Commentary

Methylation patterns and mathematical models reveal dynamics of stem cell turnover in the human colon

Simon Ro and Bruce Rannala*

Department of Medical Genetics, University of Alberta, Edmonton, AB, Canada T6G 2H7

To understand the normal aging process, as well as the role of cellular aging in diseases such as cancer, it is essential to understand the process of somatic cell development and renewal. Stem cells are undifferentiated cells, normally residing in a specific location (a niche) within a tissue. Stem cells are capable of producing a variety of somatic cell types needed for periodic tissue renewal and tissue regeneration after injury. To accomplish this, stem cells produce intermediate progenitors, called transit amplifying (TA) cells, that can divide rapidly and differentiate into various types of tissue cells. Because stem cells are the only cells capable of continuous tissue renewal, the population of stem cells must be maintained. It is still largely a mystery how stem cells maintain their numbers. Two competing models have been proposed (1, 2) (Fig. 1). The deterministic model proposes that a small number of stem cells reside in a niche, each generating exactly one stem cell and one TA cell at each (asymmetrical) cell division. The daughter TA cell leaves the niche to proliferate for tissue renewal while the daughter stem cell remains in the niche; each stem cell is "immortal" under this model. The stochastic model proposes that many stem cells exist in a niche with each stem cell division producing either two, one, or zero stem cells (and either zero, one, or two proliferating TA cells, respectively). This leads to "drift" in the numbers of descendents of each stem cell lineage over time. Eventually, a single common ancestral stem cell exists from which all stem cells in a niche are descended. The stem cell population is most likely to persist (under the stochastic model) if the probability that a stem cell division produces either two stem cells, or zero stem cells, is equal. In this issue of PNAS, Yatabe *et al.* (3) demonstrate how random changes in methylation patterns at CpG sites within particular genes can be used to study the dynamics of stem cells in human colon crypts. Methylation refers to the covalent addition of a methyl group to a DNA residue; in the mammalian genome, the addition of a methyl group to the fifth position of the cytosine ring within a CpG dinucleotide is most common (4, 5). Yatabe *et al.* (3) use their method to test deterministic and stochastic models of stem cell division and conclude that a stochastic model is more consistent with their data.

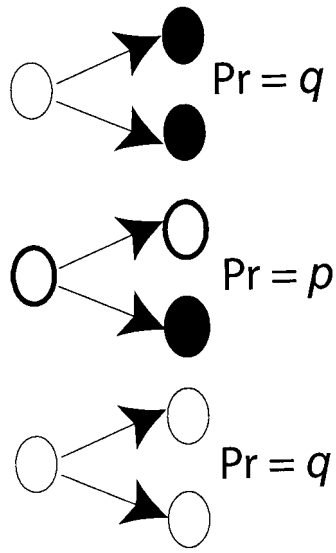

Fig. 1. Deterministic and stochastic models of stem cell population dynamics. Open circles represent stem cells and filled circles represent TA cells. The probability that either two stem cells (and no TA cells), or two TA cells (and no stem cells) result from the division of a stem cell is q. The probability that one stem cell and one TA cell result is P. The deterministic model assumes that $q = 0$ and $P = 1$ and the stochastic model assumes that $P < 1$ and $q > 0$.

Given the importance of cellular population dynamics to a complete understanding of stem cell biology, it is no surprise that cell fate mapping has become one of the hottest areas of stem cell research. Cell fate mapping methods used in mice typically involve following the fate of cells labeled with either radioactive, or histochemical tags, or following the fate of cells in chimeric mice (6–8). Alternatively, initial variability can be induced at genetic markers through mutagenesis; a subsequent loss of variation at these genetic markers (over time and space) provides information about the dynamics of cell turnover (9). Neither of these approaches can be used to study stem cell turnover in humans and it is difficult to generalize results from mice to humans because of differences in lifespan (number of cell divisions) and the larger numbers of cells that make up organs and tissues in humans. The methylation-based strategy of Yatabe *et al.* (3) provides a practical approach for studying stem cell dynamics in humans. The basic idea is that epigenetic variants with different patterns of methylation at CpG sites arise during stem cell division and the distribution of the variants among, and within, tissue regions conveys information about stem cell population dynamics.

It is well known that changes in levels of DNA methylation modify gene expression and chromatin structure (10, 11). These changes are regulated by specific enzymes and are of functional significance. However, if one considers changes in the methylation patterns of CpG sites in genes that are not expressed in cells of a particular tissue, these changes are likely to have no functional significance for cells of that tissue. The changes in methylation patterns, in this case, are not caused by gene regulation but instead arise by a random process of methylation associated with cellular aging, much like DNA mutations, but occurring at a higher rate. The patterns of methylation at particular sites in a nonexpressed gene are therefore "neutral" genetic markers that can be used to study cell fates. The basic idea is that the methylation of CpG sites in these genes occurs over time according to a stochastic process with a rate (roughly 10^{-5}) that is much higher than the rate of DNA mutation (roughly 10^{-9}).

Yatabe *et al.* (3) assume that all of the CpG sites they examine are initially unmethylated (at birth) and that sites be-

See companion article on page 10839.

*To whom reprint requests should be addressed at: Department of Medical Genetics, University of Alberta, 8–39 Medical Sciences Building, Edmonton, AB, Canada T6G 2H7. E-mail: brannala@ualberta.ca.

come methylated (at random) as cells divide over time. Evidence suggests that there is somatic inheritance of methylated sites (12). The combinations of methylated CpG sites, within each gene (referred to as tags), that accumulate within cellular lineages over time carry information about the genealogical relationships among these lineages (at the cellular level) and allow different models of stem cell turnover to be tested (by comparing predicted patterns of methylation under mathematical models of cell division, and random methylation, with observed patterns). If a target gene has N CpG sites (methylation tags) that are assayed, then each site is a binary marker (unmethylated = 0, methylated = 1). Because each site must assume one of these two states, there are 2^N possible distinct methylation patterns of the gene in any cell, yielding a large number of unique tags. For example, if $N = 8$ CpG sites occur within a gene, there are 256 unique tags. Three examples of unique tags are 00000001, 00110010, and 11101110.

Techniques for determining whether a cytosine nucleotide (C) is methylated, or unmethylated, are well established (13). Treatment of DNA with bisulfite converts unmethylated C into uracil (U), but methylated C is unaltered (14). PCR amplification and sequencing then reveals the site-specific pattern of methylation (i.e., C versus T) in a bisulfite-treated sequence. Yatabe *et al.* (3) studied the methylation status of five, eight, and nine CpG sites, respectively, in the *MYOD1*, *CSX*, and *BGN* genes (none of which are expressed in colon crypt cells). Between seven and nine normal colon crypts were obtained from each of 14 patients having undergone colectomies. Samples of at least five clones (molecules) from each crypt were analyzed to examine methylation patterns in the three genes. This allowed diversity of methylation patterns to be examined both within, and between, colon crypts. Colon crypts are straight tubular glands, which mainly consist of goblet (mucous) cells that continuously secrete mucin to lubricate the bowel, facilitating the passage of the colonic contents. A single crypt is thought to be a proliferative unit, which is renewed about every 6 days by a stem cell niche located at the bottom of a crypt. The lower third of the crypt constitutes the proliferative zone where newly generated TA cells undergo 2–3 additional divisions as they begin their migration up the crypt (15, 16).

Yatabe *et al.* (3) used several statistics to summarize the methylation patterns within and among crypts. One measure used in their analyses is the epigenetic distance between a pair of molecules. This is calculated (for a particular gene) as the total number of CpG sites at which the two molecules differ. For example, if an analysis of *MYOD1* for two molecules sampled from within a single crypt yielded 01001 and 00101, then the epigenetic distance is 2. Epigenetic distances among pairs of molecules sampled from a single crypt are called intracrypt distances; those among pairs of molecules sampled from different crypts are called intercrypt distances. The authors found that intercrypt distances were greater, on average, than intracrypt distances, but both distances were highly variable. Another important statistic is the number of unique tags per crypt; this is the total number of unique tags found in the complete sample of molecules from a crypt that are bisulfite-treated, PCR-amplified, and sequenced. In general, an increase in the relative magnitude of intracrypt versus intercrypt distances suggests a greater number of stem cells per niche.

To evaluate alternative models of stem cell turnover, Yatabe *et al.* (3) calculated the variance of the observed number of unique tags per crypt for samples from each of the 14 patients. The expected variance under each model of stem cell turnover was obtained by computer simulation. The stem cell models considered by these authors are instances of a conditional branching process model studied by Cannings (17) and can be readily simulated. The deterministic and stochastic stem cell turnover models can be viewed as special cases of a general model (Fig. 1). Define P to be the probability that one stem cell and one TA cell are produced at a cell division, and let q be the probability that either zero stem cells, or two stem cells, are produced. Note that $P + 2q = 1$ and therefore $q = (1 - P)/2$ so that P is the only free parameter. The deterministic stem cell model specifies that $P = 1$, and under the alternative (stochastic) model $P < 1$. A second parameter of both models is the number of stem cells, n, per niche. Yatabe *et al.* (3) compared the observed variances of the number of unique tags per crypt (for each patient) with the expected variance under two combinations of model parameters: a deterministic stem cell model with $P = 1$ and $n = 2$; a stochastic stem cell model with P ranging from 0.75 to 0.95 and n ranging from 4 to 512. The observed distribution of variance was most consistent with a stochastic model, but a broad range of values of P and n were possible.

Assuming one stem cell division per day, this would suggest a stem cell bottleneck, for any given crypt, about every 8.2 years, on average.

An interesting aspect of the intracrypt stem cell genealogy considered by Yatabe *et al.* (3) is the expected age of the most recent common ancestral (MRCA) stem cell of a niche (under the stochastic model). Assuming a 64-stem cell niche, they calculated that the MRCA of a crypt existed about 3,000 cell divisions in the past (on average). Assuming one stem cell division per day, this would suggest a stem cell bottleneck, for any given crypt, about every 8.2 years, on average. Similarly, the expected time to fixation, or loss, of a newly arisen mutant stem cell was predicted to be about 220 days. The authors point out that this figure roughly agrees with the clonal stabilization time of 1 year observed for human crypt heterogeneity after irradiation (18). Parameters such as the age of the MRCA of a population sample, or the time to fixation, or loss, of an allele in a population have long been a focus of interest for population geneticists, and the utility of these measures for studying population demography, and other aspects of population structure, is now widely accepted (19). The study of Yatabe *et al.* (3) demonstrates that analogous parameters may have a high degree of biological relevance for the study of population dynamics at the cellular level.

In the field of population genetics, much work has recently been devoted to the development of methods for estimating the ages of alleles (20). It has become clear that there are two potential sources of information about allele age: allele frequency (more frequent alleles tend to be older), and variation at genetic markers closely linked to the mutation defining the allele (more variation at linked markers usually suggests an older age). In some cases, natural selection favoring a mutation can greatly increase its population frequency (suggesting an old age if the allele were assumed to be neutral) despite its young age, but the lack of variation at linked genetic markers will still suggest a young age for the allele despite its high frequency. A conflict between estimates of allele age based on frequency versus variation at linked genetic markers can then be an indication of selection acting on the allele (21). An example is the ΔF508 mutation, which is the most common cause of cystic fibrosis in Europeans. This mutation appears much younger than one would expect, given its high frequency, supporting the idea that ΔF508 heterozygotes were favored by natural selection because of an increased resistance to diseases such as cholera. An intriguing possibility is that similar approaches could be used to detect somatic mutations that favor stem cell proliferation. Conflicts be-

tween the expected age of a mutant cell lineage (a cell bearing a mutation in an oncogene, for example), based on the frequency of the mutant in a tissue, and the age of the cell lineage as determined by using epigenetic markers could be used to identify cases of positive selection favoring particular mutations (mutations that increase the rate of stem cell proliferation, for example). This could provide a new tool for identifying precancerous cells.

The epigenetic tagging approach developed by Yatabe *et al.* (3) appears very promising. However, there are several outstanding questions that need to be addressed to establish the general utility and reliability of their approach. First, the accuracy (and reproducibility) of the experimental technique for inferring methylation patterns needs to be examined (22, 23). If the error rate (i.e., the chance of incorrectly inferring that a nonmethylated site is methylated) is high, this will increase variation in the number of unique tags per crypt and favor a stochastic model, even if the deterministic model is correct. Because the test of models is based on the variance in the number of distinct tags per crypt, many sources of random error in assigning tags will bias the test in favor of a stochastic model. Second, more realistic models of stem cell dynamics should be considered. Yatabe *et al.* (3) treat the number of stem cells per crypt, n, as constant in their model. If n varies across crypts within individual patients, this would increase the variance in the number of unique tags per crypt, again favoring a stochastic model.

Other potential factors that could bias test results, such as variation in methylation rates among crypts caused by endogenous, or exogenous, influences, have been at least partially addressed. For example, the percent methylation between *MYOD1* and *CSX* did not correlate within single crypts as would be expected if some common external influence were modifying methylation rates at all CpG sites within a crypt. Such correlations could also arise because of shared genealogies among genes within stem cells, however, so this is not an unambiguous test of variation in methylation rates among crypts. Despite several outstanding technical questions, methods for studying stem cell dynamics in humans using methylation-based epigenetic tags hold great promise. If these approaches prove feasible in other tissues, and especially in cancer cells, the outcomes could be revolutionary. Applying these methods to cancer cells may be particularly problematic, however, as rates of methylation may be greatly enhanced in some cell lineages (prohibiting the use of a common methylation rate of 10^{-5} for all cells) and CpG sites within, and among, genes may become uncoupled because of somatic recombination.

This research was supported by National Institutes of Health Grant HG01988 to B.R.

1. Watt, F. M. & Hogan, B. L. M. (2000) *Science* **287,** 1427–1430.
2. Loeffler, M. & Potten, C. S. (1997) in *Stem Cells*, ed. Potten, C. S. (Academic, San Diego), pp. 1–27.
3. Yatabe, Y., Tavare, S. & Shibata, D. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 10839–10844. (First Published August 21, 2001; 10.1073/pnas.191225998)
4. Robertson, K. D. & Jones, P. A. (2000) *Carcinogenesis* **21,** 461–467.
5. Robertson, K. D. & Wolffe, A. P. (2000) *Nat. Rev. Genet.* **1,** 11–19.
6. Bjerknes, M. & Cheng, H. (1981) *Am. J. Anat.* **160,** 77–91.
7. Winton, D. J., Blount, M. A. & Ponder, B. A. (1988) *Nature (London)* **333,** 463–466.
8. Hermiston, M. L., Green, R. P. & Gordon, J. I. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 8866–8870.
9. Bach, S. P., Renehan, A. G. & Potten, C. S. (2000) *Carcinogenesis* **21,** 469–476.
10. Ng, H. H. & Bird, A. (1999) *Curr. Opin. Genet. Dev.* **9,** 158–163.
11. Laird, P. W. & Jaenisch, R. (1996) *Annu. Rev. Genet.* **30,** 441–464.
12. Pfeifer, G. P., Steigerwald, S. D., Hansen, R. S., Gartler, S. M. & Riggs, A. D. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 8252–8256.
13. Oakeley, E. J. (1999) *Pharmacol. Ther.* **84,** 389–340.
14. Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., Molloy, P. L. & Paul, C. L. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 1827–1831.
15. Ross, M. H., Romrell, L. J. & Kaye, G. I. (1995) *Histology* (Williams & Wilkins, Baltimore).
16. Barkla, D. H. & Gibson, P. R. (1999) *Pathology* **31,** 230–238.
17. Cannings, C. (1974) *Adv. Appl. Prob.* **6,** 260–290.
18. Campbell, F., Williams, G. T., Appleton, M. A. C., Dixon, M. F., Harris, M. & Williams, E. D. (1996) *Gut* **39,** 569–573.
19. Hartl, D. L. & Clark, A. G. (1997) *Principles of Population Genetics* (Sinauer, Sunderland, MA).
20. Slatkin, M. & Rannala, B. (2000) *Annu. Rev. Genomics Hum. Genet.* **1,** 225–249.
21. Slatkin, M. & Rannala, B. (1997) *Am. J. Hum. Genet.* **60,** 447–458.
22. Thomassin, H., Oakeley, E. J. & Grange, T. (1999) *Methods* **19,** 465–475.
23. Warnecke, P. M., Stirzaker, C., Melki, J. R., Millar, D. S., Paul, C. L. & Clark, S. J. (1997) *Nucleic Acids Res.* **25,** 4422–4426.