Using methylation patterns for reconstructing cell division dynamics: assessing validation experiments

Daniel J. Andrews, Andy G. Lynch and Simon Tavaré[1]

Cancer Research UK Cambridge Institute, University of Cambridge, Robinson Way, Cambridge CB2 0RE, UK

Abstract

Methylation patterns present in a cell population can inform us about the way the cells are organized and how the population is sustained. Methylation is inheritable through cell divisions but changes can occur as a result of methylation replication errors. Hence, variation in methylation patterns in a cell population at a given time captures information about the history of the cell population. It is important that the observed methylation patterns are representative of those in the cell population. However, bisulfite sequencing may introduce new patterns and degradation may eliminate rare patterns. We investigate how bisulfite degradation may be expected to affect the data, and how inference could be made in light of this. A model for the data generation process makes it possible to estimate the starting number of distinct methylation patterns more accurately than simply counting the number of distinct patterns observed.

Keywords: polymerase chain reaction, lineage tracing, Markov chain Monte Carlo,

1. Introduction

Understanding the lineage relationships among cells and the dynamics of cell division is of great interest in developmental biology, cancer dynamics, stem cell dynamics, immunology, neurobiology and reproductive medicine, to name

[1]Daniel Andrews' current address: Illumina, CPC4, Capital Park, Fulbourn, Cambridge CB21 5XE, UK

just a few. The most celebrated success story is arguably the identification of the complete cell lineage tree of the nematode *C. elegans* [1]. For reviews of approaches for lineage tracing up to the end of the 20th century, see [2, 3] for example. As might have been anticipated, it has proved technically difficult to produce detailed lineage trees in higher organisms such as mouse and human. As a result, evolutionary approaches for constructing and interpreting lineage trees, which exploit changes in molecular markers during cell division as a surrogate for direct observation, have become common in the last 15 years (reviewed in [4, 5, 6] for example).

Several types of molecular marker have been used for this purpose. Microsatellite variability has been exploited in [7, 8, 9, 4, 10, 11], mitochondrial variation in [12], and variation in methylation status in [13, 14, 15, 16, 17]. In this article we focus on the use of methylation markers, which we now describe in more detail.

1.1. Using methylation patterns

The measurement of the methylation patterns present in a cell population can inform us about the way the cells are organized and how the population is sustained. Methylation is inheritable through cell divisions but changes can occur as a result of methylation replication error. Hence, variation in methylation patterns in a cell population at a given time captures information about the history of the cell population.

An example of a cell population that has been studied in this way is the human colon crypt. The colon crypt is found in the epithelium of the colon, has a cylindrical shape and consists of about 2000 cells. Residing at the bottom of the cylinder are stem cells from which originate the cell lineages of all the other cells found in the crypt. When a stem cell divides it will usually produce a daughter cell, which is committed to differentiation, as well as another stem cell. More rarely the stem cell will divide symmetrically to produce either two stem cells or two differentiating cells. Any cells committed to differentiation will move up towards the top of the cylinder, differentiating as they go, until

they become mature epithelial cells.

The methylation data we obtain are typically composed of methylation patterns obtained from bisulfite converted DNA sequenced at a small number of CpG sites in an amplicon a few hundred basepairs in length. Such patterns can be used to infer aspects of the dynamics of cells in the colon crypt by exploiting a probabilistic model for the cell population organization and for the observation process [18]. The authors fit a full probabilistic model for the stem cell genealogy, the methylation/demethylation process and the sampling, and perform inference using an MCMC algorithm. In [19] a cellular Potts model of crypt evolution is used, while the inference about stem cell structure is performed using approximate Bayesian computation (ABC). As another example, [20] infer the topological nature of the ancestral tree for tumour cells from methylation patterns and spatial data. They simulate methylation patterns and also use an ABC algorithm for the inference.

It is, of course, important that the observed methylation patterns are representative of those patterns in the cell population. However, we know that bisulfite sequencing may introduce new patterns, and degradation may eliminate rare patterns. Clearly, this is an occasion when studying the data generation process in more detail may prove beneficial. Here we investigate how bisulfite degradation may be expected to affect the data, and how inference could be made in light of this.

1.2. Bisulfite treatment

The methylation states of CpG sites can be measured by encoding this information into the DNA sequence. This is achieved by treating the DNA with bisulfite, which causes the complete deamination of cytosine to make uracil, while leaving 5mC unchanged. Thus, from a number of CpG sites, some will be indicated as methylated and others not; we call this the methylation pattern. Bisulfite treatment is followed by PCR where uracil is converted to thymine. These DNA molecules, with the methylation patterns encoded as substitutions, can be prepared and sequenced in the usual way. The sequenced reads can be

compared to a reference sequence and the methylation patterns can be inferred.

Bisulfite treatment introduces errors and biases, investigated by [21] and [22]. The method is based on the complete conversion of cytosine and the complete non-conversion of 5-methyl-cytosine. If either of these does not happen, i.e. if a cytosine fails to convert or a 5mC does convert, then an incorrect methylation pattern will be encoded, which may subsequently be sequenced.

DNA degradation is an undesired side-effect of bisulfite treatment; degraded molecules will not be sequenced and hence the methylation patterns of these will be absent from the read data. [21] conclude that complete conversion of cytosine can be achieved when the incubation of alkaline denatured DNA with a saturated bisulfite solution is performed for 4h at 55°C. They estimate that using these conditions between 84% and 96% of DNA is degraded. With such a high fraction of molecules being degraded there is a high chance that some of the more rare methylation patterns in the population may not be observed at all. This would be particularly concerning if the diversity of methylation patterns was of interest, because degradation is likely to decrease this diversity.

We consider a theoretical experiment where a single colon crypt, with approximately 2000 cells, is bisulfite-sequenced (using 454 sequencing technology). We consider methylation patterns made up from 9 CpG sites contained in a single amplicon sequence, numbers that are typical of real data (cf. [13, 18, 14, 19]). Specifically, we investigate how the measurement process affects inference of two quantities: the number of haploid genomes (out of the total of 4000) that have the most common pattern, and how many distinct patterns are present in the entire population. While we consider a specific case for illustration, the reader can generalise to other examples.

2. Errors, biases and uncertainty in bisulfite sequencing

To understand how the steps in the bisulfite sequencing protocol affect the analysis of methylation patterns we start by describing a naive statistical method for this type of data. Suppose we know that there are $N^{(0)}$ genomes in our

sample and we observe k different patterns with counts $y_1, \ldots, y_k$. Then the naive estimator of the number of distinct patterns is k and the estimator of the number of genomes with pattern i is $N^{(0)} y_i / Y$, where $Y = y_1 + \cdots + y_k$ is the total number of reads.

There are many ways that errors, biases and uncertainty can be introduced in the course of bisulfite sequencing, and which may consequently result in these estimators being biased or highly variable. For this study we limit the sources of error and bias that we consider. We consider only those biases that result from the loss of molecules from the experiment: e.g. by bisulfite degradation, by sampling, by ligation, and by bead placement. We expect that the loss of molecules from observation will affect estimates for the methylation pattern frequencies as well as estimation of the total number of distinct methylation patterns. In summary, we are focusing on the affect of losing molecules from the experiment, and we assess how small the degradation probability can be whilst still achieving small bias and variance in estimation.

3. Model for degradation and sampling

3.1. Modelling

We focus on the experimental steps of bisulfite degradation and other sampling steps; the resultant protocol can be described as follows. Let b be the number of CpGs at which the methylation status will be observed. Then there are $k = 2^b$ possible patterns which *wlog* we can label $1, \ldots, k$. A quantitative model for the protocol is as follows.

1. We start with a total of $N^{(0)}$ molecules containing the CpG sites, and with $n_i^{(0)}$ of pattern i for $i = 1, \ldots, k$; $n_1^{(0)} + \cdots + n_k^{(0)} = N^{(0)}$.
2. Bisulfite treatment causes the failure of a fraction $(1 - p)$ of the molecules leaving a total of $N^{(1)}$ molecules, which is $N^{(0)} p$ on average, and $n_i^{(1)}$ with pattern i for $i = 1, \ldots, k$.
3. PCR amplifies the number of each pattern by a constant factor M.

4. The number of reads Y will be smaller than $M \times N^{(1)}$ due to loss of molecules during ligation, bead placement, and other sampling steps that happen after PCR.

We can describe this model probabilistically as

$$n_i^{(1)} \sim \text{Binomial}(n_i^{(0)}, p) \quad \text{independently for } i = 1, \ldots, k \ ,$$

$$y|n^{(1)}, Y \sim \text{Multinomial}(Y, q) \quad \text{where} \quad q_i = \frac{n_i^{(1)}}{\sum_j n_j^{(1)}}.$$

Whether this model is accurate depends on the validity of a number of assumptions and approximations (beyond those made by ignoring the sources of bias and error as described above). These are as follows:

(i) Each molecule independently, and with probability $1 - p$, fails the bisulfite treatment.

(ii) Each molecule which survives the bisulfite treatment is independently and equally likely to be successfully read. With this assumption alone and supposing we knew the probability of a molecule being read to be r, then the second part of the model would be

$$y_i \sim \text{Binomial}(Mn_i^{(1)}, r) \quad \text{independently for } i = 1, \ldots, k \ ,$$

However, the probability r is the product of the probabilities of several (assumed) independent events: that a molecule has adapters successfully ligated to it; that a molecule is successfully hybridized to a bead; that a molecule survives any other sampling steps; that a molecule is not adsorbed onto any of the containers it is held in. Hence we will treat it as unknown. The natural way to remove r from the likelihood is by conditioning on $Y := y_1 + \cdots + y_k$, which is observed, leaving

$$p(y|n^{(1)}, Y) = \binom{MN^{(1)}}{Y}^{-1} \prod_{i=1}^{k} \binom{Mn_i^{(1)}}{y_i}.$$

In words, y is distributed as taking Y objects from an urn containing $Mn_i^{(1)}$ balls of colour i, *without* replacement (a multivariate hypergeometric random variable).

(iii) If the PCR is successful then M is large; for example, if PCR has 20–30 rounds with replication probability in 0.7-1.0, then M is between 4.1×10^4 and 1.1×10^9. We shall hence assume that $M \times N^{(1)}$ is many times larger than Y. Taking the limit as $M \to \infty$, y becomes distributed as above but *with* replacement (a multinomial random variable).

$$p(y|n^{(1)}, Y) = Y! \prod_{i=1}^{k} \frac{(n_i^{(1)} / \sum_j n_j^{(1)})^{y_i}}{y_i!}.$$

Assuming that r is unknown weakens our ability to make inference about $N^{(0)}$. However, for application to tissues such as colon crypts this is acceptable because the range of plausible values of $N^{(0)}$ is known in advance..

3.2. Simulation study: effects of degradation

Now that we have a model for the bisulfite sequencing protocol we can investigate by simulation the consequences of degradation by bisulfite treatment.

In Figure 1 we investigate the affect of the bisulfite survival probability p on the bias of the naive estimator of the starting number of distinct patterns. The more patterns that were present in the starting population (and hence the more rare patterns that are present) the more bias the estimator will have. This is because the fewer molecules that have a given pattern the more likely it is that all of them will be degraded. When $p = 0.25$ the bias is quite small, but it is very large for $p = 0.05$ and $p = 0.01$; the estimates being over 10 times too small in the latter case when there were more that 100 patterns present originally.

In Figure 2 we investigate the affect of p on the variance of the observed pattern count for a pattern originally present in 2000 out of the 4000 haploid genomes. When $p > 0.1$ this variance changes little, but as p decreases below this level the variance increases rapidly.

The consequences for estimation are that when p is small we expect large uncertainty about the frequency of any pattern and vary large uncertainty about the number of distinct patterns, especially when more than 100 patterns are observed.

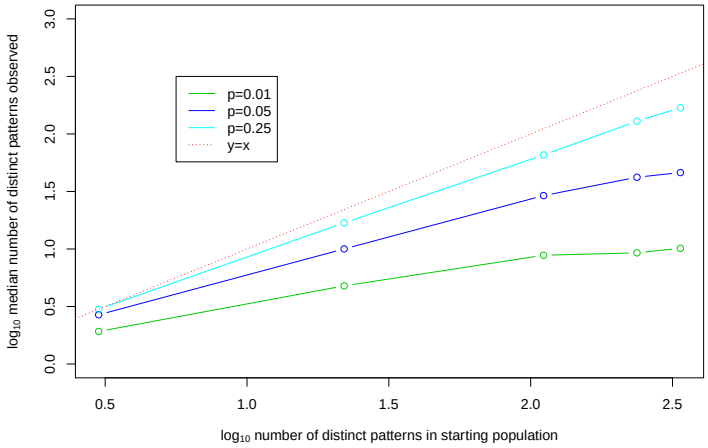

Figure 1: For 5 different starting pattern populations with number of distinct patterns, $3, 22, 111, 237, 337$ (out of a total of $2^9 = 512$), and for different degradation probabilities $p = 0.01, 0.05, 0.25$, the data y were sampled 100 times according to the model with $Y = 10^4$. The naive estimate of the number of distinct patterns is more biased for smaller p. (The starting populations are realisations of the prior discussed later in the article, with $\alpha \in (0, 1)$).

4. Statistical inference method

We have seen that the naive estimator for the starting number of distinct patterns is biased. We now present a Bayesian approach to inference, and develop an MCMC algorithm to generate samples that we will treat as being samples from the posterior.

Choice of non-informative prior We need to specify a prior for $(n_1^{(0)}, \ldots, n_k^{(0)})$. Considering this inference problem in isolation from everything known about the evolutionary process that created the population of methylation patterns, a simple three level prior would be

$$(n_i^{(0)})_{i=1}^k \sim \text{Multinomial}(N^{(0)}, (q_i)_{i=1}^k), \quad (q_i)_{i=1}^k \sim \text{Dirichlet}(\alpha \mathbf{1}_k), \quad \text{and } \alpha \sim \pi,$$

where $\mathbf{1}_k$ is a k-vector of ones, and π is a density function on $(0, \infty)$ that is yet to be decided. The priors in this family satisfy the sensible property of

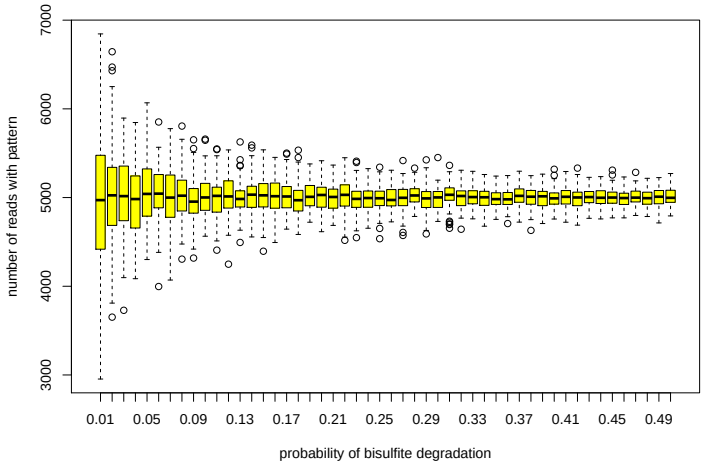

Figure 2: For an initial population $n^{(0)}$ such that the most frequent pattern was present in 50% of genomes, the data y were simulated 100 times, with $Y = 10^4$, and each for a range of p in $(0, 0.5)$. The observed number of this pattern has very large variation, but is unbiased, when p is small.

being exchangeable in $n^{(0)}$; that is, $\mathbb{P}(n_1^{(0)} = x_1, \ldots, n_k^{(0)} = x_k) = \mathbb{P}(n_{\sigma(1)}^{(0)} = x_1, \ldots, n_{\sigma(k)}^{(0)} = x_k)$ for any permutation $\sigma \in S_k$. Equivalently, our prior knowledge is invariant to the way the patterns are labelled. Simulations from

$$\pi(\alpha) = \frac{1}{3 \times 1.4} \alpha^{-2/3} \mathbf{1}_{(0,1.4^3)}(\alpha)$$

show that this prior is less informative for the number of distinct methylation patterns; see Chapter 2 of [23] for further details.

Algorithm development Figure 3 show a directed acyclic graph (DAG) representation of the model, including the prior and observation model. We shall show how $n^{(0)}$ and q can be integrated out.

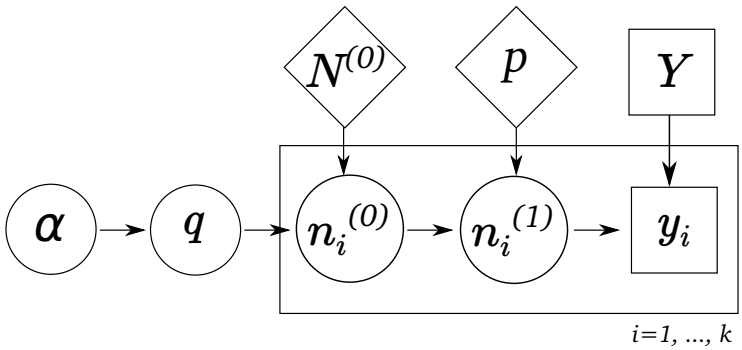

Figure 3: A directed acyclic graph (DAG) representation of the model. The circular nodes are those that are not observed; the square nodes are observed; the diamond nodes are assumed known. The arrows represent the conditional relationships: the conditional density of a node given all its ancestors is only a function of said node and its parent nodes. The nodes in the plate exist in k copies; two nodes in different plates are independent conditional on the parents of the plate-nodes.

The full conditional for $(n^{(0)}, q)$ is

$$\begin{aligned} p(n^{(0)}, q | n^{(1)}, \alpha, y) &\propto \prod_{i=1}^{k} \binom{n_i^{(0)}}{n_i^{(1)}} p^{n_i^{(1)}} (1-p)^{n_i^{(0)} - n_i^{(1)}} \mathbf{1}_{\{0 \le n_i^{(1)} \le n_i^{(0)}\}} \\ &\times N^{(0)}! \prod_{i=1}^{k} \frac{q_i^{n_i^{(0)}}}{n_i^{(0)}!} \mathbf{1}_{\{0 \le n_i^{(0)}\}} \mathbf{1}_{\{\sum_j n_j^{(0)} = N^{(0)}\}} \\ &\times \Gamma(\alpha k) \prod_{i=1}^{k} \frac{q_i^{\alpha - 1}}{\Gamma(\alpha)} \mathbf{1}_{\{0 \le q_i \le 1\}} \mathbf{1}_{\{\sum_j q_j = 1\}}, \end{aligned}$$

which is proportional (in $n^{(0)}$ and q) to

$$\begin{aligned} p(n^{(0)}, q | n^{(1)}, \alpha, y) &= (N^{(0)} - \sum_{j=1}^{k} n_j^{(1)})! \prod_{i=1}^{k} \frac{q_i^{n_i^{(0)} - n_i^{(1)}}}{(n_i^{(0)} - n_i^{(1)})!} \mathbf{1}_{\{\sum_j n_j^{(0)} = N^{(0)}\}} \mathbf{1}_{\{n_i^{(1)} \le n_i^{(0)}\}} \\ &\times \Gamma(\sum_{j=1}^{k} n_j^{(1)} + \alpha k) \prod_{i=1}^{k} \frac{q_i^{n_i^{(1)} + \alpha - 1}}{\Gamma(n_i^{(1)} + \alpha)} \mathbf{1}_{\{0 \le q_i \le 1\}} \mathbf{1}_{\{\sum_j q_j = 1\}}. \end{aligned}$$

Hence

$$n^{(0)} - n^{(1)} | n^{(1)}, q, y \sim \text{Multinomial}(N^{(0)} - \sum_{j=1}^{k} n_j^{(1)}, q)$$

and

$$q | n^{(1)}, \alpha, y \sim \text{Dirichlet}(n^{(1)} + \alpha \mathbf{1}_k).$$

Dividing $p(n^{(0)}, n^{(1)}, q, \alpha|y)$ by $p(n^{(0)}, q|n^{(1)}, \alpha, y)$ this leaves the joint posterior for $n^{(1)}, \alpha$

$$\begin{aligned} p(n^{(1)}, \alpha|y) &\propto \frac{N^{(0)}! p^{\sum_j n_j^{(1)}} (1-p)^{N^{(0)} - \sum_j n_j^{(1)}}}{(N^{(0)} - \sum_j n_j^{(1)})! (\sum_j n_j^{(1)})!} \mathbf{1}_{\{\sum_j n_j^{(1)} \leq N^{(0)}\}} \\ &\times \frac{(\sum_j n_j^{(1)})! \Gamma(\alpha k)}{\Gamma(\sum_j n_j^{(1)} + \alpha k)} \prod_{i=1}^{k} \frac{\Gamma(n_i^{(1)} + \alpha)}{(n_i^{(1)})! \Gamma(\alpha)} \left(\frac{n_i^{(1)}}{\sum_j n_j^{(1)}} \right)^{y_i} \mathbf{1}_{\{0 \leq n_i^{(1)}\}} \\ &\times \pi(\alpha), \end{aligned}$$

which cannot obviously be written in terms of simpler distributions.

The fact that the posterior distribution can be decomposed in this way suggests a program for sampling using an MCMC algorithm: get a sample from $n^{(1)}, \alpha|y$, then sample directly from the exact conditional distributions of $q|n^{(1)}, \alpha, y$, and then $n^{(0)}|n^{(1)}, q, y$. This should be better than sampling all the variables in a MCMC scheme, due to decreased correlation between samples.

The full conditionals for the random variables $\alpha, n_1^{(1)}, \ldots, n_k^{(1)}$ are

$$p(\alpha|n^{(1)}, y) \propto \frac{(\sum_j n_j^{(1)})! \, \Gamma(\alpha k)}{\Gamma(\sum_j n_j^{(1)} + \alpha k)} \quad \prod_{i=1}^{k} \frac{\Gamma(n_i^{(1)} + \alpha)}{(n_i^{(1)})! \Gamma(\alpha)} \quad \pi(\alpha)$$

and

$$\begin{aligned} p(n_i^{(1)}|s_{(-i)}, \alpha, y) &\propto \frac{(n_i^{(1)} + s_{(-i)})! \Gamma(\alpha k)}{\Gamma(n_i^{(1)} + s_{(-i)} + \alpha k)} \frac{\Gamma(n_i^{(1)} + \alpha)}{(n_i^{(1)})! \Gamma(\alpha)} \frac{{n_i^{(1)}}^{y_i}}{(n_i^{(1)} + s_{(-i)})^Y} \\ &\times \frac{N^{(0)}! p^{n_i^{(1)} + s_{(-i)}} (1-p)^{N^{(0)} - n_i^{(1)} - s_{(-i)}}}{(N^{(0)} - n_i^{(1)} - s_{(-i)})! (n_i^{(1)} + s_{(-i)})!} \mathbf{1}_{0 \leq \{n_i^{(1)} \leq N^{(0)} - s_{(-i)}\}}, \end{aligned}$$

where $s_{(-i)} = \sum_{j \neq i} n_j^{(1)}$.

One iteration of the MCMC algorithm proceeds as

1. update $n^{(1)}$ by k Metropolis-Hastings steps:
   for each $i = 1, \ldots, k$ update $n_i^{(1)}$ given $s_{(-i)}$ and α;
2. update α by a Metropolis-Hastings step;
3. sample q given $n^{(1)}$ and α;
4. sample $n^{(0)}$ given $n^{(1)}$, q and α.

There is one issue with this scheme that would prevent it working well in certain cases: if Y is large then the posterior for $n^{(0)}$ has local modes at vectors approximately integer multiples of the truth. If the algorithm is initialized near to a local mode then it will not converge to the true posterior.

Rather than altering the algorithm to get around this problem we set the initial $n^{(1)}$ to be sampled from

$$n^{(1)} \sim \text{Multinomial}\left(N^{(0)}p, \left\{\frac{n_i^{(1)}}{\sum_j n_j^{(1)}}\right\}_{i=1}^{k}\right)$$

to have a good chance of convergence to the true posterior around the global mode.

5. Simulation study: Bayesian inference

We investigate how our Bayesian inference method performs under different circumstances.

Investigating frequency estimates. For a given starting population of patterns $\{n_i^{(0)}\}_{i=1}^{512}$ we simulate the observation process 100 times, with $Y = 10^4$, each time computing the MAP estimate of the starting number of genomes with the truly most frequent pattern; this pattern is originally present with count 779 out of 4000. This we repeat with values of $p = 0.01, 0.05, 0.1, 0.25$.

Figure 4 shows that for small p the MAP estimator is biased downwards and highly variable. Both the bias and variance of this estimator decrease as p increases. The naive estimator seems to be much less biased than the MAP for smaller p, and performs equally well for larger p. The bias in the MAP estimate is due to the prior. The smaller p is, the more the posterior is more influenced by the prior, and the estimator is more biased.

Investigating diversity inference. Figure 5 shows the result of a simulation study investigating the performance of the Bayesian MAP estimator of the number of distinct patterns. When a $\log_{10}$ transformation is taken of the estimates it seems that the variance is stabilised; this means that the coefficient of variation of the

estimate does not depend on the true number of distinct patterns. The estimator variance is larger when $p = 0.01$ than when $p = 0.05$. It seems like the estimator is more-or-less unbiased, unlike the naive estimator that underestimates.

6. Discussion

We have seen in this article that degradation caused by bisulfite treatment has the potential to make the observed methylation patterns very unrepresentative of those present in the cell population. We developed a Bayesian MCMC method to infer the methylation patterns present in the cell population. We showed that the method allows us to accurately infer the number of distinct patterns originally present. However, when p is small, this method performs worse than the naive method at estimating the original count of a given pattern.

In this experiment it would seem that 99% degradation is too small to achieve accurate and precise estimation. This equates to an average of 40 molecules not being degraded. When degradation is this high the uncertainty about the original pattern counts is high and cannot be reduced by using methods based on models of the data generation process.

6.1. Different experiments

This limited study has been concerned with a particular experiment when the number of starting molecules is $N^{(0)} = 4000$, the number of CpG sites is $b = 9$, and the number of reads is $Y = 10^4$. The question remains as to how things would be different if these experimental parameters were different.

Concerning $N^{(0)}$. We expect that increasing $N^{(0)}$ and keeping the number of distinct patterns constant will have a very similar effect to increasing p, that is, more accurate inference of the pattern frequencies. If the pattern diversity also increases then there may be little improvement in the precision in estimating the starting number of distinct patterns, as the number of rare patterns may not change.

Concerning Y. The variance in the estimate of the proportion of molecules with pattern i comes from two sources: the variance in sampling $n_i^{(1)}$ and the variance in sampling y_i. Increasing Y will reduce the latter variance but not the former. Hence, there will be diminishing returns from increasing Y.

Concerning k. In this study we considered methylation patterns made up from 9 CpG sites, equivalently 512 possible patterns. If a different number of CpGs were used then it is likely that the prior would need to change in order to keep it uninformative for the number of distinct patterns. The number of possible patterns is $k = 2^b$ where b is the number of CpG sites. The MCMC algorithm is $O(2^b)$ and hence will become very slow for moderate b. This method would need to be adapted in that case.

6.2. Opportunities

In this study we rejected the uniform prior on the starting pattern counts as it was very informative for the number of patterns with non-zero count. We manufactured a prior that appeared uninformative for this function of the starting patterns, which we used for subsequent analysis. This prior is not entirely satisfactory as it biases inference of pattern frequency for small values of p; see Figure 4. Clearly this prior is informative for the marginal count of a given pattern, and may be informative for any function of the starting patterns other than the one we have considered.

We have made the assumption that $N^{(0)}$ and p are known "exactly". In practice they will have been estimated and there will be some associated uncertainty. Clearly, the more uncertainty there is about these parameters, the higher the posterior variance will be in inference about starting pattern counts and pattern diversity. It would be easy to include the uncertainty about $N^{(0)}$ and p within the Bayesian analysis by simply specifying appropriate priors and including updating steps in the MCMC algorithm.

Our simulation study in Section 5 aimed to demonstrate the inference method we had developed. As we had no suitable real data we had to simulate data,

which we did from the prior. By simulating from the prior we ignore the possibility that our model is misspecified.

We limited the scope of this study to exclude the possibility that new patterns might arise by bisulifite conversion, PCR or sequencing errors. However, it is clear that the presence of these errors will affect the ability of our method to infer the starting patterns. In particular, as under the model described in this article every observed pattern must have been present originally, every observed error pattern will at least shift the posterior distribution of the starting number of distinct patterns up by one. It is likely to have a greater effect than this, as error patterns will mostly be observed only a few times. The model will interpret seeing a few rare patterns as meaning the starting population had many rare patterns, some of which were lost to degradation, hence amplifying the bias.

6.3. Conclusions

In the introduction we claimed that understanding the data generation process would provide benefits. Have we seen any in this case? We now know that it is important that the probability of degradation $(1 - p)$ is small. Given a real experiment with some $N^{(0)}$, we could simulate the data generation process and investigate when p will be too small. Having a model for the data generation process has also made it possible for us to estimate the starting number of distinct patterns in a more accurate way.

References

[1] J. E. Sulston, E. Schierenberg, J. G. White, J. N. Thomson, The embryonic cell lineage of the nematode *Caenorhabditis elegans*, Dev Biol 100 (1) (1983) 64–119.

[2] J. D. Clarke, C. Tickle, Fate maps old and new, Nat Cell Biol 1 (4) (1999) E103–9. doi:10.1038/12105.

[3] C. D. Stern, S. E. Fraser, Tracing the lineage of tracing cell lineages, Nat Cell Biol 3 (9) (2001) E216–8. doi:10.1038/ncb0901-e216.

[4] D. Frumkin, A. Wasserstrom, S. Kaplan, U. Feige, E. Shapiro, Genomic variability within an organism exposes its cell lineage tree, PLoS Comput Biol 1 (5) (2005) e50. doi:10.1371/journal.pcbi.0010050.

[5] D. Shibata, S. Tavaré, Counting divisions in a human somatic cell tree: how, what and why?, Cell Cycle 5 (6) (2006) 610–4.

[6] D. Shibata, S. Tavaré, Stem cell chronicles: autobiographies within genomes, Stem Cell Rev 3 (1) (2007) 94–103.

[7] J. L. Tsao, J. Zhang, R. Salovaara, Z. H. Li, H. J. Järvinen, J. P. Mecklin, L. A. Aaltonen, D. Shibata, Tracing cell fates in human colorectal tumors from somatic microsatellite mutations: evidence of adenomas with stem cell architecture, Am J Pathol 153 (4) (1998) 1189–200. doi:10.1016/S0002-9440(10)65663-5.

[8] J. L. Tsao, S. Tavaré, R. Salovaara, J. R. Jass, L. A. Aaltonen, D. Shibata, Colorectal adenoma and cancer divergence. Evidence of multilineage progression, Am J Pathol 154 (6) (1999) 1815–24. doi:10.1016/S0002-9440(10)65437-5.

[9] J. L. Tsao, Y. Yatabe, R. Salovaara, H. J. Järvinen, J. P. Mecklin, L. A. Aaltonen, S. Tavaré, D. Shibata, Genetic reconstruction of individual colorectal tumor histories, Proc Natl Acad Sci U S A 97 (3) (2000) 1236–41.

[10] S. J. Salipante, M. S. Horwitz, Phylogenetic fate mapping, Proc Natl Acad Sci U S A 103 (14) (2006) 5448–53. doi:10.1073/pnas.0601265103.

[11] S. J. Salipante, J. M. Thompson, M. S. Horwitz, Phylogenetic fate mapping: theoretical and experimental studies applied to the development of mouse fibroblasts, Genetics 178 (2) (2008) 967–77. doi:10.1534/genetics.107.081018.

[12] A. Humphries, B. Cereser, L. J. Gay, D. S. J. Miller, B. Das, A. Gutteridge, G. Elia, E. Nye, R. Jeffery, R. Poulsom, M. R. Novelli, M. Rodriguez-Justo, S. A. C. McDonald, N. A. Wright, T. A. Graham, Lineage tracing reveals multipotent stem cells maintain human adenomas and the pattern of clonal expansion in tumor evolution, Proc Natl Acad Sci U S A 110 (27) (2013) E2490–9. doi:10.1073/pnas.1220353110.

[13] Y. Yatabe, S. Tavaré, D. Shibata, Investigating stem cells in human colon by using methylation patterns, Proc Natl Acad Sci U S A 98 (19) (2001) 10839–44. doi:10.1073/pnas.191225998.

[14] K. D. Siegmund, P. Marjoram, Y.-J. Woo, S. Tavaré, D. Shibata, Inferring clonal expansion and cancer stem cell dynamics from DNA methylation patterns in colorectal cancers, Proc Natl Acad Sci U S A 106 (12) (2009) 4828–33. doi:10.1073/pnas.0810276106.

[15] T. A. Graham, A. Humphries, T. Sanders, M. Rodriguez-Justo, P. J. Tadrous, S. L. Preston, M. R. Novelli, S. J. Leedham, S. A. C. McDonald, N. A. Wright, Use of methylation patterns to determine expansion of stem cell clones in human colon tissue, Gastroenterology 140 (4) (2011) 1241–1250.e1–9. doi:10.1053/j.gastro.2010.12.036.

[16] A. Sottoriva, I. Spiteri, D. Shibata, C. Curtis, S. Tavaré, Single-molecule genomic data delineate patient-specific tumor profiles and cancer stem cell organization, Cancer Res 73 (1) (2013) 41–9. doi:10.1158/0008-5472.CAN-12-2273.

[17] K. O. Koyanagi, Inferring cell differentiation processes based on phylogenetic analysis of genome-wide epigenetic information: hematopoiesis as a model case, Genome Biol Evol 7 (3) (2015) 699–705. doi:10.1093/gbe/evv024.

[18] P. Nicolas, K.-M. Kim, D. Shibata, S. Tavaré, The stem cell population of the human colon crypt: analysis via methylation patterns, PLoS Comput Biol 3 (3) (2007) e28. doi:10.1371/journal.pcbi.0030028.

[19] A. Sottoriva, S. Tavaré, Integrating approximate Bayesian computation with complex agent-based models for cancer research, in: G. Saporta, Y. Lechevallier (Eds.), COMPSTAT 2010 – Proceedings in Computational Statistics, Springer, Physica Verlag, 2010, pp. 57–66.

[20] K. D. Siegmund, P. Marjoram, D. Shibata, Modeling DNA methylation in a population of cancer cells, Stat Appl Genet Mol Biol 7 (1) (2008) Article 18. doi:10.2202/1544-6115.1374.

[21] C. Grunau, S. J. Clark, A. Rosenthal, Bisulfite genomic sequencing: systematic investigation of critical experimental parameters, Nucleic Acids Res 29 (13) (2001) E65–5.

[22] P. M. Warnecke, C. Stirzaker, J. R. Melki, D. S. Millar, C. L. Paul, S. J. Clark, Detection and measurement of PCR bias in quantitative methylation analysis of bisulphite-treated DNA, Nucleic Acids Res 25 (21) (1997) 4422–6.

[23] D. Andrews, Statistical models of PCR for quantification of target DNA by sequencing, Ph.D. thesis, University of Cambridge (2015).

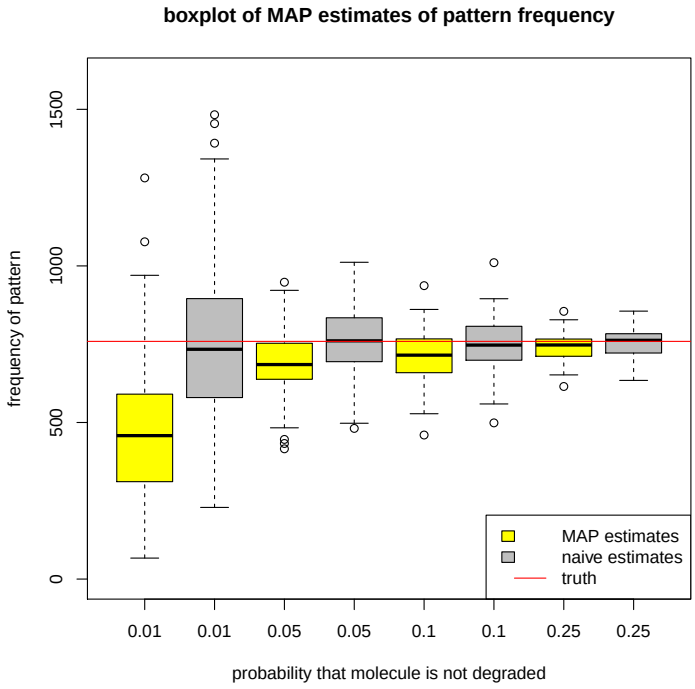

Figure 4: A starting population was simulated from the prior with $\alpha = 0.3^3$, resulting in a most frequent pattern of 779 out of 4000 and a total of 88 distinct patterns. For values of $p = 0.01, 0.05, 0.1, 0.5$, 100 sets of observed patterns were simulated from the model with $Y = 10^4$. For each of these datasets the MCMC algorithm produced MAP estimators of the frequency of the most frequent pattern. Also shown are box-plots of the naive estimates for the pattern count. For small p the MAP estimator is biased with a median of around 450. The bias and variance of the MAP estimator decreases with increasing p.

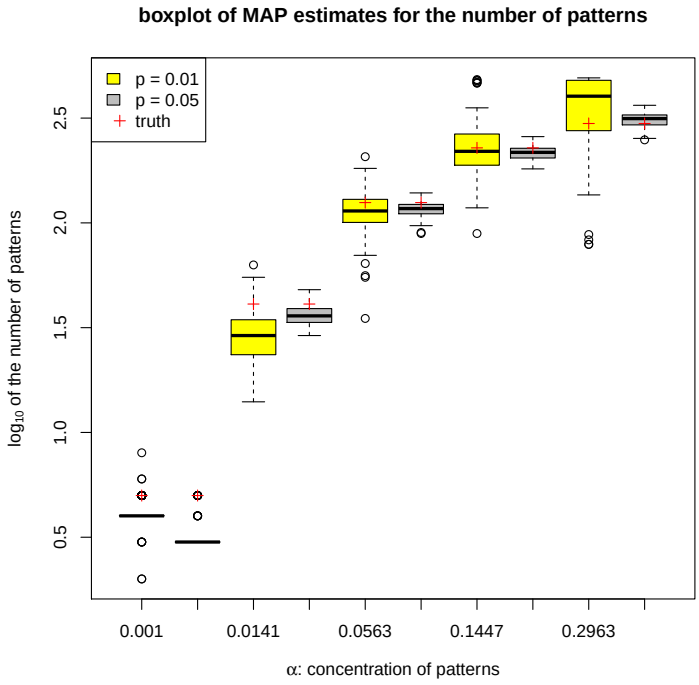

Figure 5: A range of starting populations was sampled from the prior with values of α such that $0 < \alpha < 0.3$. For each population 100 sets of observed patterns were simulated from the model with $Y = 10^4$. For each dataset our MCMC algorithm was run and the MAP estimate of the starting number of distinct patterns was computed. Shown are boxplots of $\log_{10}$ of these estimates when $p = 0.01$, and $p = 0.05$.