spatial localization of closely related cells. The effect of nonrandom sampling can be investigated by restricting the germ cell population size at an earlier stage to be smaller than it normally should be (which is 4.72 ancestral cells at the 32-cell stage). Therefore, restricting the population size at the 32-cell stage to 4, 3, and 1–2 corresponds roughly to mild, modest, and severe sampling bias, respectively. For each of these restrictions, the same likelihood analysis was carried out. The likelihood under the assumption of random sampling has the largest value. For mild sampling bias, the log-likelihood value decreases slightly and all the conclusions made under the random sampling remain the same. For modest sampling bias, the estimated mutation rate at the first cleavage is larger than that for the second cleavage, but the difference is no longer significant. The log-likelihood value with modest bias is, however, significantly smaller than that of random sampling, such that the assumption of modest bias can be rejected at the 1% level. For severe sampling bias, the log-likelihood value decreases even more substantially. Taking the results of these additional analyses into consideration, we conclude that the mutation rate at the first cleavage is high. The rates drop sharply either immediately after the first division or in the next couple of cleavages, even with the possibility that sampling at the 258-cell stage may be biased to some extent (but extremely biased sampling is very unlikely).

Our study also indicates that the mutation rate at spermatogenesis is quite high, although significantly smaller than that of the first cleavage. There appears to be good reasons why this should be expected, because part of meiosis will weaken DNA repair mechanisms. Although our experiment screens for germline mutations of the male fly, sexual differentiation occurs late in development; thus, our conclusion of a high mutation rate for the first cleavage applies to the female fly as well. Per generation mutation rate is estimated to be 1.25%, which is comparable to previous estimates of completely recessive lethal mutations [1.2% in one study by Woodruff et al. (17) and 1.9% in another study by Woodruff et al. (18)].

Although making the experiment more manageable by examining only newly matured males, our experimental data do not allow one to address the potential rate changes during aging, which is an important aspect of mutation, particularly with regard to humans. Nevertheless, the results from this study have a number of implications. It is conceivable that the first conclusion stated in the abstract is true for most (if not all) higher organisms, whereas the other three conclusions are widely applicable, although the extent may differ from species to species. Therefore, conclusions or analyses that are based on equal mutation rates during development should be taken with caution. If overwhelmingly high mutation rates of the first cleavage (or first few cleavages) hold true, cells at the early stage of development will have accumulated a large number of mutations, which will then increase the opportunity for selection to act early. It will be of great interest to see if a similar mutation pattern holds for other organisms, particularly for humans. If so, it will be necessary to reevaluate some conclusions or approaches that have been based on assumptions of equal mutation rates. For example, the so-called "male-driven evolution" (19) can be better understood in light of the present work. It has been noted from various studies that the ratio of male to female cell divisions is often considerably larger than the ratio of estimated male to female mutation rates (20), which should be so if mutation rates in the first or first few cell divisions are two or more orders of magnitude larger than those in subsequent cell divisions.

Furthermore the statistical approach developed in this paper can be adopted for studying other organisms, including the human germ-line or somatic mutational patterns. For humans, different approaches will be needed to generate mutations, and advances in the next generation of sequencing technology will undoubtedly help to accelerate the study of mutational pattern in the development of humans.

## Materials and Methods

**Experiment.** The mutation screening experiment employs a three-generation assay to screen autosomal recessive lethal or nearly lethal mutations in about 1,200 genes in *D. melanogaster* (18), which takes advantage of the balancer chromosomes that were pioneered by H. J. Muller for the purpose of maintaining newly isolated mutations, including recessive lethals, without selection (21, 22). Balancers for each of the major chromosomes of *D. melanogaster* contain multiple inversions and one or more dominant visible mutations. The inversions, which are mapped by the use of giant polytene chromosomes, act as crossover suppressors, and the clearly visible dominant mutations allow for the identification of heterozygotes. With these chromosome stocks, new lethal or nearly lethal mutations are balanced in the heterozygous state against the balancer chromosomes and the new lethal is not lost over time by recombination. Three types of autosomal haploid chromosomes (genomes), denoted by β, γ, and z, were used in the experiment, and they are

$$\beta = \mathrm{T}(2;3)\mathrm{A1}-\mathrm{W},\ \mathit{Cy\ L\ Ubx}$$
$$\gamma = \mathrm{T}(2;3)\mathrm{B18},\mathit{Pm\ Sb}$$
$$\mathrm{z} = +;+$$

The β-type balancer is homozygous lethal and is marked with the dominant visible and recessive lethal mutations, including Curly (*Cy*) wings, Lobe (*L*) eye, and Ultrabithorax (*Ubx*) enlarged halteres. It segregates as a unit and suppresses crossing over on both the second and third chromosomes (23). The γ-chromosome is also homozygous lethal and carries dominant markers. Type z represents a haploid genome with WT second and third chromosomes.

The experiment was designed to screen β/z male offspring of crosses between a single β/z male and multiple β/γ females to see if a new lethal or nearly lethal mutation occurred in chromosome z during the germ-line development of the father. Therefore, each family consists of offspring from the following:

$$\text{multiple } \beta/\gamma \text{ virgin } ♀ \times \text{single } \beta/\mathrm{z}\ ♂$$

A total of 20–40 β/z ♂ offspring were each subjected to the following assay:

$F_1$: Multiple β/γ virgin ♀ × single β/z ♂
$F_2$: Multiple β/z virgin ♀ × multiple β/z ♂
$F_3$: Observe number of z/z individuals

If a β/z male in the $F_1$ step carries a lethal or nearly lethal mutation in the z chromosome, no surviving or few (≤1%) z/z individuals will be observed among the $F_3$ offspring. The number of genes in *D. melanogaster* that harbor recessive lethal mutations is estimated (24) to be around 3,000. When there was more than one mutant in a family, allelism tests were conducted to determine if they shared the same mutation. This is done by crossing β/z offspring from different mutant lines. If the offspring of the cross have no or only a few z/z individuals, the two mutant lines can be considered to share the same mutation. The experiment was carried out at Yunnan University from October 2004 to October 2008.

A similar mating scheme as described above was used successfully in earlier assays for the occurrence of mutation clusters in several laboratories (17, 18, 25–27). It was estimated that lethal or nearly lethal mutations identified by the assay span over about 1,200 genes.

**Statistical Inference.** The germ cell divisions from a fertilized egg to sperm can be divided into $I$ intervals. Suppose the mutation rate per cell division for the $i$-th interval is $u_i$, and define $\boldsymbol{u} = (u_1, \ldots, u_I)^T$. For the genealogy of a sample, each cell division corresponds to a segment of a branch. A branch is said to be size $i$ if it has exactly $i$ descendants in the sample. Let $a_{ik}$ be the total

number of cell divisions from interval $k$ that are of size $i$, $\mathbf{a}_i = (a_{i1}, \ldots, a_{il})^T$ and $\mathbf{t} = \sum_i \mathbf{a}i$. That is, $t_k$ is the number of cell divisions from the $k$-th interval. For the genealogy shown in Fig. 1, we have $\mathbf{a}_1 = (0, 4, 5)^T$ because the branch in the first interval is of size 5. There are four cell divisions in the second interval that are size 1 (2 in the branch leading to c as well as 1 to d and e each), and all cell divisions in the third interval are of size 1. Similarly, $\mathbf{a}_2 = (0, 4, 0)^T$, $\mathbf{a}_3 = (0, 1, 0)^T$, $\mathbf{a}_4 = (0, 0, 0)^T$, and $\mathbf{a}_5 = (1, 0, 0)^T$. Direct counting leads to $\mathbf{t} = (1, 9, 5)^T$, which can also be obtained by summing $\mathbf{a}_i$ ($i = 1, \ldots, 5$).

Suppose the number of mutations in a branch follows a Poisson distribution with its parameter equal to the branch length times the mutation rate per cell division. Then, given a genealogy, the number of mutations in the genealogy is also a Poisson variable with parameter $\mathbf{t}^T\mathbf{u}$. Therefore, the probability that there is no mutation in the genealogy is given by Eq. **3**. The probability that there is only one mutation of size $i$ in the genealogy is equal to

$$\left[e^{-(\mathbf{t}-\mathbf{a}_i)^T\mathbf{u}}\right]\left[e^{-\mathbf{a}_i^T\mathbf{u}}(\mathbf{a}_i^T\mathbf{u})\right] \approx \mathbf{a}_i^T\mathbf{u} - \mathbf{u}^T\mathbf{A}_i\mathbf{u}, \quad [6]$$

where $\mathbf{A}_i = \mathbf{a}_i\mathbf{t}^T$. The probability that there are only two mutations, one being of size $i$ and another of size $j (i \neq j)$, is equal to

$$\left[e^{-(\mathbf{t}-\mathbf{a}_i-\mathbf{a}_j)^T\mathbf{u}}\right]\left[e^{-\mathbf{a}_i^T\mathbf{u}}(\mathbf{a}_i^T\mathbf{u})\right]\left[e^{-\mathbf{a}_j^T\mathbf{u}}(\mathbf{a}_j^T\mathbf{u})\right] \approx \mathbf{u}^T\mathbf{A}_{ij}\mathbf{u} - (\mathbf{a}_i^T\mathbf{u})(\mathbf{a}_j^T\mathbf{u})(\mathbf{t}^T\mathbf{u}) \approx \mathbf{u}^T\mathbf{A}_{ij}\mathbf{u}, \quad [7]$$

where $\mathbf{A}_{ij} = \mathbf{a}_i\mathbf{a}_j^T$. Note that $\mathbf{A}_i = \sum_j \mathbf{A}_{ij}$ and $\mathbf{A}_0 = \sum_i \mathbf{A}_i$. If $i = j$, we have

$$\left[e^{-(\mathbf{t}-\mathbf{a}_i)^T\mathbf{u}}\right]\left[e^{-\mathbf{a}_i^T\mathbf{u}}\frac{(\mathbf{a}_i^T\mathbf{u})^2}{2}\right] \approx \frac{1}{2}\mathbf{u}^T\mathbf{A}_{ii}\mathbf{u}. \quad [8]$$

Because, apart from a few exceptions, all the families that have mutants in the experiment harbor either one or two mutations, we will not proceed further, although the approach can be extended to cover more complex situations.

Without knowing the sample genealogy, the probabilities in Eqs. **3**–**5** have to be integrated over all possible genealogies for a sample. Therefore, the probability $p_0$ is that there is no mutation; the probability $p_i$ is that there is one mutation of size $i$; and the probability $p_{ij}$ is that there are two mutations, one of size $i$ and one of size $j$. They are, respectively, as follows:

$$p_0 \approx 1 - \bar{\mathbf{t}}^T\mathbf{u} + \frac{1}{2}\mathbf{u}^T\bar{\mathbf{A}}_0\mathbf{u}, \quad [9]$$

$$p_i \approx \bar{\mathbf{a}}_i^T\mathbf{u} - \mathbf{u}^T\bar{\mathbf{A}}_i\mathbf{u}, \quad [10]$$

$$p_{ij} \approx \frac{2 - \delta_{i-j}}{2}\mathbf{u}^T\bar{\mathbf{A}}_{ij}\mathbf{u}, \quad [11]$$

where $\delta_x = 1$ when $x = 0$ and 0 otherwise and where $\bar{\mathbf{t}}, \bar{\mathbf{A}}_0, \bar{\mathbf{a}}_i, \bar{\mathbf{A}}_i$, and $\bar{\mathbf{A}}_{ij}$ are the means of the corresponding vector or matrix. The above result thus leads to Eqs. **3**–**5**. Maximum-likelihood estimates, $\hat{\mathbf{u}}$, of $\mathbf{u}$ can be derived from $\ln(L)$, which, from Eq. **1**, is

$$ln(L) = \sum_{i=0}^{n} n_i ln(p_i) + \sum_{ij} n_{ij} ln(p_{ij}). \quad [12]$$

From Eq. **12**, the asymptotic covariance $\mathbf{V}$ of the estimates $\hat{\mathbf{u}}$ can also be obtained as

$$\left\{-\left(\frac{\partial^2 ln(L)}{\partial u_k \partial u_l}\right)\Bigg|_{\mathbf{u}=\hat{\mathbf{u}}}\right\}^{-1}.$$

Let $\mathbf{c}^T = (c_1, \ldots, c_l)$, where $c_k$ is the number of cell divisions in the $k$-th interval. Then, per generation mutation rate, $u$, can be estimated as

$$\hat{u} = c_1\hat{u}_1 + c_2\hat{u}_2 + \ldots + c_l\hat{u}_l. \quad [13]$$

The variance of this estimate is $Var(\hat{u}) = \mathbf{c}^T\mathbf{V}\mathbf{c}$. 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 $\mathbf{u}$ is $\bar{u} = M/N$, which is unbiased regardless of whether the mutation rates during development are identical (15).

Hypotheses can be tested through the use of the likelihood ratio. For example, to test the null hypothesis $H_1$, that mutation rates at different cell divisions are all equal, against the alternative hypothesis $H_8$, that rates may all be different, the log-likelihood ratio test statistic is

$$Lr = -2(ln(L_1) - ln(L_8)), \quad [14]$$

which is asymptotically a $\chi^2$ variable with $l - 1$ df.

**Estimation of Coefficients and Simulation of Genealogy.** A key to the statistical inference described above is the mean values of various coefficients in Eqs. **3–5**, namely, $\bar{\mathbf{t}}, \bar{\mathbf{A}}_0, \bar{\mathbf{A}}_i, \bar{\mathbf{a}}_i (i = 1, \ldots, l)$, and $\bar{\mathbf{A}}_{ij}$. Because of their hierarchical relationship, only $\bar{\mathbf{a}}_i$ and $\bar{\mathbf{A}}_{ij}$ are fundamental. By definition, the $j$-th element of vector $\bar{\mathbf{a}}_i$ and the ($k$, $l$) cell of matrix $\bar{\mathbf{A}}_{ij}$ are, respectively,

$$\sum_g Pr(g)a_{ij}(g), \quad \sum_g Pr(g)a_{ik}(g)a_{jl}(g),$$

where summations are taken over all possible genealogies of the sample and Pr(g) is the probability of genealogy g. Although their analytical solutions are intractable, they can be estimated with sufficient accuracy by computer simulation, which takes into consideration the developmental knowledge of the male *D. melanogaster*. Specifically, suppose M genealogies of the sample are simulated; then, the above two quantities can be estimated, respectively, by

$$\frac{1}{M}\sum_k a_{ij}(g_k), \quad \frac{1}{M}\sum_k a_{ik}(g_k)a_{jl}(g_k).$$

Adopting the common practice in population genetics, we used a discrete generation model for the cells in the germ-line lineage, which assumes that the population at the $i$-th generation consists of cells that are potentially ancestral to the spermatozoa, each of which has divided $i$ times since the fertilized egg. Let $N(i)$ be the population size at the $i$-th generation. The model further assumes that each cell divides into two daughter cells, and the $(i + 1)$-th generation is formed by sampling from the pool of these daughter cells. Developmental knowledge is used to specify the sampling schemes, which will be illustrated by example. The genealogy of a sample of *D. melanogaster* male germ cells can be simulated by a two-step process.

The first step is to simulate the composition of $i$-th population. The $N(i)$ cells at the $i$-th generation can be divided into two groups, one $[N_2(i)]$ consisting of those that have siblings and another $[N_1(i)]$ consisting of those that do not have a sibling. The simulation can be done sequentially as follows. Starting with a fertilized egg (thus $N(0) = 1$ at the 0th generation), the first division yields 2 daughter cells. Both can potentially be ancestral to the sperm cells; thus, $N_1(1) = 0$, $N_2(1) = 2$. These two cells divide into 4 cells, which then form the second generation; continuing this process will lead to $N(7) = N_2(7) = 2^7 = 128$. Among the 256 daughter cells, only 4–6 are PGCs; thus, the eighth generation consists of cells that are a sample from these 256 cells. The main result shown in this paper assumes that the PGCs are a random sample from the 256 cells, but the algorithm can easily handle nonrandom sampling (the effects of nonrandom sampling are included in *Discussion*): first, randomly select a number between 4 and 6 (say 5), and then randomly select 5 cells of these 256 cells [and record the value of $N_2$ (8) and $N_1$ (8), which form the population at the eighth generation]. These 5 cells will then divide to form generation 9 and continues, and this leads to $N(11) = 40$ and 80 cells in their daughter pool. Because it is known that $N(12)$ is between 23 and 52, similar to the previous situation, a random number between 23 and 52 is determined and the corresponding number of cells is sampled from the pool to form the 12th population. After the 14th division, the population splits into two, each consisting of 5–9 cells and starts the stem cell period, which is characterized by asymptotic division. This can be modeled by assuming for each stem cell that one of its daughter cells at each division becomes a new stem cell, with a small probability (say 0.001) of being replaced by the second daughter cell of another stem cell. After the 31st division, the derived nonstem cells go into spermatogenesis, which results in spermatozoa. We modeled this by a simple model that assumes the cells after the 31st division resume symmetrical divisions and the last 5 divisions represent the process of spermatogenesis.

The second step in the simulation of the genealogy of a sample is the coalescent process with given populations sizes at each division from the first step. A sample of $n$ cells is taken from the 36th population, and their coalescence is determined backward in time. Consider $k$ random cells taken from the $i$-th population; the number of coalescent events is then equal to the pairs of sibling cells among these $k$ cells. For example, suppose $N(i)$ 10, $N_2(i) = 4$, and $k = 4$. Then, the probability of having two coalescents going back one generation is equal to

$$1\Big/\binom{10}{4} = 1/210, \quad 2\left[\binom{6}{2} + 2\binom{5}{1}\right]\Big/\binom{10}{4} = 5/21$$

for having one coalescent, and 159/210 for having no coalescent.

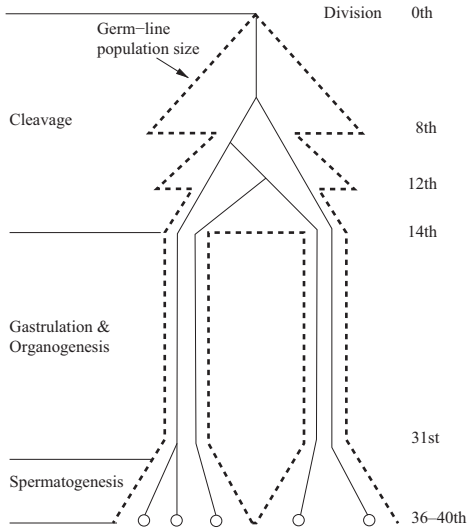


**Fig. 2.** Illustration of the relationship of developmental stages, dynamics of population sizes, and genealogy of a sample of five spermatozoa. The last five cell divisions represent spermatogenesis, which may start as early as the 32nd division.

The model of germ cell development, dynamics of the sizes of germ cell populations, and their relationship to the sample genealogy are illustrated by Fig. 2. Note that because only population sizes after each division are recorded in the first step, the genealogical relationship of the cells sampled in the second step is unknown and there are many plausible genealogies. One important feature of the sample genealogy is that it always traces back to the fertilized egg rather than stopping at the most recent common ancestor (MRCA); consequently, its height (from the time of sampling back to the fertilized egg) is a constant that is identical to the height of the germline lineage (36 divisions in our analysis). This is a marked difference from the genealogy of a sample in population genetics, where every sample may have a different age for its MRCA. Therefore, the meanings of the intervals of divisions remain the same regardless of whether one is referring to the history of the germ line or the sample genealogy.

Table 6 shows the estimates of $\bar{\boldsymbol{a}}_i^T$ for the interval divisions in the main text, with only the SEs for components of $\bar{\boldsymbol{t}}$ given because of space limitation. Because the first cell division leads to 2 cells, the second division to 4 cells, and so on, it follows that, on average, 1.889 cells of the two cells are present in sample genealogy, 2.842 of the 4 cells are present in the genealogy, and so on. The SEs of these estimates are equal to $0.317/\sqrt{500,000} = 0.317/707 = 4.5\times10^{-4}$ and $0.694/707 = 9.8\times10^{-4}$, respectively, which shows the high accuracy of estimations. In our final analysis, coefficients were estimated with at least 1 million simulated genealogies.

**Table 6. $\bar{a}_i$ and $\bar{t}$ estimated from 500,000 simulated genealogies of 20 alleles each for the division described in the main text**

| $k$ | $\bar{a}_{1k}$ | $\bar{a}_{2k}$ | $\bar{a}_{3k}$ | $\bar{a}_{4k}$ | $\bar{a}_{5k}$ |
|---|---|---|---|---|---|
| 1 | 0.040 | 0.172 | 23.330 | 82.543 | 81.151 |
| 2 | 0.060 | 0.233 | 20.061 | 60.635 | 7.620 |
| 3 | 0.074 | 0.261 | 12.846 | 28.176 | 1.017 |
| 4 | 0.088 | 0.273 | 7.697 | 9.260 | 0.123 |
| 5 | 0.099 | 0.269 | 4.895 | 2.286 | 0.012 |
| 6 | 0.108 | 0.255 | 3.345 | 0.440 | 0.001 |
| 7 | 0.114 | 0.234 | 2.347 | 0.067 | 0.000 |
| 8 | 0.120 | 0.210 | 1.640 | 0.008 | 0.000 |
| 9 | 0.123 | 0.184 | 1.117 | 0.001 | 0.000 |
| 10 | 0.125 | 0.159 | 0.747 | 0.000 | 0.000 |
| 11 | 0.123 | 0.134 | 0.487 | 0.000 | 0.000 |
| 12 | 0.120 | 0.113 | 0.307 | 0.000 | 0.000 |
| 13 | 0.114 | 0.092 | 0.191 | 0.000 | 0.000 |
| 14 | 0.108 | 0.073 | 0.115 | 0.000 | 0.000 |
| 15 | 0.099 | 0.057 | 0.070 | 0.000 | 0.000 |
| 16 | 0.088 | 0.043 | 0.040 | 0.000 | 0.000 |
| 17 | 0.074 | 0.031 | 0.023 | 0.000 | 0.000 |
| 18 | 0.060 | 0.021 | 0.012 | 0.000 | 0.000 |
| 19 | 0.040 | 0.013 | 0.006 | 0.000 | 0.000 |
| 20 | 0.111 | 0.016 | 0.004 | 0.000 | 0.000 |
| $\bar{t}^T$ | 1.889 | 2.842 | 79.278 | 183.418 | 89.924 |
| SE | (0.317) | (0.694) | (9.817) | (24.643) | (3.781) |

SE, standard error.

**ACKNOWLEDGMENTS.** We thank all who contributed to this project, particularly those who performed part of the experiment, including Ji-fen Li, Zhen Xie, Fang Chen, Jian-rui Zhou, Qun-li Wang, Lei-hua He, Qian-qian Zhao, Zong-jun Luo, Rui-lin Zhang, Ji Yao, Rui-hong Zhang, Xing-qin Yin, Xiao-qing Yang, Ji-qin Liu, Ji-xin Yang, Xing-yun Wang, Tian-fen Zhang, Yong-ping Meng, Qiu-qi Li, Yan-hong Du, Shu-li Xiong, Mei Zhang, Mei-lan Huo, Li-xian Lou, Xiao-yun Jiang, and Ya-ting Liu. This work was supported, in part, by grants from the Chinese National Science Foundation (Grant 30570248 to Y.-X.F., Grant 30460026 to J.-J.G., and Grant 30621092 to Y.-P.Z.) and the National Basic Research Program of China (973 Program, Grant 2007CB411600 to Y.-P.Z.), funds from the Bureau of Science and Technology of Yunnan Province of China (to Y.-P.Z.), and the Endowment Fund from the University of Texas (to Y.-X.F.).

1. Nei M (1987) *Molecular Evolutionary Genetics* (Columbia Univ Press, New York).
2. Li WH (1997) *Molecular Evolution* (Sinauer, Sunderland, MA).
3. Woodruff RC, Thompson JN, Jr. (1998) *Mutation and Evolution* (Kluwer, Dordrecht, The Netherlands).
4. Vogel F, Motulsky AG (1997) *Human Genetics: Problems and Approaches* (Springer, New York), 3rd Ed.
5. Glaser RL, Jabs EW (2004) Dear old dad. *Sci Aging Knowledge Environ* 3(re1):1–11.
6. Crow JF (2006) Age and sex effects on human mutation rates: An old problem with new complexities. *J Radiat Res* 47 (Suppl B):B75–B82.
7. Choi SK, Yoon SR, Calabrese P, Arnheim N (2008) A germ-line-selective advantage rather than an increased mutation rate can explain some unexpectedly common human disease mutations. *Proc Natl Acad Sci USA* 105:10143–10148.
8. Weinberg W (1912) Zur vererbung des zwergwches. *Arch Rassen Gesellschaftsbiol* 9: 710–718.
9. Muller HJ (1928) The measurement of gene mutation rate in *Drosophila*, its high variability, and its dependence upon temperature. *Genetics* 13:279–357.
10. Muller HJ, Oster II (1963) Some mutational techniques. *Drosophila. Methodology in Basic Genetics*, ed Burdette WJ (Holden–Day, San Francisco, CA), pp 249–278.
11. Drost JB, Lee WR (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.
12. Drost JB, Lee WR (1998) The developmental basis for germline mosaicism in mouse and *Drosophila melanogaster*. *Genetica* 102-103:421–443.
13. Gilbert SF (2003) *Developmental Biology* (Sinauer, Sunderland, MA), 7th Ed.
14. Ewens WJ (2004) *Mathematical Population Genetics* (Springer, New York).
15. Fu YX, Huai H (2003) Estimating mutation rate: How to count mutations? *Genetics* 164:797–805.
16. Sonnenblick BP (1965) *The Early Embryology of Drosophila melanogaster. Biology of Drosophila*, ed Demerec M (Hafner Publishing Company, New York), pp 62–167.
17. Woodruff RC, Thompson JN, Jr., Seeger MA, Spivey WE (1984) Variation in spontaneous mutation and repair in natural population lines of *Drosophila melanogaster*. *Heredity* 58:223–234.
18. Woodruff RC, Huai H, Thompson JN, Jr. (1996) Clusters of identical new mutation in the evolutionary landscape. *Genetica* 98:149–160.
19. Li WH, Yi S, Makova K (2002) Male-driven evolution. *Curr Opin Genet Dev* 12:650–656.
20. Crow JF (2000) A new study challenges the current belief of a high human male:female mutation ratio. *Trends Genet* 16:525–526.
21. Ashburner M (1989) *Drosophila: A Laboratory Handbook* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY).
22. Greenspan SF (1997) *Fly Pushing: The Theory and Practice of Drosophila Genetics* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY).
23. Lindsley DL, Zimm GG (1992) *The Genome of Drosophila melanogaster* (Academic, New York).
24. Abrahamson S, Würgler FE, DeJongh C, Meyer HU (1980) How many loci on the X-chromosome of *Drosophila melanogaster* can mutate to recessive lethals? *Environ Mutagen* 2:447–453.
25. Thompson JN, Jr., Woodruff RC (1980) Increased mutation in crosses between geographically separated strains of *Drosophila melanogaster*. *Proc Natl Acad Sci USA* 77: 1059–1062.
26. Mason JM, Valencia R, Woodruff RC, Zimmering S (1985) Genetic drift and seasonal variation in spontaneous mutation frequencies in *Drosophila*. *Environ Mutagen* 7: 663–676.
27. Brodberg RK, Mitchell MJ, Smith SL, Woodruff RC (1987) Specific reduction of N,N-dimethylnitrosamine mutagenicity in *Drosophila melanogaster* by dimethyl sulfoxide. *Environ Mol Mutagen* 10:425–432.