

Somatic evolutionary genomics: Mutations during development cause highly variable genetic mosaicism with risk of cancer and neurodegeneration

Steven A. Frank¹

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

Edited by Stephen Curtis Stearns, Yale University, New Haven, CT, and approved August 19, 2009 (received for review August 17, 2009)

Somatic mutations must happen often during development because of the large number of cell divisions to expand from a single-cell zygote to a full organism. A mutation in development carries forward to all descendant cells, causing genetic mosaicism. Widespread genetic mosaicism may influence diseases that derive from a few genetically altered cells, such as cancer. I show how to predict the expected amount of mosaicism and the variation in mosaicism between individuals. I then calculate the predicted risk of cancer derived from developmental mutations. The calculations show that a significant fraction of cancer in later life likely arises from developmental mutations in early life. In addition, much of the variation in the risk of cancer between individuals may arise from variation in the degree of genetic mosaicism set in early life. I also suggest that certain types of neurodegeneration, such as amyotrophic lateral sclerosis (ALS), may derive from a small focus of genetically altered cells. If so, then the risk of ALS would be influenced by developmental mutations and the consequent variation in genetic mosaicism. New technologies promise the ability to measure genetic mosaicism by sampling a large number of cellular genomes within an individual. The sampling of many genomes within an individual will eventually allow one to reconstruct the cell lineage history of genetic change in a single body. Somatic evolutionary genomics will follow from this technology, providing new insight into the origin and progression of disease with increasing age.

amyotrophic lateral sclerosis | Luria–Delbruck fluctuation analysis | somatic phylogenetics

A human develops from a single cell. From that single cell, an individual grows to 10^{13} to 10^{14} cells. That growth requires many cell divisions and, consequently, much somatic mutation must occur during development. Those developmental mutations likely have significant consequences for genetic mosaicism in the body and for the risk of cancer that arises from those mutations (1, 2).

In this article, I will show how to predict the amount of somatic mutation and the amount of genetic variability in the body. I will connect those calculations to the risk of cancer. I also propose that certain neurodegenerative diseases that occur later in life may often derive from early-life somatic mutations that occur during development.

Before turning to the details, let us consider in a general way the magnitude of somatic mutation during development in relation to the number of cells in the body. During development the single-cell zygote expands to $N = 10^{13}$ to 10^{14} cells. How many cell divisions occur during that expansion? Each time a cell divides, the number of cells in the body increases by one, assuming no cell death. So, to start with one cell and expand to N cells requires at least $N - 1$ cell divisions.

How much somatic mutation occurs during development? We do not have good measurements, but we can make some rough calculations. The minimum number of cells divisions is $N - 1 \approx N = 10^{13}$ to 10^{14} . Define the mutation rate per gene per cell division as u . No truly reliable estimates of somatic mutation

rates exist, but typically assumed values are of the order $u = 10^{-7}$ to 10^{-6} (3). The total number of mutational events per gene during development is the mutation rate per cell division multiplied by the number of cell divisions, $uN = 10^6$ to 10^8 . Thus, every gene in the genome mutates many times.

The value of uN measures the number of mutational events that occur in each gene. But most often, we will be interested in the number of cells that carry a mutation in a particular gene. For example, if a mutation occurs early in development, then that single mutation will carry forward to many descendant cells. By contrast, relatively few cells will carry a mutation that happens late in development. To understand the relation between the number of mutations that occur and the number of cells that carry a mutation, we must place somatic mutations in the context of cell lineage history. In other words, we must think of the body in relation to the lineage history descending from the single ancestral zygote and how mutations accumulate in that lineage history.

The accumulation of change within the lineage history of the body is somatic evolutionary genomics. With $\approx 10^{13}$ to 10^{14} cells in a body, and probably $>10^{16}$ cells produced over a lifetime, the lineage history within a single individual is much greater than for all of the hominids that have ever lived, perhaps as great as for all of the primates that have ever lived.

The tremendous evolutionary history within each human body has, until recently, been hidden by the difficulty of measuring genetic changes in cells. New high-throughput genomic technologies are just opening up the possibility of directly measuring somatic variability and evolution (4). To understand the evolutionary history of the individual and the consequences for disease, we must place somatic genomics within the context of the rate and pattern of evolutionary change in cellular lineages.

Genetic Mosaicism

In this section, I explain in more detail how mutations accumulate in cell lineages. I emphasize that the shape of lineage history differs at different times of life and in different tissues, affecting the patterns of somatic evolution. I also show how to predict the amount of genetic mosaicism in an individual and the variation in mosaicism between individuals. The following sections connect the amount and variation in genetic mosaicism to the risk of diseases such as cancer and neurodegeneration.

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Evolution in Health and Medicine" held April 2–3, 2009, at the National Academy of Sciences in Washington, DC. The complete program and audio files of most presentations are available on the NAS web site at www.nasonline.org/Sackler_Evolution_Health_Medicine.

Parts of this article were modified from earlier publications (1, 2).

Author contributions: S.A.F. performed research and wrote the paper.

The author declares no conflict of interest.

This article is a PNAS Direct Submission. S.C.S. is a guest editor invited by the Editorial Board.

¹E-mail: saf Frank@uci.edu

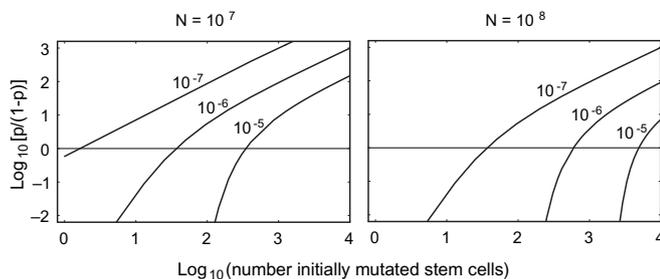


Fig. 3. Number of initially mutated stem cells at the end of development. The N initial stem cells derive by exponential growth from a single precursor cell. Each plot shows the cumulative probability, p , for the number of mutated initial stem cells. By plotting $\log_{10}[p/(1-p)]$, the zero line gives the median of the distribution. The number above each line is u_e , the mutation probability per cell added to the population during exponential growth. (I used an actual value of $10^{-5.2}$ rather than 10^{-5} because of computational limitations.) For a single gene, the mutation probability per gene per cell division, u_g , is probably $>10^{-7}$. If there are at least $L = 100$ genes for which initial mutations can influence the progression to cancer, then $u_e = Lu_g \geq 10^{-5}$. Initial mutations may, for example, occur in DNA repair genes, causing an elevated rate of mutation at other loci. Calculations were made with algorithms in Zheng (38). (Modified from figure 13.3 in ref. 2, based on the original in ref. 1.)

I have focused on mutations at a single locus. Mutations at many different loci may predispose to genetically influenced diseases such as cancer. Suppose mutations at L different loci can contribute to predisposition. We can get a rough idea of how multiple loci affect the process by simply adjusting the mutation rate per cell division to be a genome-wide rate of predisposing mutations, equal to $u_e L$. The number of loci that may affect predisposition may reasonably be around $L \approx 10^2$ and perhaps higher. Following the calculation in the previous paragraph, with $L \geq 10^2$, the number of initial stem cells carrying a predisposing mutation would on average be at least 10^5 . Some individuals might have two predisposing mutations in a single initial stem cell.

The average number of initially mutated cells tells only part of the story, because the distribution for the number of mutants is highly skewed. A few rare individuals have a great excess; in those individuals, a mutation arises early in development, and most of the stem cells would carry the mutation. Those individuals would have nearly the same risk as one who inherited the mutation.

Fig. 3 shows the distribution for the number of initially mutated stem cells at the end of development. For example, in Fig. 3 *Right*, with a mutation probability per cell division of 10^{-6} , a y-axis value of 2 means that $\approx 10^{-2}$, or 1%, of the population has $>10^4$ initially mutated stem cells at a particular locus ($L = 1$). Similarly, with a mutation probability per cell division of 10^{-7} , a y-axis value of 3 means that $\approx 10^{-3}$, or 0.1%, of the population has $>10^4$ initially mutated stem cells at a particular locus.

Amount and Variation of Genetic Mosaicism: Observations. My simple calculations show that, in a typical individual, every gene mutates somatically many times. Similarly, most cells in the body probably carry at least one somatic mutation. When we place this widespread somatic mosaicism into the context of cell lineage history, as in Fig. 2, we see that the body likely comprises variable-sized patches of somatic mosaicism throughout the genome.

Simple theory tells us that this widespread mosaicism must be present. But few measurements of mosaicism have been accomplished. The lack of measurement occurs because it is not easy to analyze genetic variation between individual cells in the huge population of cells that comprise a single body. I mentioned

earlier that the population of cells in a single body greatly exceeds that of all hominid individuals that have ever existed. A comprehensive study of the somatic evolutionary genomics of a single body would be as challenging as an evolutionary analysis of genetic variability for all humans and their hominid ancestors throughout hominid history.

A few analyses of mosaicism do exist (8). Mutation of a skin pigmentation gene in development causes skin cells to be marked by the mutation. The marked cells trace the tips of the somatic evolutionary lineage tree on the body surface. Interestingly, the patterned skin markings vary considerably. For example, several visible skin diseases follow the lines of Blaschko, which trace out what seem like contour lines or whorls over the skin surface (9–12). Other distinct patterns also occur in skin diseases (12). Speckled lentiginous naevus and Becker's naevus follow a mosaic checkerboard pattern; mosaic trisomy of chromosome 13 causes scattered leaf-like shapes of hypopigmentation.

Although somatic evolutionary genomics is difficult at present, genomic technology is advancing rapidly. The recent cancer genome project shows the potential for screening genetic changes in the somatic cells of an individual (4). Wallace's (13) work and future vision for somatic mitochondrial genetics emphasize the potential for analyzing the diseases of increasing age in the context of accumulating somatic mutations in cellular lineages.

The hematopoietic system provides a particularly promising somatic component for future study. Blood cells derived from diverse stem cell populations can easily be sampled and followed over time. Greaves (14) reviewed several lines of evidence demonstrating that developmental mutations in the hematopoietic system cause widespread somatic mosaicism. In many cases, those developmental mutations appear to be the primary cause of childhood leukemia (15–17).

Are the rare childhood forms of leukemia isolated examples, or is somatic mosaicism derived from developmental mutation a hidden risk factor in many cancers? At present, no direct evidence answers this question. The following section considers some theory by which we can predict the risk from developmental mutation. The theory also provides a framework for analyzing the data on somatic genomics that will become available in the future.

Risk of Cancer

Developmental mutations inevitably cause genetic mosaicism. Those cells carrying somatic mutations from development may be predisposed to cancer. In this section, I analyze the increased risk of cancer attributable to somatic mutations in early development.

Frank and Nowak's Model. Fig. 3 shows the probability distribution for the number of stem cells that have mutations. The number of mutated stem cells is $M = Nz$, where N is the total number of stem cells, and z is the frequency of those stem cells that carry a mutation that predisposes to disease. In this section, I focus on the average frequency of mutated stem cells, where x is the average of z , which leads to the average number of mutated stem cells, Nx . In the following section, I discuss the wide variation between individuals in the frequency of mutated stem cells. That variation in mutated stem cells may explain a significant fraction of the variation in cancer risk between individuals.

Mutations during the exponential phase of cellular growth in development cause the average frequency of stem cells with mutations to be $x \approx u_e \ln(N)$ (6), where u_e is the mutation rate during exponential cellular growth in development. Although the frequency of stem cells that start with a mutation may be small, those mutations can contribute substantially to the total risk of cancer. Suppose, by the multistage model of cancer progression, that k rate-limiting mutations are needed to cause

one or a few cells. However, Armon (25) suggested that the disease does start in a focal area and then spreads to neighboring motor neurons.

In ALS, a particular limb may show the first signs of motor neuron deterioration. The next symptoms frequently occur in the contralateral limb at the same spinal level and at contiguous spinal levels (26). Alternatively, spread may occur up and down the motor neuron system through the primary motor neurons directly connected to muscles and those neurons that feed into the primary motor system (27). Focal initiation followed by spread means that the disease may be initiated in one or a few cells, in the same way as cancer (25).

Developmental mutations and somatic mosaicism would play a role in ALS risk only if the initiating events in the focal cells derived from genetic (or epigenetic) changes in those cells. No direct evidence exists for a role of somatic mutation in initiating ALS. However, inherited germ-line mutations do strongly predispose to ALS, so it is possible that genetics plays a key role in disease (27–32). But the direct role of somatic mutation, and of what follows, remains a speculative hypothesis at present.

The most common genetic mutations associated with ALS lead to misfolded proteins (33–35). Individuals who inherit a predisposing mutation typically appear normal until disease onset usually in the age range of 25–65 years (30). As in inherited predisposition to cancer, other factors must be involved in transforming predisposed cells into the initiating focus of a disease that subsequently spreads.

Most cases seem to arise sporadically, without any evidence of inherited predisposition. Sporadic cases typically have a later age of onset in the range 40–80 years (27). The sporadic disease is mostly similar to the familial form, although some variations in the inherited form have been noted (27). This similarity between sporadic and inherited cases suggests that the sporadic cases may possibly begin from an initiating focus of cells with genetic mutations, but, again, there is no direct evidence for this.

If both inherited and sporadic cases do initiate disease from a small focus of cells that carry a genetic mutation, then, as in cancer, we can predict a continuum of risk. Those who inherit a predisposing mutation have the highest risk and earliest age of onset, with all cells carrying the mutation. Those who suffer a somatic mutation early in development have only slightly lower risk and later age of onset relative to inherited cases, with many cells carrying the mutation. Those who suffer a developmental mutation at some intermediate time in neuronal development have less risk and higher age of onset, with fewer cells, but still a significant number, carrying the predisposing mutation. Those who suffer a late developmental mutation have relatively low risk and late age of onset, with relatively few cells carrying the mutation. Finally, those with very few developmentally mutated cells may form the largest class with lowest risk and latest onset; those with very few developmental mutations may in fact have such low risk that nearly all cases can be ascribed to those who carry a significant number of developmental mutations.

If most cases derive from either inherited or developmental mutations, then the risk of ALS and perhaps other neurodegenerative diseases may be set very early in life. If sporadic cases often derive from developmental mutations, then somatic genomics will reveal an association between age of onset and the fraction of cells that carry a predisposing somatic mutation. The inherited cases would simply be the extreme of the risk continuum, in which all cells carry the predisposing mutation.

What sort of evidence would weigh in favor or against this speculative hypothesis? On the positive side: new somatic genomic measures that show a correlation between early-onset ALS and a somatic mutation widely distributed in neural tissue by descent from a single somatic mutation in a common ancestral cell; a mechanistic pathway that links the somatic mutation to the onset and progression of ALS; and further evidence that disease

spreads from a small focus of cells that have been transformed. On the negative side: no correlation between early-onset ALS and widely distributed somatic mutations known to predispose to ALS when those mutations arise in the germ line; and evidence that disease arises over a short period in many widely distributed locations rather than spreading from a small number of foci.

Potential Difficulties

The risks conferred by abundant low-penetrance genes affecting carcinogenesis may be of the same order as the risk from developmental mutations. How can we pick out the proportion of cancers that arise from developmental mutations if they are “drowned out” by population heterogeneity because of a number of weakly penetrant germ-line susceptibility genes?

We can distinguish between the risk caused by developmental mutations and weakly penetrant germ-line mutations by direct evidence from somatic genomics. Evidence favors developmental mutations when: one directly observes that somatic mutations dispersed widely in a tissue arose by normal cell lineage expansion in the development of that tissue; the dispersed somatic mutation derived from a single mutational event in an ancestral cell; and there is an association between those somatic mutations from normal developmental processes and the probability and age of cancer onset. These sorts of measurements are not easy at present, but the technology is developing very rapidly.

Another potential difficulty concerns the simple multistage model of carcinogenesis that I used in making calculations about the consequences of developmental mutations. I used the classical and most commonly cited form of the multistage model presented by Armitage and Doll (18) in 1954. As early as 1957, Armitage and Doll (36) pointed out that the age-specific incidence of certain cancers did not fit the simplest assumptions of their original multistage theory. There has followed a long history of model variants that fit particular assumptions to the age-specific incidence of particular cancers (2).

Recently, Meza et al. (37) made an important contribution to this long history by fitting explicit models of carcinogenesis to the age-specific incidence of colorectal and pancreatic cancers. They show that the classical model of Armitage and Doll (18) is not sufficient, and they develop a specific alternative that accounts for certain aspects of the biology and provides a much better fit to the data.

Given the limitations of the classical Armitage and Doll model that I used in the early sections of this article, how might my conclusions about developmental mutations be affected by alternative and more realistic models of carcinogenesis? The particular quantitative values derived from the model and shown in the figures would be changed. However, the model and the figures were meant only to illustrate the basic qualitative idea that developmental mutation potentially contributes to the risk of cancer and potentially alters the age of onset in those who suffer most strongly from widely dispersed developmental mutations. Those qualitative conclusions hold under any realistic model of carcinogenesis in which somatic mutations (or any heritable changes to cells) play a role.

My recent book (2) went into great detail about the various quantitative models of carcinogenesis in relation to the available data on age-specific incidence. Much can be learned by close study of those issues, and Meza et al. (37) have made a very significant contribution to that topic. But with regard to developmental mutations and somatic genomics, the real problems turn on the coming technological advances that will allow direct observation of how somatic mutations disperse through tissues in relation to cell lineage history and how those somatic mutations influence progression to cancer. Confidence in particular mathematical models must await those advances.

Conclusions

Skin diseases and childhood leukemia show that somatic mosaicism does occur. I have argued that those examples are just hints of the hidden and widespread mosaicism that arises from developmental mutations. Commonly accepted assumptions about cell division, cell lineage history, and somatic mutation lead inevitably to the prediction that mosaicism is common in all genes throughout the genome.

Importantly, the stochastic nature of mutation and the small number of cells in early development predict that the degree of mosaicism varies greatly between individuals. For any gene, a small fraction of individuals in the population will carry a somatic mutation in many cells, the mutation having occurred in early development. Those individuals will be at risk for diseases such as cancer and neurodegeneration that, later in life, can spread from a small focus of genetically predisposed cells.

Other individuals will carry relatively few predisposing somatic mutations derived from development. Those individuals have less risk later in life for diseases such as cancer and

neurodegeneration. If this argument is correct, then a significant fraction of risk for diseases later in life may derive from mutational events before birth.

Until recently, genetic technology has not allowed widespread sampling of somatic genomes within an individual. Thus, there is currently no direct evidence for or against this theory of widespread mosaicism and its association with the risk of disease. New technologies promise the ability to sample large numbers of genomes. The sampling of many genomes within an individual will eventually allow one to reconstruct the cell lineage history of genetic change in a single body. Somatic evolutionary genomics will follow from this technology, providing new insight into the origin and progression of disease with increasing age.

ACKNOWLEDGMENTS. I thank Susan Fitzpatrick for introducing me to the genetic basis of neurodegenerative disease and the potential analogy with cancer and Sharon Murphy for providing helpful comments and discussion, particularly with regard to childhood leukemia. My research is supported by National Science Foundation Grant EF-0822399, National Institute of General Medical Sciences MIDAS Program Grant U01-GM-76499, and a grant from the James S. McDonnell Foundation.

1. Frank SA, Nowak MA (2003) Developmental predisposition to cancer. *Nature* 422:494.
2. Frank SA (2007) *Dynamics of Cancer: Incidence, Inheritance, and Evolution* (Princeton Univ Press, Princeton).
3. Araten DJ, et al. (2005) A quantitative measurement of the human somatic mutation rate. *Cancer Res* 65:8111–8117.
4. Stratton MR, Campbell PJ, Futreal PA (2009) The cancer genome. *Nature* 458:719–724.
5. Luria SE, Delbruck M (1943) Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* 28:491–511.
6. Zheng Q (1999) Progress of a half-century in the study of the Luria–Delbruck distribution. *Math Biosci* 162:1–32.
7. Frank SA (2003) Somatic mosaicism and cancer: Inference based on a conditional Luria–Delbruck distribution. *J Theor Biol* 223:405–412.
8. Gottlieb B, Beitel LK, Trifiro MA (2001) Somatic mosaicism and variable expressivity. *Trends Genet* 17:79–82.
9. Happle R (1993) Mosaicism in human skin. Understanding the patterns and mechanisms. *Arch Dermatol* 129:1460–1470.
10. Siegel DH, Sybert VP (2006) Mosaicism in genetic skin disorders. *Pediatr Dermatol* 23:87–92.
11. Taibjee SM, Bennett DC, Moss C (2004) Abnormal pigmentation in hypomelanosis of Ito and pigmentary mosaicism: The role of pigmentary genes. *Br J Dermatol* 151:269–282.
12. Chuong CM, et al. (2006) What is the biological basis of pattern formation of skin lesions? *Exp Dermatol* 15:547–549.
13. Wallace DC (2005) A mitochondrial paradigm of metabolic and degenerative diseases, aging, and cancer: A dawn for evolutionary medicine. *Annu Rev Genet* 39:359–407.
14. Greaves M (2005) In utero origins of childhood leukemia. *Early Hum Dev* 81:123–129.
15. Greaves MF, Maia AT, Wiemels JL, Ford AM (2003) Leukemia in twins: Lessons in natural history. *Blood* 102:2321–2333.
16. Greaves MF, Wiemels J (2003) Origins of chromosome translocations in childhood leukemia. *Nat Rev Cancer* 3:639–649.
17. Mori H, et al. (2002) Chromosome translocations and covert leukemic clones are generated during normal fetal development. *Proc Natl Acad Sci USA* 99:8242–8247.
18. Armitage P, Doll R (1954) The age distribution of cancer and a multistage theory of carcinogenesis. *Br J Cancer* 8:1–12.
19. Hanahan D, Weinberg RA (2000) The hallmarks of cancer. *Cell* 100:57–70.
20. Fearon ER, Vogelstein B (1990) A genetic model for colorectal tumorigenesis. *Cell* 61:759–767.
21. Cairns J (1975) Mutation selection and the natural history of cancer. *Nature* 255:197–200.
22. Meza R, Luebeck EG, Moolgavkar SH (2005) Gestational mutations and carcinogenesis. *Math Biosci* 197:188–210.
23. Luebeck EG, Moolgavkar SH (2002) Multistage carcinogenesis and the incidence of colorectal cancer. *Proc Natl Acad Sci USA* 99:15095–15100.
24. Mitchell JD, Borasio GD (2007) Amyotrophic lateral sclerosis. *Lancet* 369:2031–2041.
25. Armon C (2005) Acquired nucleic acid changes may trigger sporadic amyotrophic lateral sclerosis. *Muscle Nerve* 32:373–377.
26. Carosio JT, Mulvihill MN, Sterling R, Abrams B (1987) Amyotrophic lateral sclerosis. Its natural history. *Neurol Clin* 5:1–8.
27. Armon C (2003) Epidemiology of ALS/MND. *Motor Neuron Disorders*, eds Shaw P, Strong M (Elsevier, Amsterdam), pp 167–206.
28. Kabashi E, et al. (2008) TARDBP mutations in individuals with sporadic and familial amyotrophic lateral sclerosis. *Nat Genet* 40:572–574.
29. Sreedharan J, et al. (2008) TDP-43 mutations in familial and sporadic amyotrophic lateral sclerosis. *Science* 319:1668–1672.
30. Strong MJ, Hudson AJ, Alvord WG (1991) Familial amyotrophic lateral sclerosis, 1850–1989: A statistical analysis of the world literature. *Can J Neurol Sci* 18:45–58.
31. Valdmanis PN, Rouleau GA (2008) Genetics of familial amyotrophic lateral sclerosis. *Neurology* 70:144–152.
32. Van Deerlin VM, et al. (2008) TARDBP mutations in amyotrophic lateral sclerosis with TDP-43 neuropathology: A genetic and histopathological analysis. *Lancet Neurol* 7:409–416.
33. Kabashi E, Durham HD (2006) Failure of protein quality control in amyotrophic lateral sclerosis. *Biochim Biophys Acta* 1762:1038–1050.
34. Meiering EM (2008) The threat of instability: Neurodegeneration predicted by protein destabilization and aggregation propensity. *PLoS Biol* 6:e193.
35. Wang Q, Johnson JL, Agar NY, Agar JN (2008) Protein aggregation and protein instability govern familial amyotrophic lateral sclerosis patient survival. *PLoS Biol* 6:e170.
36. Armitage P, Doll R (1957) A two-stage theory of carcinogenesis in relation to the age distribution of human cancer. *Br J Cancer* 11:161–169.
37. Meza R, Jeon J, Moolgavkar SH, Luebeck EG (2008) Age-specific incidence of cancer: Phases, transitions, and biological implications. *Proc Natl Acad Sci USA* 105:16284–16289.
38. Zheng Q (2005) New algorithms for Luria–Delbruck fluctuation analysis. *Math Biosci* 196:198–214.