Stochastic models for the evolution of stem cells in colon crypts

Simon Tavaré[1], Pierre Nicolas[2] and Darryl Shibata [3]

[1] Department of Biological Sciences, University of Southern California & Department of Oncology, University of Cambridge
[2] Unité Mathématique Informatique et Génome UR1077, Institut National de la Recherche Agronomique
[3] Department of Pathology, University of Southern California

1 Introduction

The human colon contains about 15 million crypts, each of which contains about 2,000 cells. Some of these are stem cells that maintain their population size when they divide, and some are transit amplifying (TA) cells that divide and differentiate to become the cells that repopulate the colon. It is of some interest to understand the behavior of stem cells in a typical crypt.

Since stem cells and TA cells cannot be readily distinguished from each other, inference about numbers of stem cells (for example) has to be indirect. One approach is to develop a stochastic model for the evolution of a crypt and infer the number of stem cells from molecular markers in a sample of cells taken from the crypt. One possibility is to type single nucleotide polymorphisms (SNPs) at a number of loci in these cells and use a comparison of the observed patterns of polymorphism as the basis for inference. This strategy, akin to those developed by population geneticists to study molecular variation in natural populations (cf. Nordborg 2001), does not work on the time scales of mitotic division because there will be essentially no variation found at each SNP. Instead, we use a marker that varies rapidly enough in mitosis to leave a signal — we follow CpG methylation patterns through mitotic division. The methylation status of each C in a CpG marker can change during mitosis; unmethylated Cs may become methylated, and methylated Cs may become demethylated. The particular CpG islands to be studied need to be chosen in regions around genes that are not expressed in the colon, so that the markers are effectively evolving neutrally. All islands are taken to be unmethylated initially.

We followed methylation in a CpG-rich region in the BGN gene on the X chromosome. This region of 77 basepairs contains 9 CpGs. Bisulfite treatment of the DNA from a crypt followed by PCR amplification, cloning and sequencing gives the methylation patterns at the 9 CpGs from a number of cells (Yatabe et al. 2001). Figure 1 illustrates these data.

2 Inferring the number of stem cells in a crypt

We sampled BGN methylation patterns from 7 male patients of ages between 40 and 87, taking 7 – 9 crypts per individual and 8 – 24 molecules per crypt. These data are described in detail in Nicolas et al. (2007).

The next stage of the analysis is a stochastic model for the evolution of the stem cells and the TA cells within a crypt, then across crypts in an individual and finally across individuals. There are two features of such a model: a description of the ancestry of the cells, and the superposition on that ancestry of the effects of methylation and demethylation. Within a given individual each

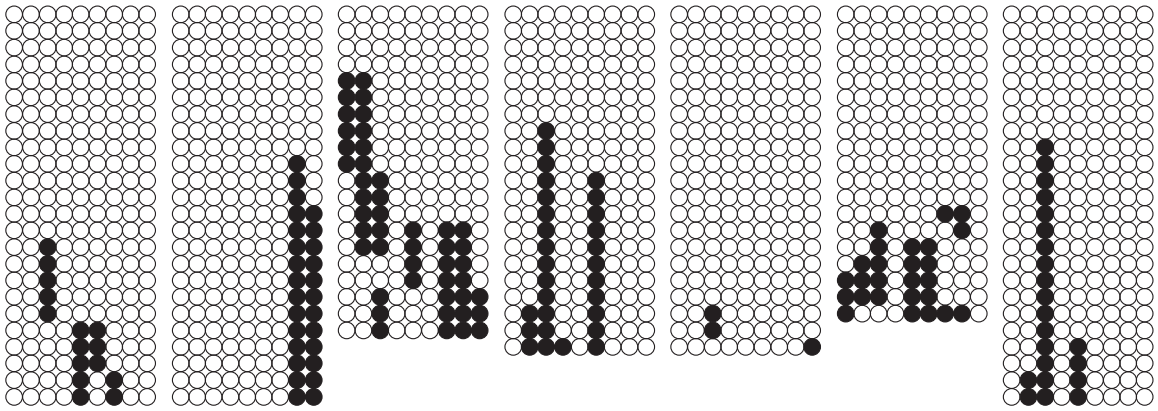

Figure 1: Methylation patterns at the BGN locus in a single individual. Each block corresponds to a different crypt. Within each block, each row represents a cell, each circle a CpG site. Filled circles denote methylated Cs, open circles unmethylated Cs.

crypt has gone through the same number of mitotic divisions, this number varying with the age of the individual. We ignore the possibility of crypt death and replacement. Stem cells divide *asymmetrically* to produce a single stem cell and a single TA cell, or *symmetrically* to produce either two stem cells or two TA cells. (A mathematician might wonder why the asymmetrical case really is called asymmetrical!) Each TA cell divides a small number of times before being lost from the crypt. We also allow for the possibility of miscalls of the methylation status in an island.

Space limitations preclude a detailed description of the model we have used; the reader is referred to Nicolas et al. (2007) for more information. In simple terms, the evolution of the stem cells within a crypt follows a coalescent-like process (Nordborg, 2001, provides a useful introduction to coalescents). We suppose that the cell content of the crypt is composed of N equal-sized sub-populations, each one corresponding to the progeny of one of the N stem cells. The genealogy of the cells sampled from the same sub-population is modelled back in time until its ancestral stem cell lineage. A number of different methylation processes were studied, including independent methylation events within a CpG island, and dependent methylation, in which methylation/demethylation rates can depend on the current methylation status of the whole island.

A Bayesian approach to model fitting and testing was implemented using Markov chain Monte Carlo. The non-stationarity of the process (numbers of divisions varying with age of the individuals), the replication over crypts and individuals, and the nature of the methylation process make this a challenging problem, our approach to which is given in Nicolas et al. (2007). One of the key steps is avoidance of peeling calculations on each coalescent tree, achieved by a variant of the ideas in Wilson and Balding (1998). Software and data can be downloaded at `http://genome.jouy.inra.fr/~pnicolas/mcmcniche/`.

Predictive assessment of model fitness based on comparing the inter-crypt average and standard deviation of a number of within-crypt statistics simulated from the posterior with those observed in the data revealed that one individual seems to behave differently from the rest. Once this individual is removed, adequate fits to the data are obtained.

For the present purposes we focus on the parameters N, the number of stem cells per crypt, and ν, the rate of the methylation process relative to the depth of the coalescent genealogy. Their

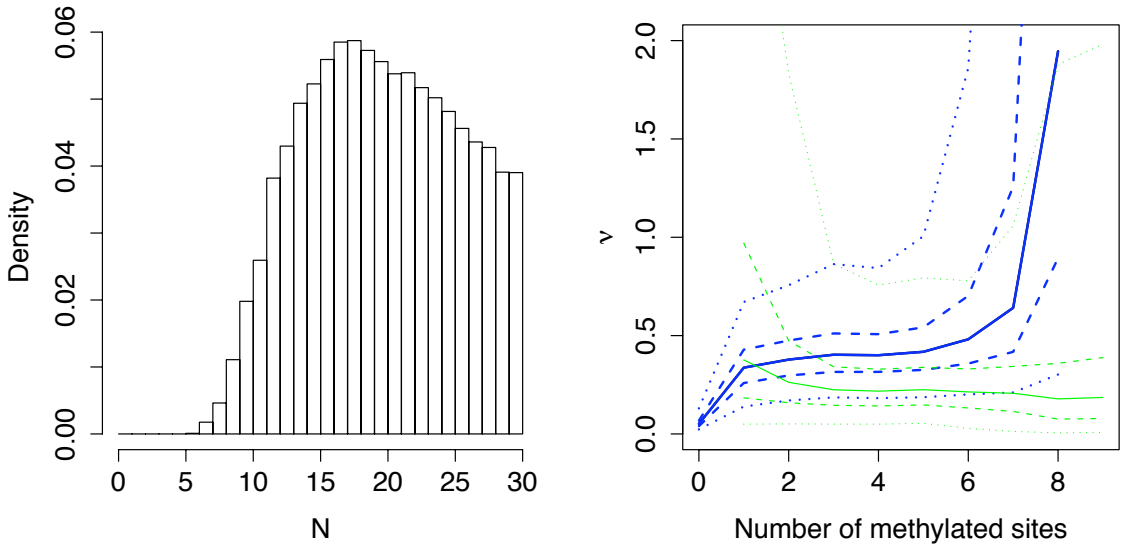

Figure 2: Posterior distribution of N and ν. Results are obtained using the model with context-dependent methylation rate. Left panel gives posterior of N, the number of stem cells per crypt. Right panel gives the posterior of ν, the scaled methylation rate. Lines show median (plain lines), first and third quartiles (dashed lines), and 95% credible interval (dotted lines) of both the methylation rate (bold dark line) and demethylation rate (light thin line) as a function of the number of methylated sites.

posterior distributions are shown in Figure 2. The posterior for N reaches its mode between 15 and 20, and provides no evidence for numbers of stem cells less than 6. The parameter ν reveals dependent methylation/demethylation across the sites of the BGN locus. The methylation rate is found to be highly dependent on the number of sites already methylated. The rate is very low when no sites are methylated and shows a more than fivefold increase when one site is already methylated. It is then relatively constant up to 7 methylated sites and then increases again. In contrast, demethylation dynamics do not seem to depend on the current level of methylation.

3 Discussion

We have used methylation patterns to track cell division in a number of other tissues, including endometrium, small intestine, hair and blood (Shibata and Tavaré, 2006). Methylation clocks appear to work well, in that the average methylation fractions increase with age, allowing mitotic history to be inferred from methylation patterns. There are however a number of drawbacks with our approach, not the least being that modelling of (in this case) crypt dynamics is required, and we are certainly not clear about all aspects of this process. For example, our current models in colon crypts do not exploit the spatial structure of the crypt.

Another drawback is the apparent paucity of data. It would certainly be advantageous to have substantially more sequences from each crypt, and to look at more CpG islands. In this

regard, the new sequencing technologies offer the opportunity to generate data on a far larger scale. At the time of writing, 454 sequencing should be able to produce 400,000 single molecule sequences from CpG islands in a collection of crypts in a single run. Whether this is feasible in terms of cost and time remains to be seen. If large numbers of observations were available from each crypt, the theory sketched earlier will need to be improved. Instead of digital read outs of methylation patterns from a few cells in a crypt, we will need to exploit accurate estimates of the proportion of each methylation haplotype in each crypt. This will doubtless provide a computational challenge.

Recently we have begun similar work on the analysis of tumor samples. The aim is to pin down the behavior of the elusive cancer stem cell — which need not be a cancerous stem cell, but rather a cell that leads to the cancer. As such, the analysis of cancer stem cells is really about common ancestry of cells. The genealogical approaches described in this paper may well turn out to be useful in this endeavor.

References

Nicolas P, Kim K-M, Shibata D & Tavaré S. (2007). The stem cell population of the human colon crypt: analysis via methylation patterns. *PLoS Computational Biology*, **3**(3), e28.

Nordborg M. (2001). Coalescent theory. *Handbook of Statistical Genetics*, 179-212. Edited by Balding DJ, Bishop MJ & Cannings C. Wiley, New York.

Shibata D & Tavaré S (2006). Counting divisions in a human somatic cell tree: how, what and why. *Cell Cycle*, **5**, 610-614.

Yatabe Y, Tavaré S & Shibata D. (2001). Investigating stem cells in human colon by using methylation patterns. *Proc. Natl. Acad. Sci. USA*, **98**, 10839-10844.

Wilson IJ & Balding DJ. (1998). Genealogical inference from microsatellite data. *Genetics*, **150**, 499-510.