Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Theoretical Biology 223 (2003) 405–412
Journal of Theoretical Biology
www.elsevier.com/locate/jtbi

# Somatic mosaicism and cancer: inference based on a conditional Luria–Delbrück distribution

Steven A. Frank*

*Department of Ecology and Evolutionary Biology, University of California, Irvine, CA 92697-2525, USA*

Received 24 December 2002; received in revised form 26 February 2003; accepted 7 March 2003

**Abstract**

Somatic mosaicism for mutations in disease-causing genes has been reported in several recent studies. Examples include hemophilia A, many skin disorders, and several cancers such as retinoblastoma and familial adenomatous polyposis. Many of these disorders require multiple mutations in order to express the disease phenotype. For example, two recessive mutations to the retinoblastoma locus are required to initiate retinoblastomal tumors. I develop a mathematical framework for somatic mosaicism in which two recessive mutations cause disease. With my framework, I analyse the following question: Given an observed frequency of cells with two mutations and an easily scored aberrant phenotype, what is the conditional frequency distribution of cells carrying one mutation and therefore susceptible to transformation by a second mutation? This question is important because a high frequency of carrier cells can cause genetic counselors to misdiagnose a mosaic as an inherited heterozygote carrier and because widespread mosaicism can lead to some germline transmission. As more data accumulate, the observed distribution of mosaics can be compared against my predicted distribution. These sorts of studies will contribute to a broader understanding of the distribution of somatic mutations, a central topic in the study of cancer.


*Keywords:* Somatic mutation; Genetic counseling; Mendelian disorders

## 1. Introduction

An adult human has about $10^{14}$ cells and even more cell divisions in a lifetime. With a mutation rate of roughly $10^{-7}$–$10^{-6}$ per gene per cell division, each gene is mutated in many cells.

Although the number of mutations is large, most individuals have only a small fraction of their cells mutated for a particular gene. For neutral mutations, an average individual has a frequency of about $nu$ cells mutated for each gene, where $n$ is the number of cell divisions in the history of a cellular lineage from the zygote to the current time, and $u$ is the mutation rate in producing each daughter cell. With $n$ on the order of 50–100, a gene is on average mutated in approximately $10^{-5}$–$10^{-4}$ of the cells. There are about $10^4$–$10^5$ genes, so each cell may carry one or more mutations.

Each individual has much genetic mosaicism in the somatic tissue. The consequences of mosaicism may sometimes be limited because each gene is mutated in only a small fraction of cells, and those mutations may be recessive or in genes with limited effect in the mutated tissue. Mosaicism is, however, an important topic for three reasons.

First, a significant proportion of severe genetic disease arises from de novo mutations during development, causing mosaicism for normal and mutant cells. For example, in eight of 61 cases of hemophilia A, an X-linked disorder, the original mutation was traced to a mother or grandmother who had leukocyte mosaicism for the mutation (Leuer et al., 2001). The cases in which the original DNA change was a point mutation had eight of 32 (25%) mosaics. In the mosaic individuals, 0.2–25% of leukocytes carried the mutation, showing wide variability in the level of mosaicism associated with germline transmission. Several other studies have found significant frequencies of somatic mosaicism in genetic diseases, including retinoblastoma (Sippel et al., 1998), familial adenomatous polyposis (Farrington and Dunlop, 1999), and various skin disorders (Nomura et al., 2001; Paller, 2001). Accurate genetic counseling requires attention to mosaicism to assess the risks of familial transmission (Gottlieb et al., 2001).

*Tel.: +1-949-824-2244; fax: +1-949-824-2181.
*E-mail address:* safrank@uci.edu (S.A. Frank).

0022-5193/03/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/S0022-5193(03)00117-6

Second, different levels of mosaicism between tissues may provide clues about the phylogenetic history of cellular lineages within an individual. An observed frequency of mosaicism in one tissue could be used as conditional information to provide a better estimate of mosaicism in other tissues such as the germline. This would be useful for genetic counseling and to understand the shared cellular history between the progenitor cells that form different tissues.

Third, the probability of progression to cancer can depend on the degree of mosaicism in a tissue. For example, the development of retinoblastoma requires two loss-of-function mutations to the retinoblastoma gene (Knudson et al., 1975; Hethcote and Knudson, 1978). The probability that the second mutation occurs increases in proportion to the frequency of cells in the developing retina that carry the first mutation.

In this paper, I develop a mathematical framework for the analysis of mosaicism. Mathematical analysis is required to understand the distribution of mosaicism, to predict the probability of germline transmission given an observed degree of mosaicism, and to study the relations between mosaicism and cancer.

With my framework, I analyse the following question: Given an observed frequency of cells with two mutations and an easily scored aberrant phenotype, what is the conditional frequency distribution of cells carrying one mutation and therefore susceptible to transformation by a second mutation?

Most genetic mosaics are discovered because they have some cells with two mutations and noticeable aberrations. My conditional distribution predicts the probability distribution for the amount of mosaicism in such cases. As more data accumulate, the observed distribution of mosaics can be compared against my predicted distribution. These sorts of studies will contribute to a broader understanding of the distribution of somatic mutations, a central topic in the study of cancer.

## 2. Extensions to the Luria–Delbrück distribution

The frequency of cells carrying at least one mutation at a particular locus follows the classical Luria–Delbrück distribution (Zheng, 1999). This distribution would, for example, help to understand the frequency of recessive X-linked mutations among the somatic cells of a female. Females with somatic mutations also have a small chance that the mutation happened early enough in development to be contained in germ cells. Thus, I emphasize the stage in development at which mutations occur.

I consider the issue of two hits to a gene, leading to an extension of the Luria–Delbrück theory. For example, if an individual has a frequency $f$ of aberrant cells caused by two hits, then what is the conditional frequency of cells with at least one hit? I assume that all mutations are neutral with regard to birth and death rates. I use this simplifying assumption so that I can emphasize the structure of the problem. Under neutrality, I can develop some simple approximations that highlight key processes shaping the distributions of somatic mutations.

Cancer is certainly not a neutral phenotype. However, my analyses still provide insight into the distribution of mosaicism in cancer mutations. For example, in many cancers, the loss of function of one tumor-suppressor locus by mutation to both alleles is not sufficient to cause increased cell proliferation. It may be possible to assay apparently normal tissue for the frequency of cells with loss of function at a tumor-suppressor locus. The frequency of this phenotype can then be used to analyse the conditional distribution of cells carrying one recessive mutation at the tumor-suppressor locus.

Alternatively, a tumor-suppressor locus may transform a cell lineage into an expanding cancer clone after it suffers mutations to both alleles. The frequency of cells with two mutations then increases by selection, and the neutrality assumption no longer applies. Therefore, I focus on the time in cellular history at which the second of two mutations to a cell first occur. Assuming that the first mutation to a tumor-suppressor gene is neutral, the time of the second mutation is the same under models that assume neutrality or allow for selection. Estimating the time of the second mutation can be accomplished by measuring or making assumptions about the increase in cell proliferation in the expanding cancer lineage.

## 3. Frequency of mutant cells

I start with the classical Luria–Delbrück problem as background for my two-hit model and to motivate thinking about the distributions in terms of the timing of mutations in development. I also introduce a very simple approximation which, although rough, shows the processes involved in understanding the time of mutations in development.

I use a branching process model to describe the processes that shape the distribution of the number of mutant cells. For the gene of interest, let each cell carry either no mutations or at least one mutation. All cells with a mutation give rise to cells with a mutation, that is, there is no back mutation. Thus, one can describe each cell division as a branching in which a parent cell with no mutations produces 0, 1, or 2 daughter cells with no mutations, with probabilities $p_0$, $p_1$, and $p_2$, respectively. This allows me to track the probability of having $C$ cells with no mutations after $n$ rounds of cellular division. Starting with one cell and assuming two daughter cells in each round of division, the total number of cells is

$T = 2^n$, thus $1 - C/T$ is the frequency of cells with at least one mutation.

I calculate the probability distribution for $C$ by first writing

$$p_0 = u^2 \approx 0, \tag{1a}$$

$$p_1 = 2u(1-u) \approx 2u, \tag{1b}$$

$$p_2 = (1-u)^2 \approx 1 - 2u, \tag{1c}$$

where $u << 1$ is the probability of a mutation along each branch of the binary tree. The probability generating function for the branching process is

$$\phi(s) = p_0 + p_1 s + p_2 s^2. \tag{2}$$

The coefficient on the $j$-th power of $s$ gives the probability of $j$ descendants with no mutations, that is, the probability that $C = j$. The generating function can be expanded recursively, such that

$$\phi_{k+1}(s) = \phi_k(\phi(s)), \tag{3}$$

with $\phi_0 \equiv s$ and $\phi_1 \equiv \phi(s)$. Thus, $\phi_n$ provides an exact probability distribution for $C$ among the $2^n$ cells after $n$ cell divisions deriving from a single common ancestor. However, the generating function provides little insight and becomes time consuming to calculate as $n$ increases.

There are many technical papers on the Luria–Delbrück distribution (Zheng, 1999). Rather than expand on that technical work, I develop a simple approximation to clarify the processes that shape the distribution. I also focus on what I call the *effective time* of the first mutation, defined as $M = -\log_2(m)$, where $m$ is the frequency of mutant cells. If there is only one mutation, then $M$ is the round of cellular division at which the mutation occurred. If there is more than one mutation, then $M$ is the effective time of mutation.

An approximate distribution for the frequency of mutant cells can be built as follows. Start with a single, initial cell. In the $i$-th round of division, the number of new cells is $2^i$, the expected number of mutations is $\gamma_i = 2^i u$, and the probability of one mutation is $\gamma_i e^{-\gamma_i}$. A mutation in the $i$-th round of cell division leads to a mutation carried by at least $1/2^i$ of the total cells because the early mutation carries forward through time in $1/2^i$ of the lineages. The total number of branches in the cellular history is $2(2^n - 1) \approx 2^{n+1}$. If $2^{n+1}u << 1$, then nearly all histories have either no mutations or one mutation, and the probability that a frequency of $1/2^i$ of the cells is mutated is $\gamma_i e^{-\gamma_i}$ for $i = 1, \ldots, n$.

If $2^{n+1}u$ is greater than one, then cellular histories often have more than one mutation. Each round of cellular division contributes on average a frequency of $u$ mutations to the final total because the expected number of mutants in the $i$-th round is $2^i u$ and each mutation carries forward in $1/2^i$ lineages. Thus, $k$ rounds of cellular division add on average $ku$ to the frequency of mutants.

In the early rounds of cellular division, the accumulation of mutations is highly stochastic. However, once the expected number of mutations per round of cellular division rises to about two, one can add to the mutation frequency in a roughly deterministic way. Thus, if one separates between stochastic and deterministic accumulation at the round of division $x$ such that $2^x u \approx 2$ and $x \approx 1 - \log_2(u)$, then an approximate distribution is that a frequency of $m = 1/2^i + (1 - 1/2^i)(n - x - 1)u$ of the cells is mutated with probability $\gamma_i e^{-\gamma_i}$, with $i \geqslant 1$ increasing until the cumulative probability reaches one. Fig. 1 shows the cumulative probability distribution for $m$ as a function of the effective time of mutation, $M = -\log_2(m)$.

## 4. Frequency of mutant cells given two hits in some cells

I extend my analysis by studying the frequency $m$ of cells with at least one hit given an observation of two mutations in a frequency $f$ of cells. This is important because greater mosaicism for the first mutation increases the risk of multiple cancer origins and germline transmission. In the extreme case, the first mutation happens very early in development and the affected individual will often have symptoms similar to cases in which the defective allele was inherited. It is useful to know the probability that a case with hereditary symptoms is in fact a consequence of somatic mosaicism.

Consider a binary tree for cellular history with $2^n$ total cells after $n$ rounds of cellular division. Our problem concerns the fraction $m$ of the final cells that carry at least one mutation given that a frequency $f$ of the final cells that carry two mutations. Let $1 \leqslant y \leqslant n$ be the time step during which the earliest case of a second mutation occurs, and let $1 \leqslant x \leqslant y$ be the time step during which the first mutation occurs in the lineage that eventually has the earliest second mutation.

The first hit at time $x$ will cause at least $2^{-x}$ of the final cells to carry at least one mutation—this is the degree of mosaicism, $m \geqslant 2^{-x}$. If the second hit occurs at time $y$, then the frequency of cells with two hits will be $f \geqslant 2^{-y}$. To translate frequencies into effective times of mutation, I use $M = -\log_2(m)$ and $F = -\log_2(f)$. I define effective time intervals based on $x$ and $y$ such that $x - 1 < M \leqslant x$ and $y - 1 < F \leqslant y$, where $x$ and $y$ derive from the frequencies $m$ and $f$ rather than the actual times of the first and second hits.

I seek the conditional probability that $x - 1 < M \leqslant x$ given that $y - 1 < F \leqslant y$. This is roughly the same as finding the probability for the effective time of the first mutation, $x$, given the effective time of the second mutation, $y$, where $x \leqslant y$.

I present a branching process formulation in the appendix. The generating function from that

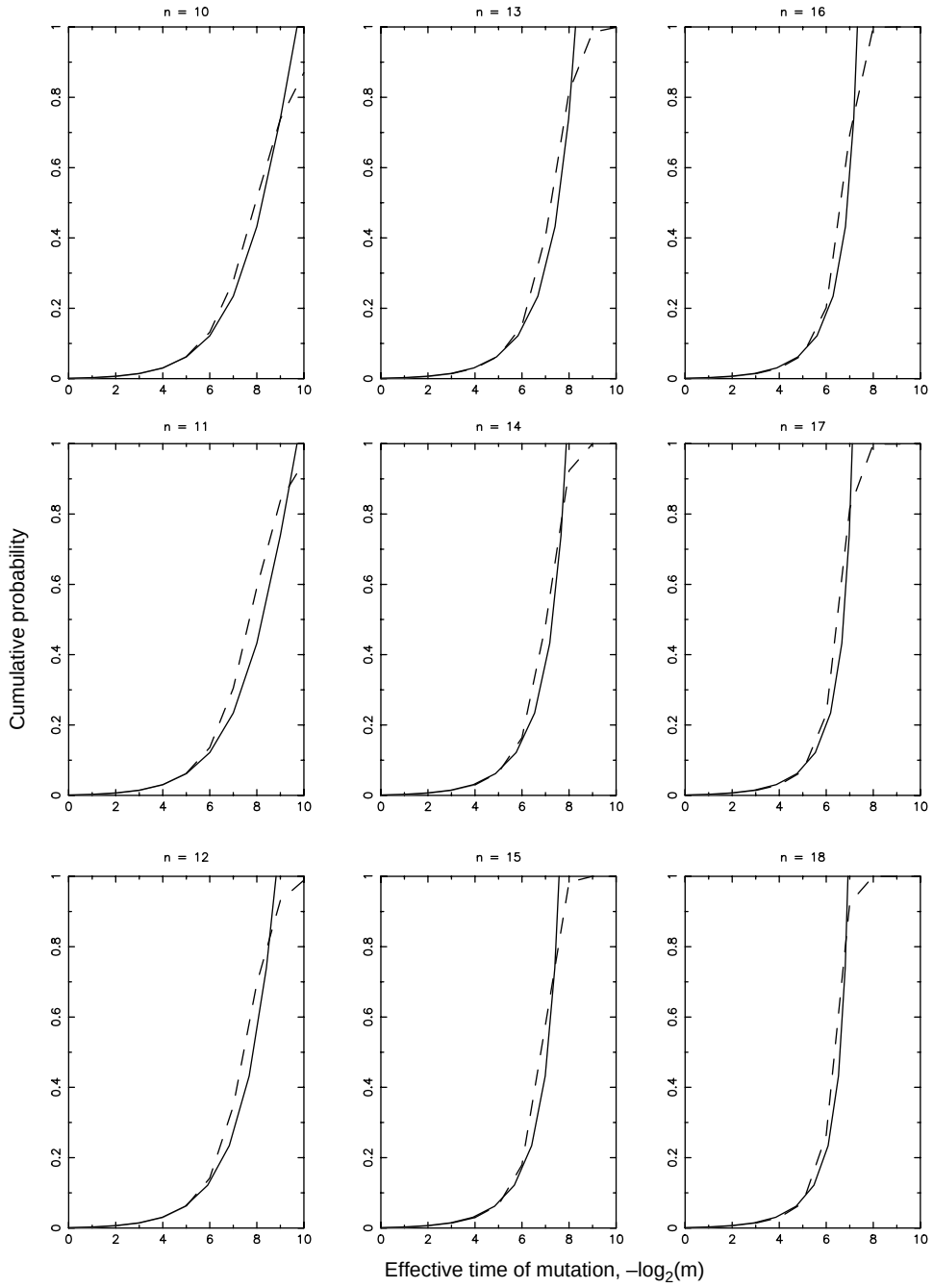


Fig. 1. Probability distribution of mosaicism for cells with at least one mutation in a particular gene. The $x$-axis is the frequency $m$ of cells with mutations, given as the effective time of mutation. The solid line shows the approximation given in the text. The dashed line shows the observed distribution in a Monte Carlo simulation of the branching process for 500 000 replications with $u = 10^{-3}$. I used a large mutation rate so that I could accumulate enough mutations in a reasonable amount of computer time. Each panel shows a different value of $n$, the total rounds of cellular division leading to $2^n$ cells. This is a form of the classical Luria–Delbrück distribution.

formulation gives the exact probability distribution function. However, the generating function becomes very time consuming to calculate as $n$ increases, and it provides little insight into the factors that influence the distribution.

I build up an approximation by estimating the probability, $p_x$, that the frequency, $m$, among the final $2^n$ cells with one or two hits is $2^{-x} \leqslant m < 2^{-x+1}$, given that the frequency, $f$, of cells with two hits is $2^{-y} \leqslant f < 2^{-y+1}$. I developed this approximation by listing factors that contribute to the probability of interest, and then dropping terms that did not significantly affect the fit of the approximation to computer simulations. I discuss only the significant terms.

Suppose that a second mutation occurred at time $y$. Then a first mutation must have occurred in the lineage leading up to the second mutation sometime during the $y$ cell divisions in that lineage. The mutation probability per daughter cell deriving from a parent with zero mutations is $u$ for $x<y$. For $x=y$, given that the second mutation occurs, the probability of the first mutation in that same cell is $u/2$. Thus, the approximate probability of a first mutation occurring at time step $x=1,\ldots,y$ is proportional to $\alpha(x)/2$, with $\alpha(z)=0$ for $z>y$, $\alpha(z)=1$ for $z=y$, $\alpha(z)=2$ for $z<y$.

This approximation must be adjusted to account for several additional factors. First, in the lineage descending from the first mutation, another mutation must not occur before $y$ or in two descendants at time $y$. The number of descendants in which there must be zero mutations is $2(2^{y-x}-1)$. Second mutations occur at rate $u/2$, so the expected number of mutations in this class of descendants is $\lambda_1=(u/2)2(2^{y-x}-1)$ and the probability of zero mutations is $P_0(\lambda_1)=e^{-\lambda_1}$, with $P_i(\lambda)$ as the Poisson probability of $i$ occurrences of a process with mean $\lambda$.

Second, no mutations can occur at any branch at time $t\leqslant x$ except the first mutation at time $x$ in the particular branch leading to the focal second mutation at time $y$. The probability of zero mutations of this sort is $P_0(\lambda_2)$, with $\lambda_2=u(2(2^x-1)-1)$.

Third, another alternative is for the mutation in the focal lineage to occur at any time $t>x$ and another mutation to occur at time $x$ in a lineage different from the focal lineage, which occurs with probability $P_1(\lambda_3)$ with $\lambda_3=u(2^x-1)$. One also has to deduct for the probability of no mutations before $x$ with $\lambda_4=u(2(2^x-1)-1)$, no second mutations in the lineage with the first mutation at $x$, $\lambda_5=(u/2)(2(2^{y-x}-1)+1)$, and no second mutations after $t$ and before the given second mutation at $y$, where such second mutations occur on average at $\lambda_6(t)=(u/2)(2(2^{y-t}-1))$.

Fourth, one has to add to $p_x$ those cases in which enough mutations happen at rounds of cell division greater than $x$ to cause $2^{-x}\leqslant m<2^{-x+1}$. I handle this by adding to $p_x$ all probabilities $p_z$ such that $z>x$ and $z\in Z$, where $Z$ is the set of indices that map to $x$. Elements in $Z$ are obtained by starting with those $z$ for which $2^z u>0.2$, which limits the mapping to those $z$ over which multiple mutations are reasonably probable. If $z$ satisfies that condition, then $z$ maps to $x=ceil[z-(z-k)/2]$, where $k=-\log_2(2^{-z}+(1-2^{-z})(n-z+1)u)$, and $ceil(a)$ rounds to the nearest integer greater than or equal to $a$. Here, $k$ accounts for the nearly deterministic accumulation of mutations after time $x$, in a way similar to my approximation for the Luria–Delbrück problem. The use of $x=ceil[z-(z-k)/2]$ causes the later mutations at $z$ to flow back one-half the distance toward the effective time, $k$. I chose the particular form of this flowback factor by fitting to distributions from Monte Carlo simulations.

Combining terms,

$$Sp_x=\alpha(x)P_0(\lambda_1)P_0(\lambda_2)+P_1(\lambda_3)P_0(\lambda_4+\lambda_5)\sum_{k=x+1}^{y}\alpha(k)P_0(\lambda_6(k))+\sum_{z\in Z}p_z$$

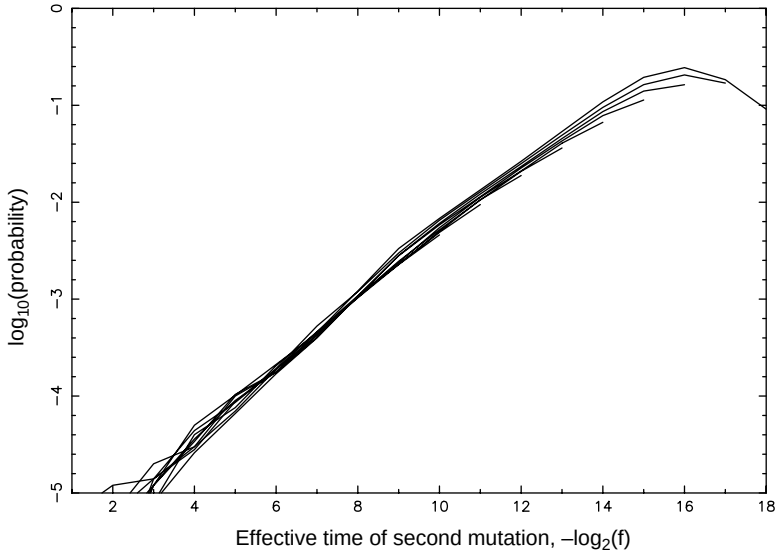


Fig. 2. Probability distribution of individuals having a frequency of cells $f$ with two mutations. The $x$-axis is given as the effective time of the second mutation, $-\log_2(f)$. This is the unconditional distribution, showing how often individuals carry a particular burden of doubly mutated cells. The plot shows results from Monte Carlo simulation of the branching process for 500 000 replications with $u=10^{-3}$. Each line shows a different value of $n=10,\ldots,18$.

for $x = 1, \ldots, y$, with $S = \sum p_x$ and the definition of $\alpha(z)$ given above.

Fig. 2 shows the unconditional probability distribution for individuals having a frequency $f$ of cells with two mutations. This shows how often phenotypic aberrations of a particular abundance would arise in individuals from two hits within a cellular lineage.

Fig. 3 shows the conditional distribution of individuals having a frequency of cells $m$ with at least one mutation given a frequency of cells with two mutations of $2^{-y} \leqslant f < 2^{-y+1}$ or, equivalently, an effective time of a second mutation of $y$. My simple approximation matches very well the observed conditional distributions.

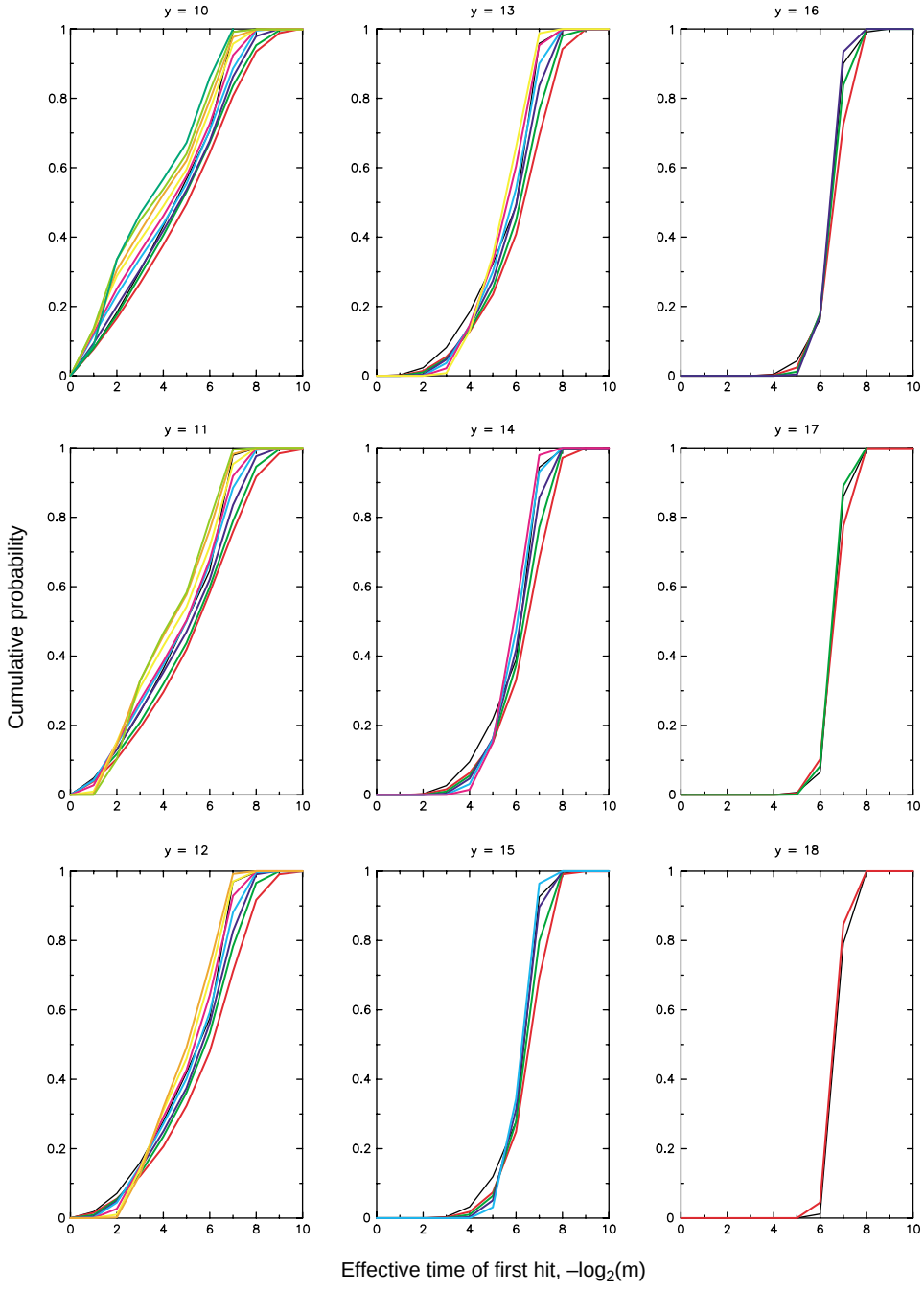


Fig. 3. Probability distribution of individuals having a frequency of cells $m$ with at least one mutation given a frequency of cells with two mutations of $2^{-y} \leqslant f < 2^{-y+1}$ or, equivalently, an effective time of a second mutation of $y$. Each panel shows the conditional distribution given a value of $y = 10, \ldots, 18$. Colored lines show results for $n = y, \ldots, 18$ from a Monte Carlo simulation of the branching process for 500 000 replications with $u = 10^{-3}$. Each black line is based on the theoretical approximation given in the text for the average value of $n$ in the panel, $n = (18 + y)/2$.

## 5. Discussion

The distribution of mutation frequency in a population contributes fundamentally to many aspects of population biology and evolution. The Luria–Delbrück distribution is the well-known description for the frequency of mutations at a single locus in haploid populations. I have extended the problem to consider the relations between the frequency of cells with two mutations and the frequency of cells with at least one mutation.

Consider, for example, a patient with retinoblastoma cancer. This cancer arises following mutations to both alleles at the retinoblastoma locus (Knudson, 1993). The developed retina has on the order of $10^6$ cells. If one assumes a mutation rate on the order of $10^{-6}$ or smaller, then the frequency of cells with two hits at the time of the second hit is at least $u$. Given a frequency of cells $u$ or greater with two hits at the time of the second hit, at what stage in retinal development did the first hit occur? Or, equivalently, assuming that the first hit occurred during retinal development, what frequency of normal retinal cells carries the first mutation and is at risk for transformation by a second mutation?

My analysis shows that, given a frequency of at least $u$ cells expressing a mutant recessive phenotype with the second hit occurring at the $y$-th round of cellular division, there is a nearly uniform probability that the first hit occurs at any time over the cellular generations $1, \ldots, z \leqslant y$. This can be seen in the upper left panel of Fig. 3, in which $y = 10$ and, to speed calculations, $u = 10^{-3}$. A second hit at $y = 10$ means that at least $1/2^y \approx u$ cells have two hits. The nearly linear cumulative distribution up to $z \approx 8$ shows an approximately uniform distribution over time for the occurrence of the first hit.

When the second hit occurs later, and the frequency of cells with two hits at the time of the second hit is much smaller than $u$, the effective time of the first hit becomes concentrated within one or a few rounds of cellular division. For $u = 10^{-3}$ and the second hit at $y = 18$, the effective time of the first hit occurs mostly at or just after the 7-th round of cellular division (Fig. 3, lower right panel). Two opposing forces concentrate the frequency of cells with one hit near $1/2^7$. First, earlier first mutations would almost certainly have second hits before $y = 18$. Second, later first hits are augmented in frequency by the nearly deterministic accumulation of first hits in other lineages at times $i$ for $2^i u > 1$. This can be seen in the Luria–Delbrück distribution for the unconditional times of the first hit (Fig. 1, lower right panel).

From these results and from simulations with $u = 10^{-4}$ (not shown), when the second hit occurs late, $1/2^y << u$, it appears that the effective time of the first hit concentrates near $i$ such that $2^i u \approx 0.1$. This matches the observed simulation values of $i \approx 7$ for $u = 10^{-3}$ and $i \approx 10$ for $u = 10^{-4}$, and predicts that $i \approx 17$ for $u = 10^{-6}$.

These various results show how the effective time of the second mutation provides information about the effective time of the first mutation. When the first mutation happens early in the cellular history, then the cellular population carries a high frequency of cells susceptible to transformation by a second mutation. If the cellular history under consideration goes back early into development, then different tissues will share a high frequency of cells carrying the mutation. The carrier frequency is particularly important when considering the probability that the germline has the mutation.

The degree to which different tissues carry the same early mutation also provides information about the cellular history of early development. This is particularly important in understanding when the germline differentiates, how many cells form the primordial germ tissue, and what is the cellular history of those primordial cells. These questions are, of course, empirical problems. But the empirical problems are greatly sharpened and more productively studied by understanding the quantitative relations between mutation and the history of cellular populations.

## Acknowledgements

National Science Foundation Grant DEB–0089741 and National Institutes of Health Grant AI24424 support my research.

## Appendix

In this appendix, I present a branching process model for the joint probability distribution of cells with one and two hits. The conditional distribution of cellular histories with $m$ cells carrying at least one mutation given $f$ cells with two mutations can be obtained from the joint distribution.

In this model, any previously unmutated allele mutates during transmission from parent cell to daughter cell with probability $u/2$. Each mutation event affects only one of the four alleles in the two diploid daughter cells. Thus, if a parent has two unmutated alleles, then it produces one mutant copy with a probability of approximately $4(u/2) = 2u$. If a parent has one mutated allele and one unmutated allele, then the second allele is mutated in one progeny with a probability of approximately $2(u/2) = u$.

One can track the number of cells with 0, 1, or 2 hits as $C_0$, $C_1$, and $C_2$ in a three-type branching process. Let $p_i(x, y, z)$ be the probability that an individual with $i = 0, 1, 2$ hits produces $x$ daughter cells with zero hits each,

$y$ daughter cells with one hit each, and $z$ daughter cells with two hits each. Every cell produces two progeny after division. The values for the transition probabilities are

$$p_0(2,0,0) = (1-u/2)^4 \approx 1 - 2u + 3u^2/2,$$

$$p_0(1,1,0) = 4(u/2)(1-u/2)^3 \approx 2u - 3u^2,$$

$$p_0(0,2,0) = 4(u/2)^2(1-u/2)^2 \approx u^2,$$

$$p_0(1,0,1) = 2(u/2)^2(1-u/2)^2 \approx u^2/2,$$

$$p_0(0,1,1) = 4(u/2)^3(1-u/2) \approx 0,$$

$$p_0(0,0,2) = (u/2)^4 \approx 0,$$

$$p_1(0,2,0) = (1-u/2)^2,$$

$$p_1(0,1,1) = 2(u/2)(1-u/2),$$

$$p_1(0,0,2) = (u/2)^2,$$

$$p_2(0,0,2) = 1.$$

The one-step probability generating functions for individuals with $i = 0, 1, 2$ hits are

$$\phi^{(0)}(s,t,w) = p_0(2,0,0)s^2 + p_0(1,1,0)st + p_0(0,2,0)t^2 + p_0(1,0,1)sw,$$

$$\phi^{(1)}(s,t,w) = p_1(0,2,0)t^2 + p_1(0,1,1)tw + p_1(0,0,2)w^2,$$

$$\phi^{(2)}(s,t,w) = p_2(0,0,2)w^2,$$

with $\phi_1^{(i)}(s,t,w) \equiv \phi^{(i)}(s,t,w)$ and, assuming that one starts at time zero with a single cell that has 0 hits, $\phi_0^{(0)} \equiv s$.

The generating function is built recursively for successive time steps as

$$\phi_{k+1}^{(0)}(s,t,w) = \phi_k^{(0)}[\phi^{(0)}(s,t,w), \phi^{(1)}(s,t,w), \phi^{(2)}(s,t,w)].$$

The coefficient for $s^{C_0}t^{C_1}w^{C_2}$ in $\phi_n^{(0)}(s,t,w)$ gives the probability that after $n$ rounds of cell division a lineage has $C_0$ cells with zero hits, $C_1$ cells with one hit, and $C_2$ cells with two hits. This generating function gives an exact probability distribution, but it becomes very time consuming to calculate as $n$ rises above 10.

## References


Farrington, S.M., Dunlop, M.G., 1999. Mosaicism and sporadic familial adenomatous polyposis. Am. J. Hum. Genet. 64, 653–658.

Gottlieb, B., Beitel, L.K., Trifiro, M.A., 2001. Somatic mosaicism and variable expressivity. Trends Genet. 17, 79–82.

Hethcote, H.W., Knudson, A.G., 1978. Model for the incidence of embryonal cancers: application to retinoblastoma. Proc. Natl Acad. Sci. USA 75, 2453–2457.

Knudson, A.G., 1993. Antioncogenes and human cancer. Proc. Natl Acad. Sci. USA 90, 10914–10921.

Knudson, A.G., Hethcote, H.W., Brown, B.W., 1975. Mutation and childhood cancer: a probabilistic model for the incidence of retinoblastoma. Proc. Natl Acad. Sci. USA 72, 5116–5120.

Leuer, M., Oldenburg, J., Lavergne, J.-M., Ludwig, M., Fregin, A., et al., 2001. Somatic mosaicism in hemophilia A: a fairly common event. Am. J. Hum. Genet. 69, 75–87.

Nomura, K., Umeki, K., Hatayama, I., Kuronuma, T., 2001. Phenotypic heterogeneity in bullous congenital ichthyosiform erythroderma. Arch. Dermatol. 137, 1192–1195.

Paller, A.S., 2001. Expanding our concepts of mosaic disorders of skin. Arch. Dermatol. 137, 1236–1238.

Sippel, K.C., Fraioli, R.E., Smith, G.D., Schalkoff, M.E., Sutherland, J., et al., 1998. Frequency of somatic and germ-line mosaicism in retinoblastoma: implications for genetic counseling. Am. J. Hum. Genet. 62, 610–619.

Zheng, Q., 1999. Progress of a half century in the study of the Luria–Delbrück distribution. Math. Biosci. 162, 1–32.