Models and Inference of Transmission of DNA Methylation Patterns in Mammalian Somatic Cells

Qiuyan Fu

A dissertation submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Washington

2008

Program Authorized to Offer Degree: Statistics

University of Washington
Graduate School

This is to certify that I have examined this copy of a doctoral dissertation by

Qiuyan Fu

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Chair of the Supervisory Committee:

_____
Matthew Stephens

Reading Committee:

_____
Charles D. Laird

_____
Matthew Stephens

_____
Elizabeth A. Thompson

Date: _____

University of Washington

Abstract

Models and Inference of Transmission of DNA Methylation Patterns in Mammalian Somatic Cells

Qiuyan Fu

Chair of the Supervisory Committee:
Professor Matthew Stephens
Department of Statistics, University of Washington

DNA methylation is critical for normal development of humans and many other organisms. Here we study the transmission process of methylation patterns over somatic cell division in mammals, with the goal of understanding: (1) to what extent methylation patterns are transmitted faithfully over cell generations, and (2) whether methylation enzymes, which are the forces in shaping methylation patterns, exhibit processivity, a property common to many enzymes, along DNA sequence.

We develop statistical models and inference methods for methylation data on pairs of parent and daughter strands from individual molecules, although which strand is which is unknown. To tackle the first question, we estimate the rates of methylation events and assess variability in these rates across CpG sites, formulating the question as a latent variable problem and developing multi-site models to infer the latent strand type. For the second question we take a different approach and model each type of methylation event as a hidden Markov chain. Average sojourn times (or distances in our case) of a Markov chain are used as a measure of the level of processivity. Our models for both questions can easily incorporate and estimate experimental error rates, which have not been accounted for in previous analyses of methylation patterns. Markov chain Monte Carlo techniques under a Bayesian framework are used

for inference.

We analyse double-stranded methylation data collected from the promoter region of the *FMR1* locus on the hypermethylated X chromosome in females and conclude that: (1) the average inappropriate bisulfite conversion error rate is 1–3% with little variation across sites; (2) the average failure of maintenance rate is 2-6% also with little variation across sites; (3) de novo events occur at a high rate at a few sites on the parent strand, and occur at some of the other sites on either the parent or the daughter strand; (4) the maintenance process is processive – on average, the stretch of maintenance events is 347–1716 nucleotides (nt), or equivalently, 52–256 CpG sites, whereas the stretch of failure of maintenance events is 6–17 nt (1–3 CpG sites).

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank Matthew Stephens for not only introducing me to the fascinating problems in this thesis, but also making research and thesis writing a fun process for me.

This thesis is based on collaborative work originated with Matthew Stephens in the Department of Statistics, and Charles Laird and Diane Genereux in the Department of Biology at the University of Washington. In particular, part of Chapters 2 and 3 is based on a manuscript co-authored by the four of us. Additionally, Charles and Diane drew our attention to the connections between spatial dependence and processivity of methylation enzymes, which gave rise to Chapter 4.

This thesis uses extensively the *FMR1* data collected by Reinhard Stöger and Brooks Miner in the Laird Lab; they have kindly granted permission of usage prior to complete publication, for which I am quite grateful.

I would also like to thank Matthew Stephens, Elizabeth Thompson, Charles Laird and Diane Genereux for carefully reading several versions of the thesis and for providing valuable comments and suggestions for revisions.

I further acknowledge the Department of Statistics and the Center for Studies in Demography and Ecology for computing support, and Charles Laird and the Department of Statistics for financial support.

DEDICATION

To my parents and brother;

To all my dear friends.

Chapter 1

INTRODUCTION

Epigenetic mechanisms are phenomena that lead to heritable changes in phenotypes and in gene functions *without* altering DNA sequence itself (Bird, 2002; Strachan and Read, 2004). DNA methylation is one such mechanmism. A good example that illustrates this concept is a calico cat which has yellow and black patches in her fur. A calico cat is always female, having two X chromosomes. Genes responsible for those coloured patches are located on the X chromosomes. The genotypes in the skin cells of a calico cat can be written as X^yX^b: one X chromosome carries the "yellow" allele and the other the "black" allele; alleles are alternative versions of a gene. Methylation stably inactivates the "black" allele in cells in yellow patches and the "yellow" allele in black patches, thus resulting in patches of two different colours in the calico cat. DNA methylation occurs in plants, bacteria and fungi as well: it can change the morphology of flowers (e.g. Cubas et al., 1999 on *Linaria vulgaris*), and protect bacteria from infection (Peterson and Reich, 2006). Epigenetic mechanisms are quite different from the genetic ones. The latter involve mutations in DNA sequence that result in heritable and nearly irreversible changes in phenotypes. For example, a single base mutation can cause sickle cell anemia that is inherited through many generations. In contrast, epigenetic changes are meta-stable (i.e. marginally stable), and are not always inherited through generations.

DNA methylation plays an important role in normal human development, such as genetic imprinting (Barlow et al., 1991), X chromosome inactivation (Lyon, 1972), embryonic development (Jaenisch and Bird, 2003) and cell differentiation (Reik et al.,

2001). Aberrant methylation patterns, however, can lead to developmental diseases (Dittrich et al., 1992; Laird, 1987) and cancers (Jones and Baylin, 2002; Clark and Melki, 2002). C. Laird and his colleagues studied methylation patterns (Stöger et al., 1997) in a family affected by fragile X syndrome, a developmental disease, the symptoms of which include mental retardation. In particular, they looked at methylation patterns from the promoter region of the *FMR1* locus; this locus is responsible for this syndrome and is located on the X chromosome. Figure 1.1 displays the cognitive status of six male offspring in the family and their methylation density in this region. Five of the six offspring are affected to various degrees. The figure shows that disparity in methylation patterns can result in striking differences in IQ scores.

DNA methylation occurs when a methyl group is attached to a cytosine in a DNA molecule. Cytosines (Cs) are one of the four types of bases that constitute the genome, the other three types being adenine (A), guanine (G) and thymine (T). Methylated cytosines are sometimes referred to as the "fifth base", since they are functionally different from unmethylated cytosines and play an important role in normal development in organisms. In this thesis we study the transmission process of methylation patterns over somatic cell division in mammals by analysing double-stranded sequence methylation data collected using the hairpin-bisulfite PCR technique. We develop statistical models and inference methods for rates of methylation events and processivity of methylation enzymes during the transmission process. The thesis is organised as follows:

- Chapter 1 introduces the basic terminology related to and current knowledge of DNA methylation and the transmission process of methylation patterns; these terms will be used repeatedly throughout the thesis. We will describe hairpin-bisulfite PCR (Laird et al., 2004) which is used to obtain double-stranded methylation data and issues of statistical interest that arise in the experiment.

- Chapter 2 lays out the model that describes the transmission process of methy-

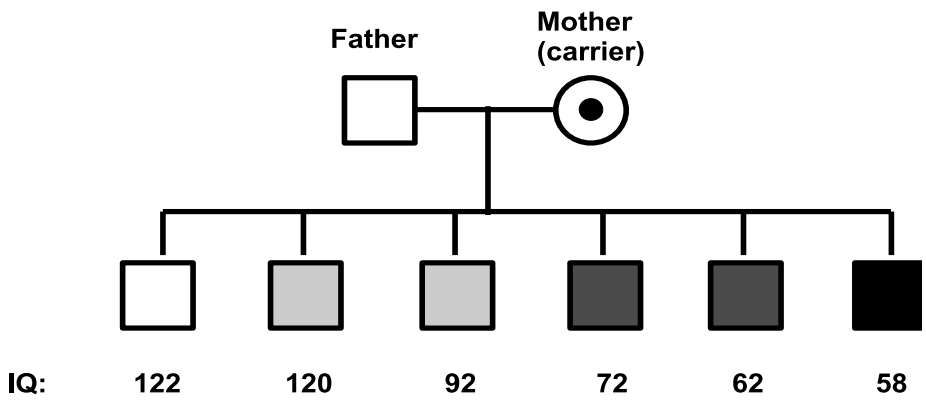

Figure 1.1: The fragile X family. The parents are unaffected, although the mother is a pre-mutation carrier, i.e. she is not affected by fragile X syndrome, but one of her X chromosomes is in the pre-mutation state. The shading indicates the methylation density level in the promoter region of the FMR1 locus on the only X chromosome in a male. The darker the shading, the higher the methylation density, and the lower the IQ score. One son (first from left) inherited the normal X chromosome from the mother, whereas the other five inherited the pre-mutated X chromosome from the mother. Full mutation developed in some of these five sons, as indicated by high methylation density in this region. The second and third sons from left are non-penetrant, even though they also show an elevated methylation level. The three sons from right have low IQ scores, a symptom of fragile X syndrome. Figure adapted from Figure 2 in Stöger et al. (1997).

lation patterns, and defines methylation events and their rates. We review existing methods for estimating those rates from double-stranded methylation data. We then introduce a simple statistical model for rate estimation that uses the methylation density and multi-site information readily available in the double-stranded data. We use the expectation-maximisation (EM) algorithm for inference. We will look at how this model performs on simulated data as

well as data from the *FMR1* locus (collected by Brooks Miner and Reinhard Stöger in the Laird Lab at the University of Washington), and what biological insight we gain from using this simple model.

- Chapter 3 extends the simple multi-site model developed in Chapter 2 to hierarchical models. They allow for rates of methylation events to vary across sites, account for experimental errors and estimate their rates, and incorporate temporal stationarity, an important assumption underlying several existing methods. We adopt a Bayesian framework and use Markov chain Monte Carlo (MCMC) for inference. These more sophisticated multi-site models are applied to the *FMR1* data.

- Chapter 4 concerns the possibility of methylation enzymes behaving in a processive manner. Many enzymes that bind to molecules act in a processive manner; that is, the enzyme binds to the molecule and stay on it for a while before falling off. Methylation enzymes play a key role in shaping methylation patterns over cell generations. Dnmt1, commonly referred to as the maintenance enzyme, is generally thought to be processive. We describe hidden Markov models developed for this property using the double-stranded in vivo methylation patterns. We compare our analysis of the *FMR1* data with other studies of processivity. To our knowledge this is the first statistical analysis for the *in vivo* processivity.

- Chapter 5 summarizes the features of the models and the scientific findings. It also points to possible future directions.

1.1 Biological background

1.1.1 The basics of DNA methylation

We begin by briefly describing basic properties of DNA molecules. A key property is that they are double-stranded, each strand consisting of a sequence of nucleotide bases of the four types. The two strands are complementary to one another: C on one strand pairs with G on the other, and A with T, by hydrogen bonding. For example, below is a double-stranded DNA sequence,

```
5'-ATCGG-3'
3'-TAGCC-5'
```

in which 5' and 3' indicate the chemical polarity of a strand. It is conventional to read each strand from its 5' end to its 3' end. To measure the length of a sequence, we will use base pairs (bp) for double-stranded sequences, or nucleotides (nt) for single-stranded sequences. Hence, the above sequence is 5 bp long.

In somatic cells of mammals, DNA methylation occurs almost exclusively on the 5' cytosines in CG dinucleotides, also known as a CpG. The lower case p in CpG represents the phosphate that links the two bases together. Thus, it stresses that the two bases are physically adjacent on the same strand. CpGs appearing on opposite strands as follows:

```
5'-CG-3'
3'-GC-5',
```

which are termed a CpG/CpG, a CpG dyad, or a CpG site, used interchangeably throughout this thesis. Neither, one or both cytosines at a CpG site can be methylated, resulting in three types of CpG dyads: unmethylated, hemimethylated, or methylated, respectively. On average, about 70–80% of CpG sites in mammalian genomes are methylated (Bird, 2002).

1.1.2 Transmission of methylation patterns, methylation enzymes and methylation events

An important property of DNA methylation is its highly accurate transmission over somatic cell division in mammals. Somatic cells are any cells except for the sperm and egg cells. The process of somatic cell replication and division is also called mitosis. The transmission process of methylation patterns is tied closely to DNA replication, but usually occurs just after the cytosine is incorporated into DNA. The calico cat provides a good example of the faithful transmission of methylation. Her skin cells go through many rounds of cell division over her lifetime, but the colour of a patch does not switch from yellow to black, or vice versa, indicating that the methylation patterns remain relatively stable. Here we sketch the DNA replication mechanism, which provides a platform for methylation events.

At each of several initiation points of DNA replication, the double-stranded DNA molecule unwinds and forms a replication bubble, which consists of two replication forks pointing in opposite directions. The single-stranded segments on those forks are "templates". DNA polymerase enzymes add to the new strand a base complementary to the base on the template strand, moving from the 5' end of the new strand to its 3' end. Replication bubbles enlarge and merge, producing four strands, each being a copy of the genome. Thus, in preparation for cell division, the original double-stranded molecule gives rise to two double-stranded molecules, each composed of a template strand, termed the *parent* strand, and its newly synthesised, complementary strand, termed the *daughter* strand.

The strict base-complementarity rules help ensure that this replication process is highly accurate. However, these rules do not ensure that patterns of methylation in the parent strand are also accurately transmitted. This is because methyl groups are not integrated during DNA replication; instead, they must be added later, as mentioned earlier. Although the transmission of methylation patterns is less well un-

derstood than the transmission of nucleotides, a widely-accepted model is that, just after the production of the daughter strand, the parent strand has the same methylation patterns as before, where the daughter strand is not methylated at all. Thus CpG dyads at this stage are either unmethylated or hemimethylated on the parent strand. Dnmt1, a methyltransferase (i.e. methylation enzyme), mostly targets hemimethylated CpG dyads and adds methyl groups to the daughter strand at those CpG sites (Vilkaitis et al., 2005). In particular, it is suggested that Dnmt1 may couple with the replication machinery and move also from the 5' end to the 3' end of the daughter strand, although this coupling may not be essential to the maintaining of methylation patterns (Schermelleh et al., 2007). This process is highly accurate in mammalian somatic cells, but imperfect transmission does occur. The daughter strand sometimes remains unmethylated even when the parent is methylated, an event we refer to as *failure of maintenance*. On the other hand, two other methyltransferases, namely, Dnmt3a and Dnmt3b, are mainly responsible for introduction of methylation at CpG dyads unmethylated before replication, events we will refer to as *de novo* methylation events. Since maintenance methylation is imperfect, de novo methylation events are essential to avoid long-term declines in methylation density. Sequence data indicate that de novo events must occur on at least daughter strands (Genereux et al., 2005). It is not known, however, whether de novo events also occur on parent strands.

1.2 Hairpin-bisulfite PCR and double-stranded sequence methylation data

To investigate rates and processivity of methylation events, we need to identify locations of methylated cytosines in genomic DNA. Conventional sequencing methods do not provide information about methylation states. Instead, bisulfite treatment of DNA, the most general method for detecting methylation patterns on DNA molecules, is used. It is based on the fact that sodium bisulfite, when applied to single-stranded DNA, converts unmethylated cytosines to a different base, uracil. Although uracil

does not naturally occur in DNA, it does pair with adenine like the natural DNA base thymine does. Subsequent PCR amplification further produces thymine where uracil is. Hence, identifying thymine in the PCR-derived sequence of the bisulfite-treated DNA effectively identifies the locations of unmethylated cytosines in the original molecule.

Two types of conversion error are possible during the bisulfite treatment. One is failure of conversion: it means that bisulfite treatment fails to convert unmethylated cytosines and such cytosines would be misidentified as methylated. This type of error has received a considerable amount of attention (Grunau et al., 2001), because it is a measure of the performance of bisulfite treatment. The other type of error is inappropriate conversion error (Shiraishi and Hayatsu, 2004); it means that bisulfite converts methylated cytosines and these cytosines would be identified as unmethylated. These errors are still under investigation (Genereux et al., 2008). They add noise to the observed data, and can have a significant impact on the inference. Therefore, analysis methods should take those errors into consideration.

Because bisulfite conversion requires separation of the double-stranded molecules into their single-stranded components, a process called denaturation, standard bisulfite PCR generates methylation patterns for a number of single-stranded sequences when applied to a population of DNA molecules. Most current technologies thus yield information on only either the top or bottom strands in a population of molecules, and provide no information on the strands' relationships to one other. To preserve this important information from double-stranded molecules, Laird et al. (2004) introduced a new technique, hairpin-bisulfite PCR (Figure 1.2), which uses a short DNA sequence as a "hairpin" to link together parent and daughter strands within a cell before denaturation. This step ensures that the strands, when separated by denaturation, will remain attached to form one extended single-stranded molecule, to which bisulfite PCR can then be applied, and the methylation patterns read (Figure 1.2). This technique therefore yields data on methylation patterns for *pairs* of parent-

daughter strands, although which strand is the parent and which the daughter is not determinable by current technologies. Miner et al. (2004) further employed quality control procedures to ensure that the double-stranded data from the hairpin-bisulfite PCR technique are authenticated; that is, the data are representative of the population of molecules analysed, and not from contaminated or redundant molecules. More details on quality control are in Appendix A.

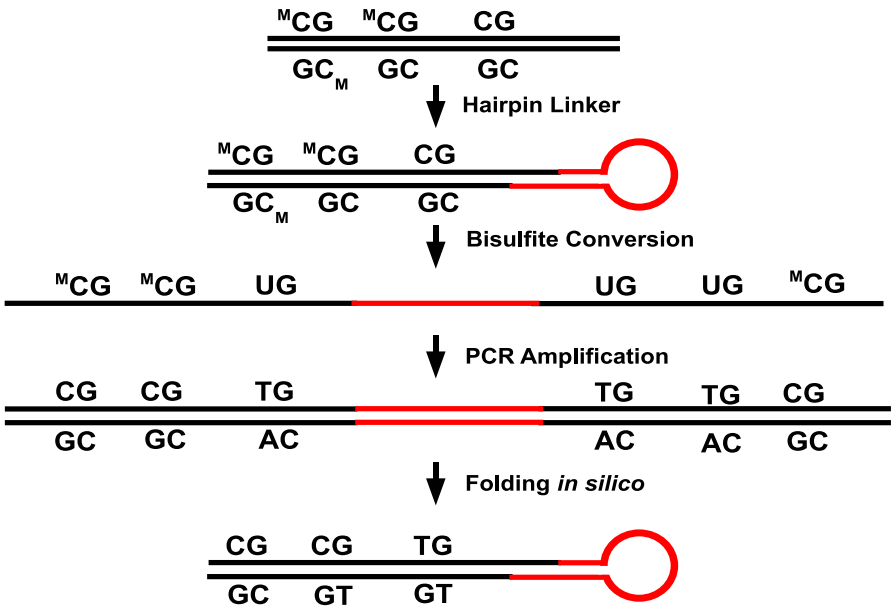

Figure 1.2: An illustration of major steps of the hairpin-bisulfite PCR technique. Three CpG dyads on a hypothetical double-stranded molecule are methylated, hemimethylated and unmethylated. ^{M}C represents methylated cytosines. After restriction enzyme cuts the two strands in a molecule to leave a "sticky" end on one of the strands, the hairpin linker binds to the two strands with the help of DNA ligase. Denaturation breaks down hydrogen bonding between parent and daughter strands, but the hairpin linker holds the ends of the strands together. Bisulfite treatment converts unmethylated cytosines to uracils. The long single strand after denaturation and bisulfite treatment then goes through many rounds of PCR amplification. Although we are unable to determine which strand is the parent strand and which the daughter strand, it is possible to identify hemimethylated CpG sites. Figure adapted from Figure 1 in Riggs and Xiong (2004), which is based on Laird et al. (2004).

Another possible source of error is PCR sequencing, during which PCR crossovers may occur. That is, a strand from molecule A swaps segments with a strand from molecule B. The nucleotide bases on two strands in either resulting sequence still show complementarity to each other, but the methylation patterns have changed. However, PCR crossovers happen very infrequently, at the rate of about 1% (Burden et al., 2005). Therefore we ignore this error in our analyses.

Hairpin bisulfite PCR is currently one of the few techniques for obtaining methylation patterns from double-stranded DNA molecules (see, for example, Bird (1978) for an alternative). However, several aspects of this technique limit the size of the available data set: (i) bisulfite conversion is highly destructive of genomic DNA templates, so even large amounts of DNA yield only a limited number of intact templates amplifiable by PCR; (ii) Hairpin bisulfite PCR is time-consuming, requiring many more steps than conventional bisulfite PCR; (iii) due to PCR contamination and sampling redundancy, the number of *unique, useful* sequences obtained is sometimes far less than the number actually sequenced. These issues make the development of efficient statistical analysis methods particularly appealing, as they can make full use of the available data.

Methylation data that arise from hairpin-bisulfite PCR differ fundamentally from data produced by genome-wide methylation profiling techniques. Profiling techniques (see Ordway et al., 2006 for example) provide a summary of which genomic regions tend to be methylated in a *population* of DNA molecules, and so are tremendously useful for identifying loci whose alternate methylation states may account for phenotypic differences among cell lines or individuals. By contrast, hairpin-bisulfite PCR reveals methylation patterns at loci on individual double-stranded sequence, and thus preserves information about the methylation states of multiple CpG sites on *individual* DNA molecules. This information is essential for analysing methylation inheritance, and for investigating possible site-site correlations in methylation events.

1.3 ***The*** *FMR1* ***data***

As mentioned at the beginning of this chapter, the Laird Lab has been investigating the role of DNA methylation at the *FMR1* locus on the X chromosome, in part because of its role in fragile X syndrome. In XX females (for example the calico cat mentioned before) genes on one of the two X chromosomes are in general methylated and inactivated to compensate for the extra X chromosome in XX females compared to XY males. On the other hand, methylation of the *FMR1* locus on the active X chromosome in women and on the only (and hence active) X chromosome in men can inactivate this gene and cause the disease (Laird, 1987; Stöger et al., 1997). As part of the fragile X investigation, the Laird Lab collected double-stranded methylation data from the promoter region at the *FMR1* gene on the hypermethylated X chromosome in females unaffected by fragile X syndrome, using the hairpin-bisulfite PCR technique.

The Laird Lab collected two sets of data, each in a segment of the promoter region from two groups of females, although three individuals were included in both groups. Below are methylation patterns for four double-stranded sequences at 22 CpG sites taken from three different females (F1, cF3 and F7):

```
F1,0,0,1,0,1,1,1,0,1,1,1,1,1,1,1,1,1,0,0,1,1,1
F1,0,0,0,1,1,1,1,1,1,1,0,0,1,1,1,1,1,1,1,1,0,1

F1,0,0,0,1,1,1,1,1,1,1,0,0,1,1,1,1,1,1,1,1,1,1
F1,0,0,0,1,1,1,1,1,1,1,0,0,1,1,1,1,1,1,1,1,0,1

cF3,0,1,1,1,1,1,1,1,1,1,1,0,1,1,1,1,1,0,0,1,0,1
cF3,0,1,1,1,1,1,0,1,1,1,0,1,1,1,0,1,1,1,1,1,1,1

F7,0,1,0,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,0
F7,0,1,0,0,1,1,1,0,1,1,1,1,1,1,1,1,1,1,1,0,0,0
```

where 0 denotes an unmethylated CpG and 1 a methylated one. Summary statistics of the two data sets are given in Table 1.1. These high-level summary statistics, however,

Table 1.1: Summary statistics of the two double-stranded *FMR1* data sets collected by the Laird Lab using hairpin-bisulfite PCR. The data were from two segments in the promoter region of the *FMR1* gene on the hypermethylated X chromosome in females unaffected by fragile X syndrome. Each double-stranded (ds) sequence has two strands. Each CpG site can be methylated (M), hemimethylated (H), or unmethylated (U).

	Data Set 1 (Sites 1–22)	Data Set 2 (Sites 25–52)
Length of segment (bp)	141	183
No. of CpG sites	22	28
No. of females represented in data	6	6
No. of ds seqs.	169	83
Proportions of (M,H,U) (%)	(82, 6, 12)	(64, 8, 28)

reveal little information of site variation, while the four sequences we showed above suggest that proportions of methylated (M), hemimethylated (H) and unmethylated (U) CpG dyads vary a lot across CpG sites, resulting in varying methylation density. To reflect this variation we summarise counts of (M,H,U) and methylation density at each site in Tables 1.2 and 1.3.

Table 1.2: Site-specific counts of methylated (M), hemimethylated (H) and unmethylated (U), as well as methylation densities of the *FMR1* data at sites 1–22. There are 169 double-stranded sequences. $\hat{m}$ is the proportion of cytosines at a site that are methylated.

Site Index	M	H	U	$\hat{m}$
1	132	15	22	0.83
2	117	12	40	0.73
3	146	11	12	0.90
4	116	15	38	0.73
5	115	14	40	0.72
6	146	9	14	0.89
7	145	10	14	0.89
8	154	8	7	0.93
9	154	10	5	0.94
10	151	17	1	0.94
11	136	7	26	0.83
12	143	8	18	0.87
13	154	7	8	0.93
14	151	17	1	0.94
15	152	14	3	0.94
16	157	12	0	0.96
17	151	5	13	0.91
18	117	7	45	0.71
19	118	6	45	0.72
20	141	12	16	0.87
21	115	13	41	0.72
22	135	8	26	0.82

Table 1.3: Site-specific counts of methylated (M), hemimethylated (H) and unmethylated (U), as well as methylation densities of the *FMR1* data at sites 25–52. There are 169 double-stranded sequences. $\hat{m}$ is the proportion of cytosines at a site that are methylated.

Site Index	M	H	U	$\hat{m}$	Site Index	M	H	U	$\hat{m}$
1	24	9	50	0.34	15	51	6	26	0.65
2	30	5	48	0.39	16	70	8	5	0.89
3	38	4	41	0.48	17	54	7	22	0.69
4	20	4	59	0.27	18	65	6	12	0.82
5	26	11	46	0.38	19	77	3	3	0.95
6	63	10	10	0.82	20	39	7	37	0.51
7	77	3	3	0.95	21	51	6	26	0.65
8	73	9	1	0.93	22	55	6	22	0.70
9	79	4	0	0.98	23	54	6	23	0.69
10	63	4	16	0.78	24	52	7	24	0.67
11	55	4	24	0.69	25	40	8	35	0.53
12	68	7	8	0.86	26	58	6	19	0.73
13	64	9	10	0.83	27	28	6	49	0.37
14	50	8	25	0.65	28	60	6	17	0.76

1.4 Questions of interest

We use the double-stranded methylation data to address two fundamental questions about DNA methylation:

1. To what extent are methylation patterns faithfully transmitted from one cell generation to the next? Since imperfect transmission is due to failure of maintenance and de novo methylation events, we approach this first question by estimating rates of those events as well as assessing variation in those rates across CpG sites.

   In terms of the failure of maintenance rate, Bird (1978) used restriction enzymes to detect methylation status, and data from his experiments indicate this rate to be less than 2%. Studies based on bisulfite treatment have estimated this rate to be 4–5% (Laird et al., 2004; Vilkaitis et al., 2005; Genereux et al., 2005) or as low as below 1% (for example, 0.1% in Pfeifer et al., 1990; 0.08–0.15% in Ushijima et al., 2003). Additionally, Genereux et al. (2005) obtained site-specific estimates of failure of maintenance rate and concluded that there is little variation in this rate across sites. The questions are then, which is closer to the truth and why are these estimates so different?

   By comparison, it is more difficult to estimate de novo rates and the results so far have been quite variable. For example, Pfeifer et al. (1990) estimated a total de novo rate to be about 5%. whereas Laird et al. (2004) assumed no parent de novo events and estimated the daughter de novo rate to be around 0.17. Genereux et al. (2005) estimated de novo rates under each of the following two assumptions: (1) the parent de novo rate is 0; and (2) the parent and daughter de novo rates are the same. The estimates at most CpG sites in either case vary from 0.02 to 0.24, and are 1 or close to 0 at a few sites; this variation may be due to either true variation in the de novo rates or an artefact of the estimation

methods. Both Laird et al. (2004) and Genereux et al. (2005) analysed data collected from the *FMR1* locus; they also provided the basic framework for our modelling of the transmission process. Hence, we will look at those two methods in greater detail in Chapter 2. Issues related to rate estimation also include: (1) how to account for experimental errors (Genereux et al., 2008), which turned out to have a large impact on inference but have been generally ignored in existing analyses of methylation patterns; (2) how to estimate directly variation in the rates; and (3) how to account for temporal stationarity, an assumption underlying many existing methods, in a more flexible manner. We address these issues in Chapters 2 and 3.

2. Do methyltransferases exhibit processivity, meaning that the enzymes bind to the molecule and move along it to methylate cytosines, perhaps unidirectionally from the 5' end of the daughter strand to its 3' end? There is evidence that Dnmt1, mostly a maintenance methyltransferase, acts in this way at least *in vitro* (Vilkaitis et al., 2005; Goyal et al., 2006). Goyal et al. (2006) further employed a random walk model for Dnmt1 and estimated the "diffusion" distance to be about 6000 bp, or about 566 CpGs for an average physical distance of 10.6 bp. Our observation of clustering of hemimethylated CpG dyads in the *FMR1* data suggests the possibility of quantifying processivity of methylation enzymes using the *in vivo* double-stranded methylation data. We will address this question in Chapter 4.

Chapter 2

A MULTI-SITE MODEL FOR ESTIMATION OF RATES OF METHYLATING EVENTS USING HAIRPIN BISULFITE PCR DATA

In this and next chapters, we present models to draw inference about the rates of methylation events. We introduce notation and basic distribution assumptions for the double-stranded methylation data at the beginning, and then review two existing methods. Those methods have been applied to the *FMR1* data summarised in Chapter 1, and have provided general information of the rates. The first method (Laird et al., 2004) considers several CpG sites simultaneously but is limited in its deterministic approach to assigning parent and daughter strands. On the other hand, the second method (Genereux et al., 2005) derives maximum likelihood estimates for the rates assuming that data at CpG sites are independent and the that the transmission process has attained stationarity. Based on the framework of those two methods, we formulate the rate estimation problem as a latent variable problem and develop a multi-site model that infers probabilistically the latent strand type and pre-replication parent strand. The strand type refers to which strand is the parent and which is the daughter. Here we assume that the rates of methylation events are constant across CpG sites. We further assume no experimental errors and ignore temporal stationarity for now. The expectation-maximisation algorithm (Dempster et al., 1977) provides an easy way to find the maximum likelihood estimates in this latent variable problem. We assess the model performance using simulated data and then apply this model to the *FMR1* data. Discussion of the results and of the model is at the end of the chapter.

2.1 Notation and modelling

Consider data collected using hairpin-bisulfite PCR, consisting of methylation states at S CpG sites on N double-stranded DNA molecules. If we denote an unmethylated CpG site by 0, and a methylated CpG by 1, then the data are N pairs of binary strings, each string being of length S. Each pair of binary strings can be thought of as representing the methylation patterns on a parent and daughter strand after a round of DNA replication (Figure 2.1).

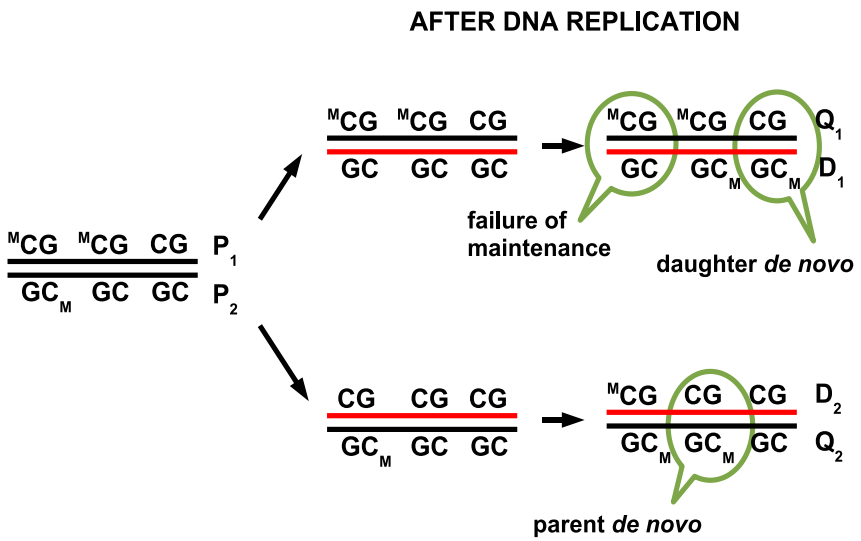

Figure 2.1: The transmission process of DNA methylation patterns in mammalian somatic cells. The two strands in a DNA molecule (left column) become parent strands during DNA replication, each giving rise to a daughter strand (red lines). There is a short, intermediate stage (middle column) during which the methylation patterns on parent strands are the same as those before replication and the daughter strands are completely unmethylated. Subsequently methyl groups are added to cytosines (right column). Failure of maintenance and de novo methylation events can occur. We use $\mathbf{P}_1$ and $\mathbf{P}_2$ to denote the methylation patterns on pre-replication parent strands, $\mathbf{Q}_1$ and $\mathbf{Q}_2$ to denote methylation patterns on post-replication parent strands, and $\mathbf{D}_1$ and $\mathbf{D}_2$ to denote methylation patterns on daughter strands.

As noted in Section 1.1, current technologies are not able to identify the strand type – which strand is the parent and which is the daughter. For simplicity of presentation, however, we assume initially that this information is available, and let $\mathbf{Q}_i$ and $\mathbf{D}_i$ denote the patterns of methylation on the parent strand and daughter strand, respectively. $\mathbf{Q}_i$ and $\mathbf{D}_i$ are each potentially imperfect copies of the patterns of methylation on the unobserved pre-replication parent strand, which we denote $\mathbf{P}_i$. Differences between $\mathbf{P}_i$ and $\mathbf{D}_i$ can arise due to failure of maintenance, or de novo methylation on the daughter strand; differences between $\mathbf{P}_i$ and $\mathbf{Q}_i$ can arise due to de novo methylation on the parent strand. We assume here and in Chapter 3 that these three types of event occur independently of one another, and independently across individuals and across sites. We denote the probabilities of these events at site j by $1-\mu$, δ_d and δ_p respectively. Thus,

$$\Pr(D_{ij}=0|P_{ij}=1)=1-\mu \quad \text{(failure of maintenance)}, \tag{2.1.1}$$

$$\Pr(Q_{ij}=1|P_{ij}=0)=\delta_p \quad \text{(de novo methylation on parent)}, \tag{2.1.2}$$

$$\Pr(D_{ij}=1|P_{ij}=0)=\delta_d \quad \text{(de novo methylation on daughter)}, \tag{2.1.3}$$

where P_{ij}, Q_{ij} and D_{ij} are elements of vectors $\mathbf{P}_i$, $\mathbf{Q}_i$ and $\mathbf{D}_i$, respectively. We are interested in estimating failure of maintenance and de novo methylation rates.

Although recent publications (Métivier et al., 2008; Kangaspeska et al., 2008) suggest the possibility that transcriptionally *active* loci can have very rapid changes in methylation patterns which may be due to active removal of methyl groups from the template DNA, there is no evidence so far that this active removal occurs at *inactive* loci in leukocytes, the cell type from which the *FMR1* data were collected. Hence, consistent with the models in Laird et al. (2004) and Genereux et al. (2005), we assume that, if the parent strand is methylated before replication, then it will also be methylated after replication:

$$\Pr(Q_{ij}=1|P_{ij}=1)=1. \tag{2.1.4}$$

Equations (2.1.1)–(2.1.4), together with the assumption that events occur independently of one another across sites, determine the conditional distribution of $(\mathbf{Q}_i, \mathbf{D}_i)$ given $\mathbf{P}_i$. Specifically, given $\mathbf{P}_i$, the elements of $\mathbf{Q}_i$ and $\mathbf{D}_i$ are independent, with probability $h_\lambda(q, d; p) = \Pr\left((Q_{ij}, D_{ij}) = (q, d) | P_{ij} = p\right)$ being determined by μ, δ_d and δ_p as in Table 2.1.

Table 2.1: Conditional probabilities of methylation events in the i-th sequence at the j-th site, $h_\lambda(q, d; p) = \Pr\left((Q_{ij}, D_{ij}) = (q, d) | P_{ij} = p\right)$ under the simple multi-site model. Here P_{ij} Q_{ij} and D_{ij} are methylation states on the pre-replication parent strand, the post-replication parent strand and the daughter strand, respectively. Additionally, 0 represents unmethylated and 1 methylated.

$(Q_{ij}, D_{ij}) = (q, d)$	$P_{ij} = p$	$h_\lambda(q, d; p)$
(0, 0)	1	0
(0, 1)	1	0
(1, 0)	1	$1 - \mu$
(1, 1)	1	μ
(0, 0)	0	$(1 - \delta_p)(1 - \delta_d)$
(0, 1)	0	$(1 - \delta_p)\delta_d$
(1, 0)	0	$\delta_p(1 - \delta_d)$
(1, 1)	0	$\delta_p\delta_d$

To complete the specification of the distribution of (Q_{ij}, D_{ij}), we further assume that P_{ij}'s are independent and identically distributed Bernoulli random variables with parameters m_j, where m_j is the methylation probability on the pre-replication parent strand at site j; in other words,

$$\Pr(P_{ij} = 1) = m_j. \tag{2.1.5}$$

2.2 Existing methods for analysing double-stranded methylation data

The earliest analysis of double-stranded methylation data may trace back to Bird (1978), who used restriction enzymes rather than hairpin bisulfite PCR to detect the methylation status, and examined the proportion of hemimethylated CpG dyads at a hypermethylated locus. He concluded that this proportion might be less than 2%. The laboratory techniques Bird had used allowed him to focus on failure of maintenance events. Using the proportion of hemimethylated dyads of 2% and the methylation density of 89%, we may calculate the failure of maintenance rate as follows

$$1 - \mu = \Pr(D_{ij} = 0 | P_{ij} = 1) = \frac{\Pr(P_{ij} = 1, D_{ij} = 0)}{\Pr(P_{ij} = 1)} = \frac{2\%}{89\%} = 2.2\%. \tag{2.2.1}$$

This proportion can be taken as an upper bound on the failure of maintenance rate in Bird's experiment.

With the invention of hairpin bisulfite PCR, Laird et al. (2004) collected the first few double-stranded sequences from the *FMR1* locus. Under the assumption of no de novo events on the parent strand, these data enabled the authors to investigate not only failure of maintenance, but also daughter de novo events. The authors obtained the range and mean for each rate by assigning half of the strands to be parent strands and half to be daughter strands and then counting events of the two types. This method is multi-site in essence, although the assignment of strand type is deterministic.

Genereux et al. (2005) took a different approach to part of the *FMR1* data analysed here, aiming at studying the rates *and* their variability across CpG sites. The authors developed a single-site maximum likelihood (ML) method to obtain site-specific estimates. They first summarised the data as observed counts of methylated, hemimethylated and unmethylated dyads, denoted as (M, H, U), for each CpG site. They then found the likelihood assuming that data at each site are independent and that (M, H, U) at each site forms a multinomial sample with parameters (p_M, p_H, p_U), where $p_M + p_H + p_U = 1$. Further assuming the methylation process has attained

temporal stationarity – stationarity over cell division – they derived a set of equations from which the maximum likelihood estimates of maintenance and de novo rates at each site could be solved:

$$p_M = \frac{\mu(\delta_p + \delta_d - \delta_p\delta_d) + \delta_p\delta_d}{1 + \delta_p + \delta_d - \mu}, \tag{2.2.2}$$

$$p_H = \frac{2(\delta_p + \delta_d - \delta_p\delta_d)(1 - \mu)}{1 + \delta_p + \delta_d - \mu}, \tag{2.2.3}$$

$$p_U = 1 - p_M - p_U = \frac{(1 - \delta_p)(1 - \delta_d)(1 - \mu)}{1 + \delta_p + \delta_d - \mu}. \tag{2.2.4}$$

The stationarity equations (2.2.2) through (2.2.4) of the three unknowns have only two degrees of freedom. The main task then is to separate the two types of de novo events. The double-stranded sequence data indicate that de novo events must occur sometimes on the daughter strands (Genereux et al., 2005). However, those data contain no direct information on the occurrence of parent strand de novo events. Therefore, Genereux et al. (2005) imposed two alternate constraints: $\delta_p = 0$ and $\delta_p = \delta_d$. Otto and Walbot (1990) first derived a set of stationarity equations similar to (2.2.2) – (2.2.4), although they assumed that de novo events occur simultaneously on the parent and daughter strands. On the other hand, Pfeifer et al. (1990) obtained differential equations for methylated and unmethylated CpG dinucleotides in single-stranded methylation data. Details of those two methods can be found in Appendix B.

This single-site ML approach in Genereux et al. (2005) is different from the approaches in Bird (1978) and Laird et al. (2004) in two ways: (1) it makes a probabilistic (multinomial) assumption of the data and hence acknowledges the uncertainty in the estimation; and (2) it fully exploits the temporal stationarity assumption, which is both reasonable and useful for modelling the transmission process, reasonable because there is evidence suggesting stability of methylation patterns in somatic cells, and useful because this assumption makes finding equations (2.2.2) – (2.2.4) and solving for the rates possible.

Nonetheless, this method has its limitations. First, the method ignores the multi-site information in the data and thus implicitly assumes independence of data across CpG sites. This is particularly limiting in this context given that the current technology can produce only low-throughput data. The inefficiency in estimation imposes constraints on the kind of inference we can perform. For example, one such limitation is that the two de novo rates are unidentifiable under this approach. Second, the temporal stationarity condition, under which the single-site ML approach and other existing methods (Otto and Walbot, 1990; Pfeifer et al., 1990) are derived, is a rather stringent assumption. This condition, together with the assumption of site independence, gives an estimate of 1 for de novo rates (under either constraint on de novo rates) at sites where there are methylated and hemimethylated dyads but no unmethylated dyads. For now, however, we will ignore the temporal stationarity assumption and focus on constructing a simple multi-site model that effectively uses the multi-site information in the data.

2.3 The multi-site model

The data collected contain double-stranded sequences and each sequence has a top and a bottom strand. Assume that the strand type in a double-stranded sequence is known; that is, we know which strand is the parent strand and which the daughter strand. Consider the case where the top strand is the parent strand and the bottom the daughter strand. Also assume that methylation events occur independently across sites. Then, at each CpG site, the methylation probability, along with maintenance and de novo rates, gives rise to the methylation states on pre- and post-replication parent strands and those on the daughter strand. Recall that $\mathbf{P}_i$, $\mathbf{Q}_i$ and $\mathbf{D}_i$ denote methylation patterns on the i-th pre-replication parent strand, the post-replication parent strand and the daughter strand, respectively. We further denote the i-th top and bottom strand by $\mathbf{x}_i$ and $\mathbf{y}_i$, respectively. We will also use brackets to denote an ordered set of variables, whereas curly brackets an unordered set. For example,

the order matters in $(\mathbf{x}_i, \mathbf{y}_i)$ but not in $\{\mathbf{x}_i, \mathbf{y}_i\}$. The probability of observing this double-stranded sequence with the known strand type is then

$$
\begin{aligned}
d_\lambda(\mathbf{x}_i, \mathbf{y}_i) &\equiv \Pr((\mathbf{Q}_i, \mathbf{D}_i) = (\mathbf{x}_i, \mathbf{y}_i); \lambda) && (2.3.1)\\
&= \prod_{j=1}^{S} \Pr((Q_{ij}, D_{ij}) = (x_{ij}, y_{ij})) \quad \text{(independence of methylation events)} && (2.3.2)\\
&= \prod_{j=1}^{S} \sum_{z_{ij}=0}^{1} \Pr((Q_{ij}, D_{ij}) = (x_{ij}, y_{ij}) | P_{ij} = z_{ij}) \Pr(P_{ij} = z_{ij}) && (2.3.3)\\
&= \prod_{j=1}^{S} \sum_{z_{ij}=0}^{1} h_\lambda(x_{ij}, y_{ij}; z_{ij}) m_j^{z_{ij}} (1 - m_j)^{1-z_{ij}}, && (2.3.4)
\end{aligned}
$$

in which z_{ij} is the value of random variable P_{ij} and $h_\lambda(x_{ij}, y_{ij}; z_{ij})$ can be computed from Table 2.1. The probability for the case where the top strand is the daughter strand and the bottom the parent strand, $d_\lambda(\mathbf{y}_i, \mathbf{x}_i)$, can be defined and calculated similarly.

The strand type, however, is unknown in the observed data. When the top and bottom strands in a sequence are different, there are always two possibilities: the top strand is the parent strand and the bottom the daughter strand, and vice versa. The probability of observing each sequence needs to sum over these two possibilities. Let the parameters be $\lambda = \{\mu, \delta_p, \delta_d\}$. The likelihood of λ given the data is then

$$
\begin{aligned}
L(\lambda; \{\mathbf{x}, \mathbf{y}\}) &= \prod_{i=1}^{N} \Pr(\{\mathbf{Q}_i, \mathbf{D}_i\} = \{\mathbf{x}_i, \mathbf{y}_i\}; \lambda) && (2.3.5)\\
&= \prod_{i=1}^{N} \Pr\Big((\mathbf{Q}_i, \mathbf{D}_i) = (\mathbf{x}_i, \mathbf{y}_i) \text{ or } (\mathbf{Q}_i, \mathbf{D}_i) = (\mathbf{y}_i, \mathbf{x}_i); \lambda \Big) && (2.3.6)\\
&= \prod_{i=1}^{N} \Big(\Pr((\mathbf{Q}_i, \mathbf{D}_i) = (\mathbf{x}_i, \mathbf{y}_i); \lambda) + \mathbf{1}(\mathbf{x}_i \neq \mathbf{y}_i) \Pr((\mathbf{Q}_i, \mathbf{D}_i) = (\mathbf{y}_i, \mathbf{x}_i); \lambda) \Big) && (2.3.7)\\
&\equiv \prod_{i=1}^{N} \Big(d_\lambda(\mathbf{x}_i, \mathbf{y}_i) + \mathbf{1}(\mathbf{x}_i \neq \mathbf{y}_i) d_\lambda(\mathbf{y}_i, \mathbf{x}_i) \Big), && (2.3.8)
\end{aligned}
$$

where $\mathbf{1}()$ is the indicator function, taking on value 1 if the condition in the brackets is met and 0 otherwise.

2.4 Maximum likelihood estimation

2.4.1 Introduction to the expectation-maximisation (EM) algorithm

Since the likelihood function (2.3.8) involves summation inside multiplication, the usual approach of taking derivatives to find the maximum likelihood estimates is not straightforward. The expectation-maximisation (EM) method (Dempster et al., 1977), however, provides an easy alternative.

The idea of the EM method is to introduce latent variables so that these latent variables and the observed data constitute the "complete data", the likelihood function of which is easy to manipulate. For example, the likelihood function of the complete data can be written as multiplication. The usual technique of setting the derivatives to zero to maximise the function is then applicable. This method updates the complete and observed data likelihood as well as the estimates iteratively, until the observed data likelihood converges. The theory behind this method guarantees that maximising the complete data likelihood leads to maximisation of the likelihood of the observed data and that the estimates at convergence are then maximum likelihood estimates.

Specifically, denote the observed data by $\mathbf{z}_o$ with likelihood function $L(\lambda; \mathbf{z}_o)$ and complete data by $\mathbf{z}_c$ with likelihood function $L(\lambda; \mathbf{z}_c)$, where λ is the parameter vector of length k. The complete data $\mathbf{z}_c$ contain the observed data in addition to latent variables, and are only indirectly observable through the observed data. The EM algorithm then involves, at the t-th iteration, the following two steps:

THE EXPECTATION (E) STEP, in which we compute the expected complete data log likelihood given observed data $\mathbf{z}_o$ and current parameter values $\lambda^{(t)}$

$$l(\lambda; \lambda^{(t)}, \mathbf{z}_o) = E(\log L(\lambda; \mathbf{z}_c) | \mathbf{z}_o, \lambda^{(t)}), \tag{2.4.1}$$

where the expectation is taken over the complete data.

THE MAXIMISATION (M) STEP, in which we maximise $l(\lambda; \lambda^{(t)}, \mathbf{z}_o)$ with respect to λ to obtain new estimates $\lambda^{(t)}$.

The algorithm cycles through the E and M steps and climbs up the observed data likelihood surface until it reaches a local maximum. This algorithm carries the risk of missing the global maximum or other local maxima. Running the algorithm from different starting values usually helps identifying multimodality of the observed data likelihood surface.

When the complete data likelihood $L(\lambda; \mathbf{z}_c)$ has the form of a regular exponential family

$$L(\lambda; \mathbf{z}_c) = b(\mathbf{z}_c) \exp(\lambda r(\mathbf{z}_c)^T)/a(\lambda) \tag{2.4.2}$$

where $r(\mathbf{z}_c)$ is a $1 \times k$ vector of sufficient statistics based on the complete data for the parameter vector λ, maximising

$$l(\lambda; \lambda^{(t)}, \mathbf{z}_o) = E(-\log a(\lambda) + \log b(\mathbf{z}_c) + \lambda r(\mathbf{z}_c)^T | \mathbf{z}_o, \lambda^{(t)}) \tag{2.4.3}$$

is equivalent to maximising

$$E(-\log a(\lambda) + \lambda r(\mathbf{z}_c)^T | \mathbf{z}_o, \lambda^{(t)}) = -\log a(\lambda) + E(\lambda r(\mathbf{z}_c)^T | \mathbf{z}_o, \lambda^{(t)}). \tag{2.4.4}$$

That means, we can simplify the E step by estimating the expectation only of the sufficient statistics $r(\mathbf{z}_c)$ given the observed data and current parameter values. Indeed, in our case and many other problems of this type, we try to identify "missing" variables such that the complete data likelihood has the form of an exponential family and that we need to evaluate just the expectation of the sufficient statistics in the E step.

2.4.2 The EM algorithm for estimating λ

The key to an effective EM algorithm is therefore identification of latent variables that facilitate working with the complete data likelihood function. Our description

of the biological process illustrated in Figure 2.1 suggests that, had we known the methylation states on the parent strand before replication and the strand type in the observed double-stranded data, we would be able to determine the type of methylation event. Hence, the latent variable in the multi-site model is $\mathbf{H}_{ij}(q,d,p) \equiv \mathbf{1}(Q_{ij} = q, D_{ij} = d, P_{ij} = p)$, where $\mathbf{1}()$ is the indicator function. Subscripts i and j are omitted from $\mathbf{H}()$ when there is no ambiguity. We also use $\mathbf{H}$ to denote the complete data as a whole. Its likelihood can then be written as

$$
\begin{aligned}
L(\lambda;\mathbf{H}) =& \prod_{i=1}^{N}\prod_{j=1}^{S}\left\{\left(\Pr\left((Q_{ij},D_{ij},P_{ij}) = (x_{ij},y_{ij},z_{ij})\right)\right)^{\mathbf{H}(x_{ij},y_{ij},z_{ij})}\right\} \\
&\times\left\{\left(\Pr\left((Q_{ij},D_{ij},P_{ij}) = (y_{ij},x_{ij},z_{ij})\right)\right)^{\mathbf{H}(y_{ij},x_{ij},z_{ij})}\right\}^{\mathbf{1}(\mathbf{x}_i\neq\mathbf{y}_i)} \qquad (2.4.5)\\
=& \prod_{i=1}^{N}\prod_{j=1}^{S}\left(h_\lambda(x_{ij},y_{ij};z_{ij})m_j^{z_{ij}}(1-m_j)^{1-z_{ij}}\right)^{\mathbf{H}(x_{ij},y_{ij},z_{ij})} \\
&\times\left(h_\lambda(y_{ij},x_{ij};z_{ij})m_j^{z_{ij}}(1-m_j)^{1-z_{ij}}\right)^{\mathbf{H}(y_{ij},x_{ij},z_{ij})\mathbf{1}(\mathbf{x}_i\neq\mathbf{y}_i)} \qquad (2.4.6)\\
=& \prod_{i=1}^{N}\prod_{j=1}^{S} m_j^{z_{ij}}(1-m_j)^{1-z_{ij}}h_\lambda(x_{ij},y_{ij};z_{ij})^{\mathbf{H}(x_{ij},y_{ij},z_{ij})} \\
&\times h_\lambda(y_{ij},x_{ij};z_{ij})^{\mathbf{H}(y_{ij},x_{ij},z_{ij})\mathbf{1}(\mathbf{x}_i\neq\mathbf{y}_i)}, \qquad (2.4.7)
\end{aligned}
$$

and the complete data log likelihood is

$$
\begin{aligned}
\log L(\lambda;\mathbf{H}) =& \sum_{i=1}^{N}\sum_{j=1}^{S} z_{ij}\log m_j + \left(NS - \sum_{i=1}^{N}\sum_{j=1}^{S} z_{ij}\right)\log(1-m_j) \\
&+ \sum_{i=1}^{N}\sum_{j=1}^{S}\mathbf{H}(x_{ij},y_{ij},z_{ij})\log h_\lambda(x_{ij},y_{ij};z_{ij}) \\
&+ \sum_{i=1}^{N}\sum_{j=1}^{S}\mathbf{1}(\mathbf{x}_i\neq\mathbf{y}_i)\mathbf{H}(y_{ij},x_{ij},z_{ij})\log h_\lambda(y_{ij},x_{ij};z_{ij}). \qquad (2.4.8)
\end{aligned}
$$

The complete data likelihood is based on a multinomial distribution and the complete data are taken to be the sufficient statistics. The EM algorithm starts at some initial values of the parameters and iterates through the following E and M steps.

THE E STEP We compute the expectation of $\mathbf{H}()$ given the observed data and current parameter values from the t-th iteration as the following:

$$K_{ij}^{(t)}(q,d,p;\lambda) \equiv E(\mathbf{H}_{ij}(q,d,p)|\{\mathbf{x},\mathbf{y}\},\lambda^{(t)}) \tag{2.4.9}$$

$$= \Pr((Q_{ij},D_{ij},P_{ij}) = (q,d,p)|\{\mathbf{x},\mathbf{y}\},\lambda^{(t)}) \tag{2.4.10}$$

$$= \Pr(Q_{ij}=q, D_{ij}=d|\{\mathbf{x},\mathbf{y}\},\lambda^{(t)}) \Pr(P_{ij}=p|Q_{ij}=q, D_{ij}=d,\lambda^{(t)}). \tag{2.4.11}$$

The first term in (2.4.11) infers the strand type given the data and current estimates, whereas the second term the methylation states on the pre-replication parent strand. Hence, the E step can be broken down into two sub-steps:

1. Calculating the expectation of the strand type given the data and current estimates of the parameters.

$$\Pr((\mathbf{Q}_i,\mathbf{D}_i) = (\mathbf{x}_i,\mathbf{y}_i)|\{\mathbf{x},\mathbf{y}\},\lambda^{(t)}) = \frac{d_\lambda^{(t)}(\mathbf{x}_i,\mathbf{y}_i)}{d_\lambda^{(t)}(\mathbf{x}_i,\mathbf{y}_i) + \mathbf{1}(\mathbf{x}_i \neq \mathbf{y}_i) d_\lambda^{(t)}(\mathbf{y}_i,\mathbf{x}_i)} \tag{2.4.12}$$

   where $d_\lambda^{(t)}(\mathbf{x}_i,\mathbf{y}_i)$ is defined in (2.3.4) and $d_\lambda^{(t)}(\mathbf{y}_i,\mathbf{x}_i)$ defined similarly.

2. Calculating the expectation of methylation states on the pre-replication parent strand $\mathbf{P}_i$ given the strand type and current parameter estimates.

$$\Pr(P_{ij}=1|Q_{ij}=q, D_{ij}=d,\lambda^{(t)}) = \frac{h_\lambda^{(t)}(q,d;1)\, m_j^{(t)}}{h_\lambda^{(t)}(q,d;1)\, m_j^{(t)} + h_\lambda^{(t)}(q,d;0)\,(1-m_j^{(t)})}. \tag{2.4.13}$$

THE M STEP This step maximises the expected complete data log likelihood with respect to the parameters. The resulting estimates are then used as current parameter values in the next EM iteration. Take the maintenance rate μ for example. We demonstrate below the procedure to obtain its estimate. Similar procedures can be applied to get the estimates of the de novo rates.

The expected complete data log likelihood can be written as a function of μ as the following

$$\begin{aligned}
E(\log L(\lambda, \mathbf{H})|\{\mathbf{x}, \mathbf{y}\}, \lambda^{(t)}) =& \log(1-\mu) \sum_{i=1}^{N} \sum_{j=1}^{S} K_{ij}^{(t)}(x_{ij}, y_{ij}, 1; \lambda) \mathbf{1}(x_{ij} = 1, y_{ij} = 0) \\
&+ \log \mu \sum_{i=1}^{N} \sum_{j=1}^{S} K_{ij}^{(t)}(x_{ij}, y_{ij}, 1; \lambda) \mathbf{1}(x_{ij} = y_{ij} = 1) \\
&+ \log(1-\mu) \sum_{i=1}^{N} \sum_{j=1}^{S} \mathbf{1}(\mathbf{x}_i \neq \mathbf{y}_i) K_{ij}^{(t)}(y_{ij}, x_{ij}, 1; \lambda) \\
&\times \mathbf{1}(x_{ij} = 0, y_{ij} = 1) + \log \mu \sum_{i=1}^{N} \sum_{j=1}^{S} \mathbf{1}(\mathbf{x}_i \neq \mathbf{y}_i) \\
&\times K_{ij}^{(t)}(y_{ij}, x_{ij}, 1; \lambda) \mathbf{1}(x_{ij} = y_{ij} = 1) + \text{const.} \qquad (2.4.14) \\
\equiv& N_{\lambda}^{(t)}(1, 0; 1) \log(1-\mu) + N_{\lambda}^{(t)}(1, 1; 1) \log \mu, \qquad (2.4.15)
\end{aligned}$$

where

$$\begin{aligned}
N_{\lambda}^{(t)}(q, d; p) =& \sum_{i=1}^{N} \sum_{j=1}^{S} K_{ij}^{(t)}(x_{ij}, y_{ij}, p; \lambda) \mathbf{1}(x_{ij} = q, y_{ij} = d) \\
&+ \sum_{i=1}^{N} \sum_{j=1}^{S} \mathbf{1}(\mathbf{x}_i \neq \mathbf{y}_i) K_{ij}^{(t)}(y_{ij}, x_{ij}, p; \lambda) \mathbf{1}(x_{ij} = d, y_{ij} = q). \qquad (2.4.16)
\end{aligned}$$

Recall that $K_{ij}^{(t)}(q, d, p; \lambda)$ is the joint probability of the methylation states at site j on the i-th pre- and post-replication parent strands and the daughter strand, conditional on the data and estimates from the t-th iteration. Then $N_{\lambda}^{(t)}(q, d; p)$ is the expected number of transitions from p to (q, d), also conditional on the data and estimates from the t-th iteration; in other words, it is the conditional expectation of a certain methylation event, defined by $(q, d; p)$, across sites and sequences.

Differentiating (2.4.15) with respect to μ and setting it to 0, we have

$$\frac{\partial}{\partial \mu} E(\log L_c|\{\mathbf{x}, \mathbf{y}\}, \lambda^{(t)}) = -\frac{N_{\lambda}^{(t)}(1, 0; 1)}{1-\mu} + \frac{N_{\lambda}^{(t)}(1, 1; 1)}{\mu} = 0, \qquad (2.4.17)$$

which gives

$$\mu^{(t+1)} = \frac{N_\lambda^{(t)}(1,1;1)}{N_\lambda^{(t)}(1,0;1) + N_\lambda^{(t)}(1,1;1)}. \tag{2.4.18}$$

Intuitively, this procedure counts the maintenance events on each sequence, treating the top strand as the parent strand and the bottom the daughter. When the top and bottom strands are different, we swap the two strands and add the maintenance events to the existing number.

The same intuition applies to the de novo rates and methylation probabilities. The explicit forms of their estimates are then:

$$\delta_p{}^{(t+1)} = \frac{N_\lambda^{(t)}(1,0;0) + N_\lambda^{(t)}(1,1;0)}{\sum_{q=0}^{1}\sum_{d=0}^{1} N_\lambda^{(t)}(q,d;0)}; \tag{2.4.19}$$

$$\delta_d{}^{(t+1)} = \frac{N_\lambda^{(t)}(0,1;0) + N_\lambda^{(t)}(1,1;0)}{\sum_{q=0}^{1}\sum_{d=0}^{1} N_\lambda^{(t)}(q,d;0)}; \tag{2.4.20}$$

and

$$m_j^{(t+1)} = \frac{\sum_{q=0}^{1}\sum_{d=0}^{1} N_{\lambda,j}^{(t)}(q,d;1)}{N}. \tag{2.4.21}$$

2.5 Simulation studies

To validate the EM algorithm, we simulated a data set (labelled as A1 hereinafter) of 169 sequences and 22 sites, the same size as the *FMR1* data at sites 1–22. In simulation we used the EM estimates from the *FMR1* data at sites 1–22. Specifically, we set $\mu = 0.95$, $\delta_p = 0.001$ and $\delta_d = 0.15$. The methylation probability m_j followed a Beta($\alpha = 7.2, \beta = 1.8$) distribution, which has mean 0.8 and standard deviation 0.13. We ran the EM algorithm from different starting points and set the stopping criterion to be 10^{-6} for the difference between the observed data log likelihoods from the $t-1$-st iteration to the t-th iteration. The results from multiple runs are consistent and in good agreement with the simulation truth. We list in Table 2.2 results from the run that gives the highest observed data likelihood.

To investigate the performance of the EM algorithm under different scenarios, we considered 6 models and simulated 5 data sets for each model. Again, each data set contains 169 sequences and 22 sites.

- Data sets (A1)-(A5): $\mu = 0.95$, $\delta_p = 0.001$, $\delta_d = 0.15$;
- Data sets (B1)-(B5): $\mu = 0.95$, $\delta_p = 0.15$, $\delta_d = 0.001$;
- Data sets (C1)-(C5): $\mu = 0.99$, $\delta_p = 0.001$, $\delta_d = 0.15$;
- Data sets (D1)-(D5): $\mu = 0.99$, $\delta_p = 0.15$, $\delta_d = 0.001$;
- Data sets (E1)-(E5): $\mu = 0.95$, $\delta_p = \delta_d = 0.001$;
- Data sets (F1)-(F5): $\mu = 0.95$, $\delta_p = \delta_d = 0.075$.

We ran the EM algorithm multiple times from different starting values on each data set and listed the results from the run with the highest observed data log likelihood in Table 2.2. The EM algorithm generally recovers the truth well. In cases where the truth is very small, for example 0.001, the EM estimates are at least close to the 0 boundary. Nevertheless, there is some difference among these models: Model C tends to produce data with multimodal likelihood surfaces (Table 2.3). Four out of the five data sets under this model have multimodal likelihood surfaces. For example, the log likelihood surface for δ_p and δ_d from data set (C2), as shown in Figure 2.2, has clearly two peaks, whereas the second peak in Figure 2.3 from data set (C1) is not so visible.

To better understand the confounding among $1 - \mu$, δ_p and δ_d, we considered just one CpG site and simulated data under Models A and C, fixing $m = 0.8$. Each data set contains 100 "sequences", or, since there is only one site, 100 CpG dyads. Counts of methylated, unmethylated, and hemimethylated dyads is in Table 2.4. The log likelihood surface for $1 - \mu$ versus δ_p and that for δ_p versus δ_d are plotted in Figure

Table 2.2: Results from the EM algorithm under the multi-site model for five data sets under each of six simulation models. See text for simulation models. The EM algorithm is run multiple times for each data set, but results only from the run giving the highest observed data likelihood are reported here.

	Model A			Model B		
	$1-\mu$	δ_p	δ_d	$1-\mu$	δ_p	δ_d
Truth	0.05	0.001	0.15	0.05	0.15	0.001
Data set 1	0.056	0.005	0.127	0.046	0.138	2.07e−83
Data set 2	0.047	1.47e−7	0.151	0.053	0.169	7.92e−60
Data set 3	0.041	0.058	0.123	0.072	0.105	0.002
Data set 4	0.051	7.75e−6	0.149	0.048	0.192	0.002
Data set 5	0.054	0.032	0.140	0.052	0.142	2.57e−61
	Model C			Model D		
	$1-\mu$	δ_p	δ_d	$1-\mu$	δ_p	δ_d
Truth	0.01	0.001	0.15	0.01	0.15	0.001
Data set 1	0.008	0.002	0.176	0.025	0.118	0.002
Data set 2	0.007	0.017	0.162	0.015	0.141	6.34e−89
Data set 3	0.010	2.81e−7	0.129	0.009	0.141	0.004
Data set 4	0.016	8.40e−8	0.144	0.018	0.113	0.002
Data set 5	0.014	4.46e−7	0.156	0.011	0.134	1.05e−54
	Model E			Model F		
	$1-\mu$	δ_p	δ_d	$1-\mu$	δ_p	δ_d
Truth	0.05	0.001	0.001	0.05	0.075	0.075
Data set 1	0.045	4.07e−4	0.002	0.045	0.098	0.077
Data set 2	0.047	3.38e−7	0.002	0.049	0.105	0.060
Data set 3	0.047	0.005	1.50e−6	0.046	0.051	0.100
Data set 4	0.049	1.76e−7	7.32e−84	0.030	0.134	0.069
Data set 5	0.057	7.44e−6	0.002	0.059	0.044	0.071

Table 2.3: EM estimates and corresponding observed data log likelihood from different starting values for data sets simulated under Model C.

	Model C			
	$1-\mu$	δ_p	δ_d	$\log L$
Truth	0.01	0.001	0.15	
Data set 1	0.008	0.002	0.176	−1717.666
	0.008	0.117	0.078	−1720.582
Data set 2	0.007	0.017	0.162	−2156.289
	0.006	0.138	0.052	−2156.939
Data set 3	0.010	2.81e−7	0.129	−2027.558
Data set 4	0.016	8.40e−8	0.144	−2252.797
	0.020	0.073	0.068	−2258.321
Data set 5	0.014	4.46e−7	0.156	−2129.696
	0.019	0.081	0.066	−2131.354
	0.000	0.115	0.115	−2136.672

Table 2.4: Data simulated for one site under Models A and C. $m = 0.8$ and $N = 100$. See text for values of $1-\mu$, δ_p and δ_d.

	M	U	H (parent methylated)	H (daughter methylated)
Model A	78	13	4	5
Model C	78	19	0	3

2.4. Except for the log likelihood surface for δ_p versus δ_d under Model C, other log likelihood surfaces are unimodal, although the peaks can be quite far from the truth. Comparison with the good estimation results in Table 2.2 under Model A and with the estimates for $1-\mu$ under Model C in the same table suggests that data from

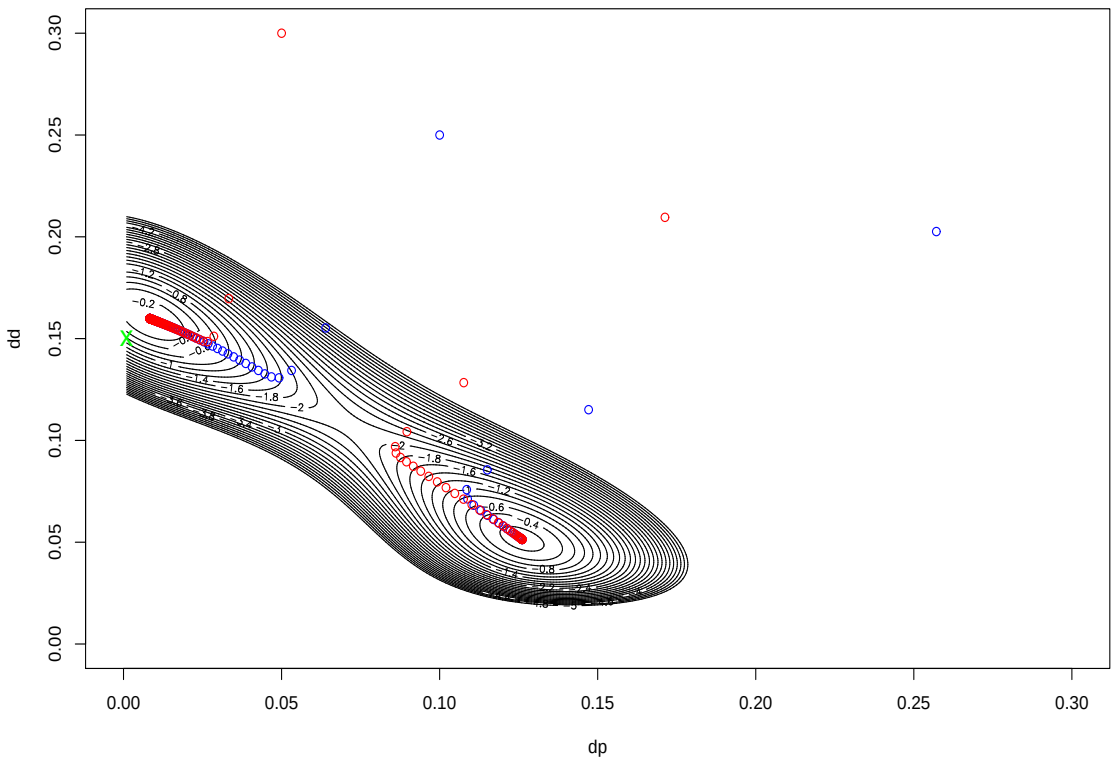

Figure 2.2: Log likelihood surface for δ_p and δ_d from simulated data set (C2). In analysis $1 - \mu$ and m are fixed to the truth. The surface is re-scaled so that the maximum is at 0 and the contours are shown to the 5 log likelihood units below 0. The green cross indicates the simulation truth. Red and blue circles are estimates from EM iterations starting at different values. There are two peaks in this plot with very similar log likelihood.

a large number of sites can significantly increase the sharpness of the log likelihood surface and move its peak closer to the truth. However, having many sites does not help too much when the log likelihood surface at one site is already multimodal, as is the case with the two de novo rates under Model C.

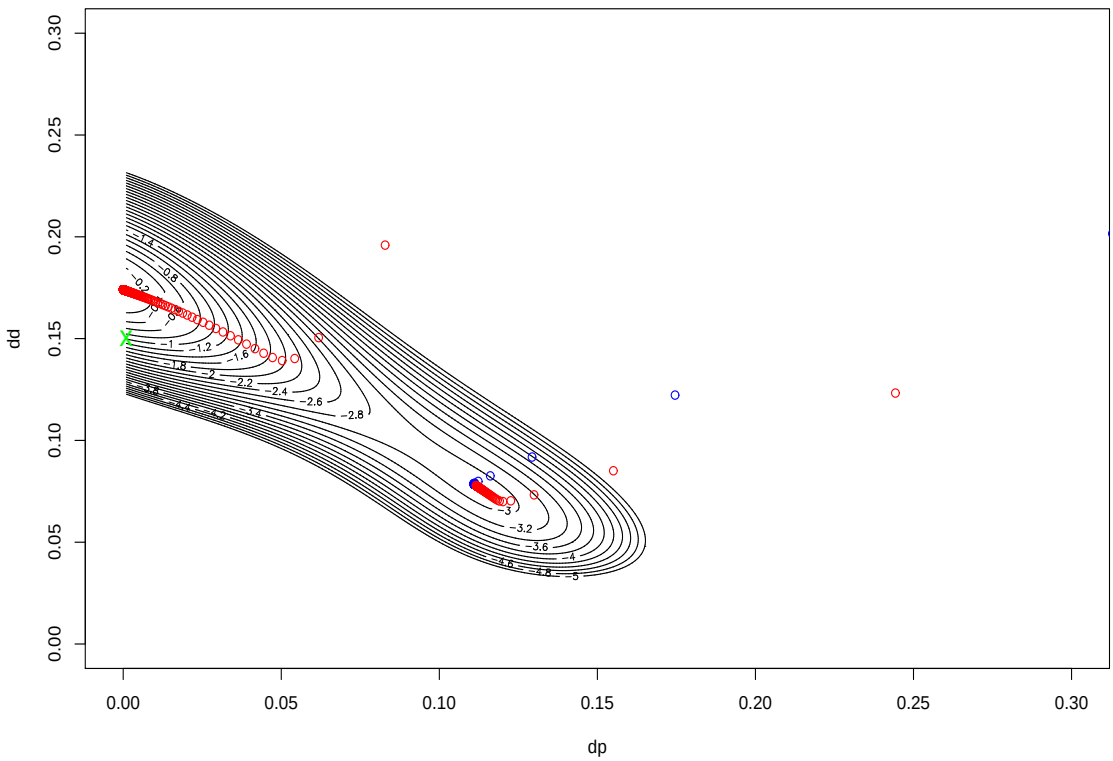

Figure 2.3: Log likelihood surface for δ_p and δ_d from simulated data set (C1). In analysis $1-\mu$ and m are fixed to the truth. The surface is re-scaled so that the maximum is at 0 and the contours are shown to 5 log likelihood units below 0. The green cross indicates the simulation truth. Red and blue circles are estimates from EM iterations starting at different values. The area around $(0.1, 0.1)$ contains the second peak, although not quite visible due to the resolution of contour lines.

2.6 *Analysis of the FMR1 data*

Applying the multi-site model to the *FMR1* data sets at sites 1–22 and at 25–52 separately, we obtain the results via the EM algorithm and list these point estimates in Table 2.5. We further investigate variability in these estimates using profile likelihood functions.

In high-dimensional inference problems, profile likelihood functions are often used

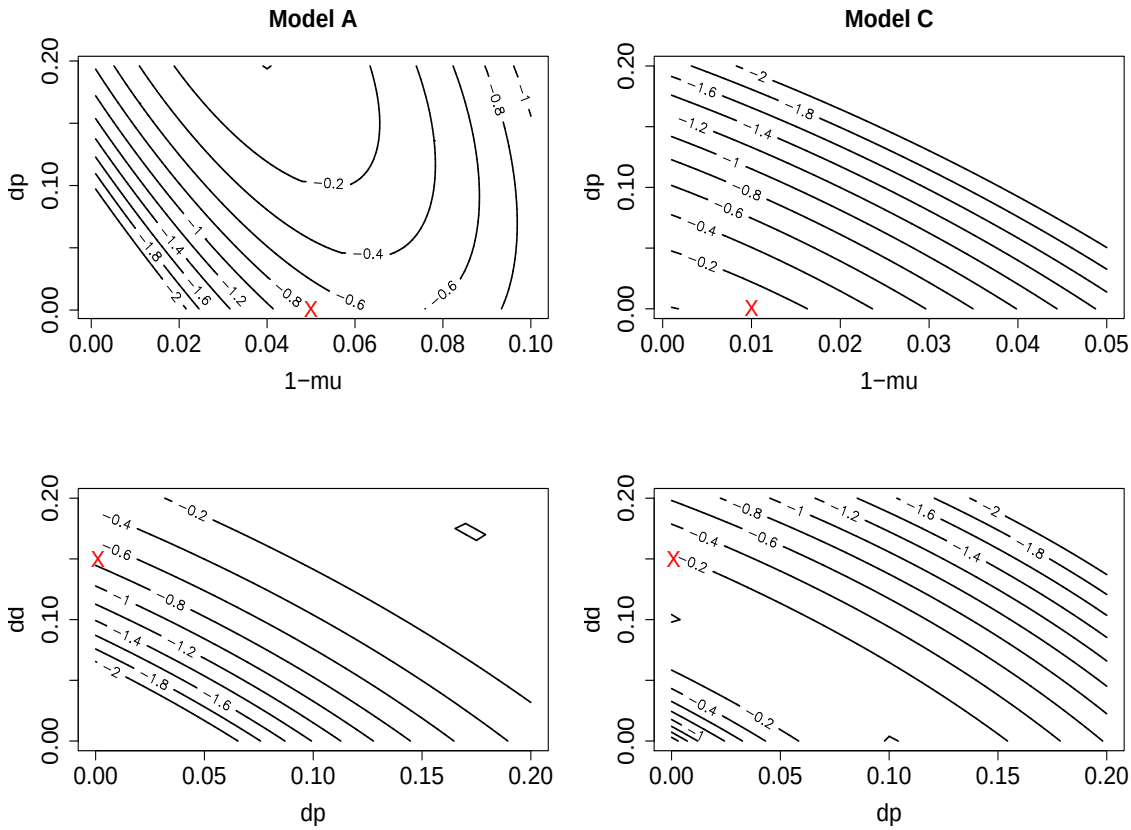

Figure 2.4: Log likelihood surfaces for data simulated for one site under Models A and C. For both data sets, the true value of m is 0.8. The log likelihood surface is evaluated for $1-\mu$ and δ_p with δ_d and m fixed to the truth, or for δ_p and δ_d with $1-\mu$ and m fixed to the truth. Red crosses indicate the simulation truth. The surface is re-scaled so that the maximum is at 0 and the contours are shown to 2 log likelihood units below 0.

for exploration of one parameter at a time. For example, let failure of maintenance rate $1-\mu$ be the parameter of interest and all other parameters be nuisance parameters η. The profile likelihood function for $1-\mu$ is

$$L_p(1-\mu) = \sup_{\eta} L(1-\mu, \eta), \tag{2.6.1}$$

where $L(1-\mu, \eta)$ is the likelihood function of all the parameters. In other words, the above profile likelihood function is the likelihood function under the full model

Table 2.5: Results under the simple multi-site model via the expectation-maximisation algorithm for the *FMR1* data at sites 1–22 and at sites 25–52. The 95% confidence intervals in brackets are obtained from the profile likelihood functions.

	Sites 1–22	Sites 25–52
$1-\mu$	0.05 (0.044, 0.061)	0.07 (0.062, 0.087)
δ_p	1.93×10^{-4} (0, 0.033)	0.02 (0, 0.06)
δ_d	0.14 (0.101, 0.176)	0.08 (0.055, 0.107)

maximised with $1-\mu$ fixed to be a particular value. It has been shown that a profile likelihood function can be used in many ways similar to the usual likelihood function. As pointed out by Murphy and van der Vaart (2000), three aspects are particularly useful in inference. We illustrate these points with the failure of maintenance rate example.

1. The maximiser of the profile likelihood function $L_p(1-\mu)$ is the maximum likelihood estimator of $1-\mu$.

2. One can still carry out hypothesis testing. The likelihood ratio test statistic for $H_0 : \mu = \mu_0$ is

$$2(L_p(1-\mu) - \sup_{\mu} L_p(1-\mu)), \tag{2.6.2}$$

   which has a χ^2_1 distribution asymptotically when μ_0 is not on the boundary. The acceptance region for a test of size α can then be inverted to construct a confidence interval at $1-\alpha$ level.

3. One may also compute the curvature at the maximum on a profile likelihood curve to approximate the variability in the maximum likelihood estimate, in just the same way the second derivative is calculated for a full likelihood function.

Detailed discussion of profile likelihood can also be found in Severini (2000).

Here we compute the profile log likelihood for each methylation event rate and plot it in Figure 2.5. Each profile log likelihood curve is re-scaled to 0 and shown down to -2 log likelihood units. The values corresponding to the -2 log likelihood units then form a 95% confidence interval for the rate. These confidence intervals are summarised in Table 2.5. The profile log likelihood function for the parent de novo rate attains its maximum at the boundary at sites 1–22. This may raise concerns of the actual length of the confidence interval because the distribution of the maximum in this case may in fact be a mixture (Self and Laing, 1987). Nevertheless, it is clear from the plot in Figure 2.5 that 0 should be included in the interval.

Log likelihood surfaces for δ_p versus δ_d with μ and m fixed to the estimated values are plotted in Figures 2.6 and 2.7. Consistent with the profile log likelihood plots, the log likelihood surface for either data set has only one mode. One may also compare with the log likelihood surfaces on simulated data, such as Figures 2.2 and 2.3, where two modes are easily detected.

To investigate the hypotheses of equal de novo rates, $\delta_p = \delta_d$, and of no parent de novo event, $\delta_p = 0$, we obtain the maximum log likelihood under either hypothesis and compare it with that under the unconstrained model. The results are listed in Table 2.6. We apply the likelihood ratio test for either hypothesis. Under $\delta_p = \delta_d$ the null distribution of the test statistic $-2 \log L_0/L$ is a χ^2_1. Under $\delta_p = 0$, which is on the boundary, the null distribution is a 50:50 mixture of point mass at 0 and a χ^2_1 distribution (Self and Laing, 1987). P values are also listed in Table 2.6. Based on these p values, we reject the null hypothesis of equal de novo rates for either data set, with strong evidence for sites 1–22 and weaker evidence for sites 25–52. On the other hand, we do not reject the hypothesis of no parent de novo events, and again, the evidence for the null at sites 1–22 is stronger than that at sites 25–52.

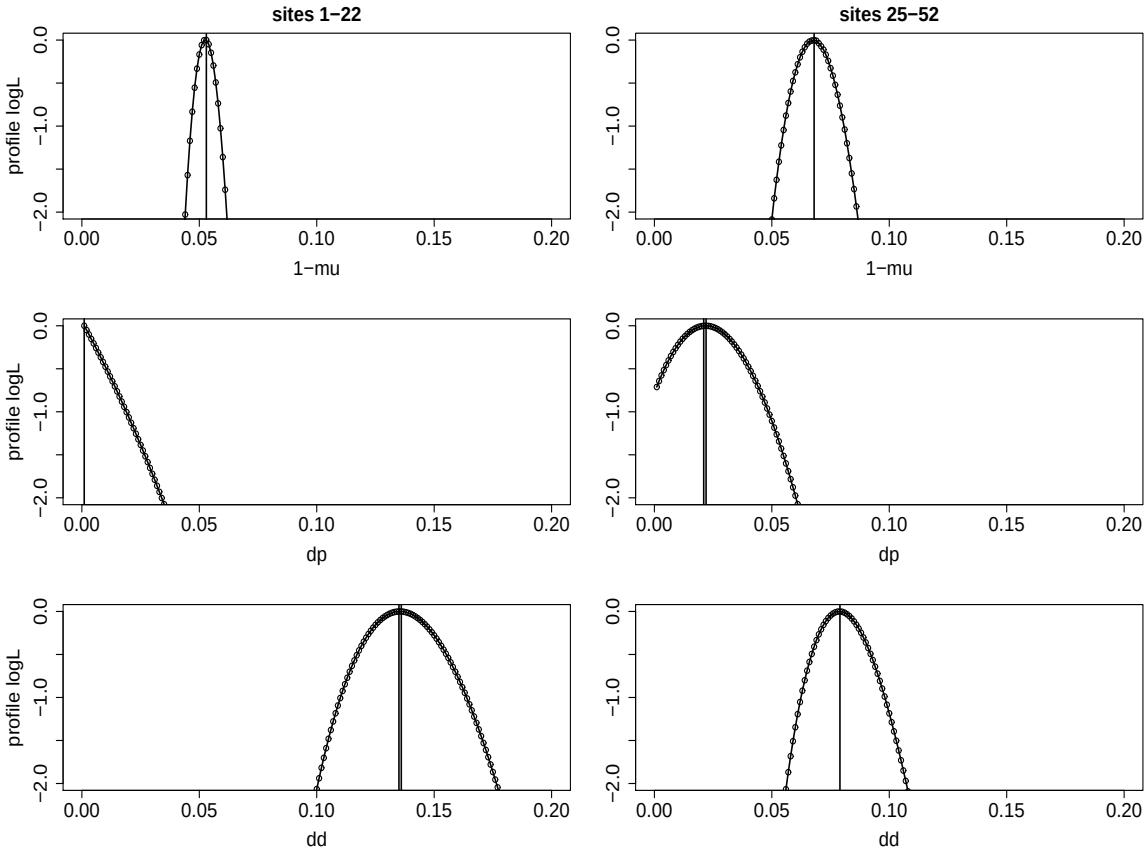

Figure 2.5: Profile log likelihood for each of $1-\mu$, δ_p and δ_d for the two *FMR1* data sets under the simple multi-site model. The left three plots are based on data set 1 (sites 1–22), and the right three plots data set 2 (sites 25–52). Each circle (x, y) on the curve is the log likelihood maximised over all other parameters with one parameter fixed to x. Vertical lines indicate the maxima. The log likelihood is re-scaled in each plot such that the highest log likelihood is 0.

2.7 Summary and discussion

In this chapter we formulated the rate estimation problem as a latent variable problem, in which two pieces of information are missing, one being the methylation states on the pre-replication parent strand, $\mathbf{P}$, and the other the strand type – which strand is which. We presented a simple multi-site model to infer the latent variables and then estimate the failure of maintenance and de novo rates. Methylation probabilities m_j

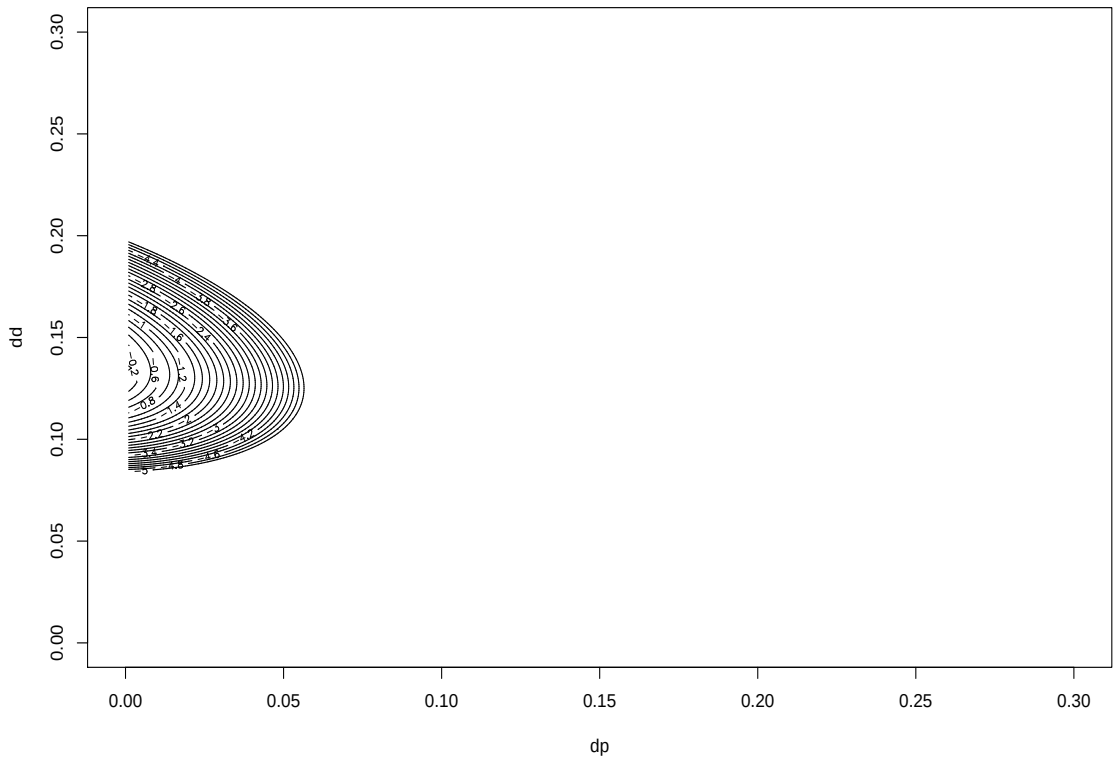

Figure 2.6: Log likelihood surface for δ_p versus δ_d for the *FMR1* data at sites 1–22 under the simple multi-site model. μ and m are fixed to the EM estimates. The log likelihood surface is re-scaled to have maximum at 0 and the contour lines are shown down to 5 log likelihood units below 0.

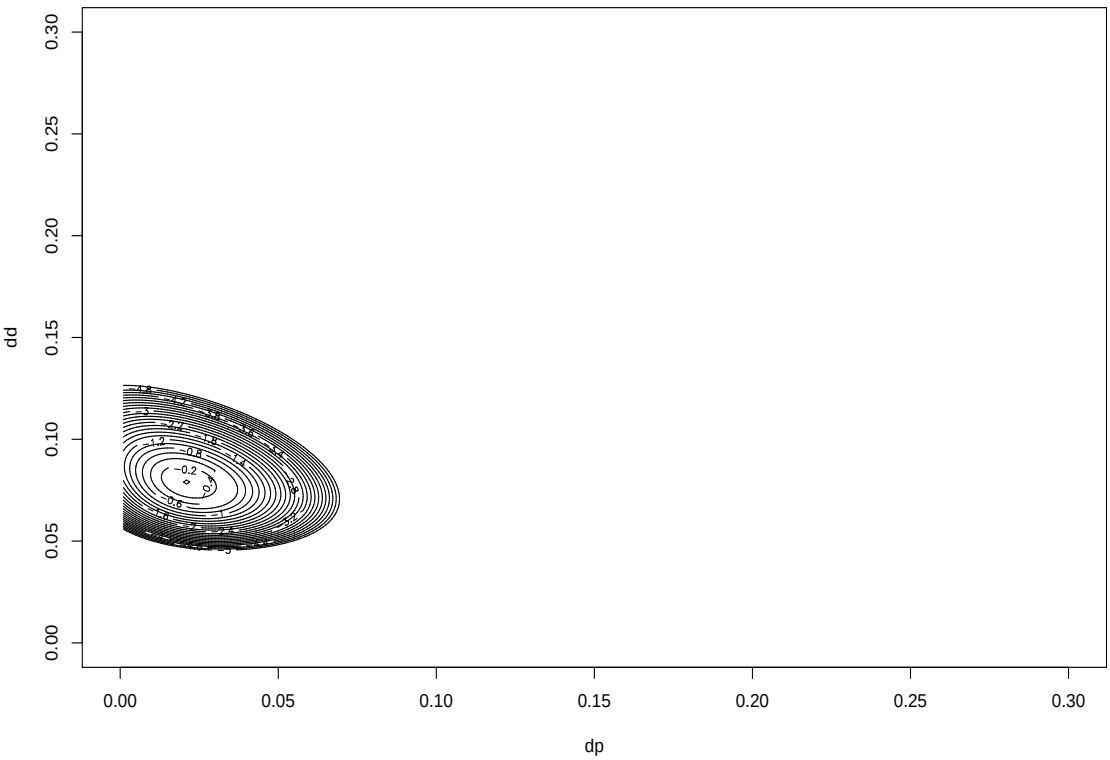

Figure 2.7: Log likelihood surface for δ_p versus δ_d for the *FMR1* data at sites 25–52 under the simple multi-site model. μ and m are fixed to the EM estimates. The log likelihood surface is re-scaled to have maximum at 0 and the contour lines are shown down to 5 log likelihood units below 0.

Table 2.6: Maximum log likelihood under $\delta_p = \delta_d$, under $\delta_p = 0$ and under the unconstrained model, respectively, for the two *FMR1* data sets under the simple multi-site model. Numbers in brackets are p values under the likelihood ratio tests. The null distribution is χ^2_1 under $\delta_p = \delta_d$ and is a 50:50 mixture of point mass at 0 and a χ^2_1 under $\delta_p = 0$.

	FMR1 Data	
Model	Sites 1–22	Sites 25–52
$\delta_p = \delta_d$	-2127.820 ($p = 10^{-5}$)	-1774.948 ($p = 0.018$)
$\delta_p = 0$	-2118.048 ($p = 1$)	-1772.950 ($p = 0.105$)
unconstrained	-2118.048	-1772.165

plays an important role in distinguishing different types of methylation events. Since m_j is defined as $\Pr(P_{ij} = 1)$, the probability that the pre-replication parent strand is methylated and can be estimated by the proportion of methylated dyads in the data, it is then possible to infer probabilistically $\mathbf{P}$ from those methylation probabilities. For example, a hypermethylated region would indicate that $\mathbf{P}$ is likely to be methylated at most CpG sites. Together with the assumption of no loss of methylation on the parent strand, we can further infer that the post-replication parent strand $\mathbf{Q}$ is also mostly methylated. If there were no de novo events and only failure of maintenance events, then the daughter strand $\mathbf{D}$ would always be methylated less than $\mathbf{Q}$. In the case where de novo events also occur, the more methylated strand in an observed double-stranded sequence is still more likely to be the post-replication parent strand, unless the daughter de novo rate greatly exceeds the failure of maintenance rate and the parent de novo rate put together. Indeed, this last scenario may explain the multimodality of the likelihood surface under Model C ($1 - \mu = 0.01$, $\delta_p = 0.001$ and $\delta_d = 0.15$) in simulation studies (Figures 2.2 and 2.3). Once the latent variables are inferred probabilistically, we can then infer the type of methylation event at each site. Specifically, we can infer daughter de novo events based on probabilistic

assignment of the strand type. We can also distinguish failure of maintenance events and parent de novo events, because the former occur at previously methylated sites and the latter previously unmethylated sites. The inference carries uncertainty, and accounting for experimental errors and other factors can aggravate confounding. But this simple multi-site model illustrates what information we can extract from the data for parameter estimation, and this capability is fully demonstrated in the simulation studies. This simple multi-site model forms the foundation of more elaborate models in the next chapter.

We applied this model to the two *FMR1* data sets and obtained results summarised in Table 2.5. Results from the likelihood ratio tests further suggest that the parent de novo rate is likely to be 0 in both regions (Table 2.6). Laird et al. (2004) and Genereux et al. (2005) also analysed data collected from the *FMR1* data (see Section 2.2 for details of their methods). A comparison of our results with theirs would shed light on how different model assumptions affect the inference and therefore how we should interpret our findings.

The comparison of our results with that from Laird et al. (2004) can be found in Table 2.7. Recall that Laird et al. (2004) also used the entire sequence for rate

Table 2.7: Comparison of results from the simple multi-site model via the expectation-maximisation (EM) algorithm and those from Laird et al. (2004). The multi-site EM method was applied to the *FMR1* data at sites 1–22, whereas the results from Laird et al. (2004) were for a different data set at the same locus. Numbers in brackets for the multi-site EM method are 95% confidence intervals obtained from profile likelihood functions. Numbers in brackets for results from Laird et al. (2004) are ranges.

	Multi-site Model via EM	Laird *et al.* (2004)
$1-\mu$	0.05 (0.044, 0.061)	0.04 (min. 0.02; max. 0.05)
δ_p	1.93×10^{-4} (0, 0.033)	assumed to be 0
δ_d	0.14 (0.101, 0.176)	0.17 (min.0.11; max. 0.23)

estimation and assumed that the failure of maintenance rate and the daughter de novo rate each were constant across CpG sites, but they did not assign which strand is which in a probabilistic way. In addition, they assumed that there were no parent de novo events. The points estimates for $1 - \mu$ and δ_d from the two studies are very close.

When comparing with the single-site maximum likelihood method (Genereux et al., 2005), we applied their approach to both *FMR1* data sets and plotted the estimates against those from the multi-site EM method in Figure 2.8. Since the single-site ML estimate of methylation probability m_j is just the observed proportion of methylated CpG dinucleotides at site j, this estimate does not depend on the constraint imposed on de novo rates. As seen in the top two plots in Figure 2.8, estimated methylation probabilities under the two approaches are in good agreement as expected. The other four plots show boxplots of the single-site ML estimates obtained under $\delta_p = 0$ (the middle row) and under $\delta_p = \delta_d$ (the bottom row), and red dots indicating the multi-site EM estimates. Although the two approaches are not directly comparable, with the single-site ML method estimating site-specific rates and imposing temporal stationarity and additional constraints on de novo rates, $1-\hat{\mu}$ and $\hat{\delta}_d$ under the multi-site model are similar to the medians of the estimates under the single-site approach; the difference is even smaller when $\delta_p = 0$ is imposed under the single-site model.

These comparisons suggest consistency in results for the failure of maintenance and daughter de novo rates across three studies. They also suggest that, when we ignore variability and experiment errors, parent de novo events are not necessary to explain the imperfection in methylation patterns transmitted over cell generations. In Chapter 3, however, we will see that, as we extend the multi-site model to account for variability, experimental errors and temporal stationarity, the data no longer strongly support the hypothesis of no parent de novo events. Hence, model assumptions have a substantial impact on the inference of de novo rates.

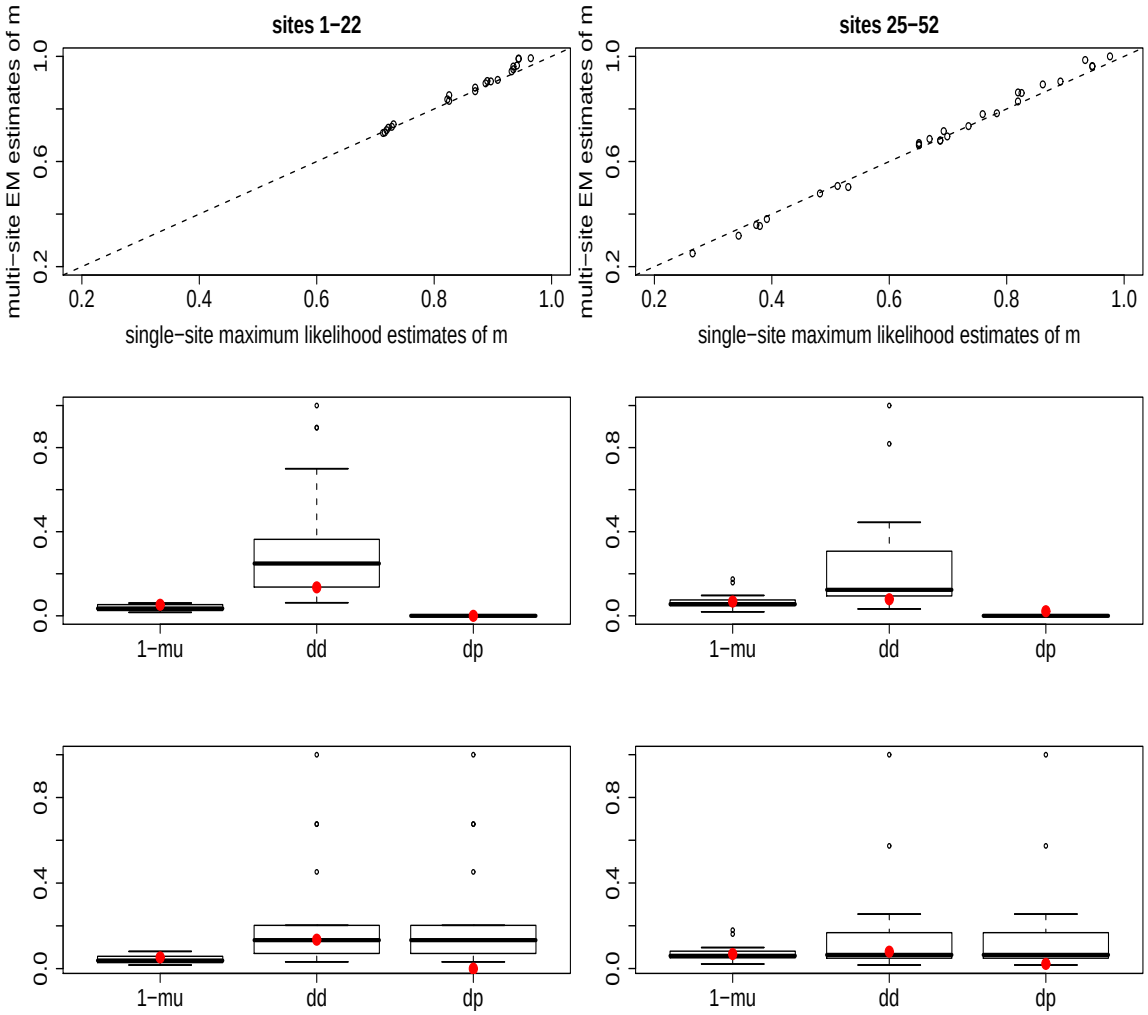

Figure 2.8: Comparison of results obtained from the simple multi-site model via the expectation-maximisation (EM) algorithm and those from the single-site maximum likelihood (ML) approach under two constraints for the two real data sets. The left column shows the results for sites 1–22, whereas the right column sites 25–52. **Top Row**: estimates for methylation probabilities m_j. These estimates under the single-site ML approach do not change regardless of the constraint imposed on de novo rates. **Middle Row:** the multi-site EM estimates (red dots) versus the single-site ML estimates under $\delta_p = 0$ (boxplots). **Bottom Row:** the multi-site EM estimates (red dots) versus the single-site ML estimates under $\delta_p = \delta_d$ (boxplots). Specifically, the multi-site EM estimates are: at sites 1–22, $1 - \hat{\mu} = 0.05$, $\hat{\delta}_p = 1.93 \times 10^{-4}$, $\hat{\delta}_d = 0.14$; at sites 25–52, $1 - \hat{\mu} = 0.07$, $\hat{\delta}_p = 0.02$, $\hat{\delta}_d = 0.08$.

Chapter 3

MODEL EXTENSIONS: RATE VARIABILITY, TEMPORAL STATIONARITY AND EXPERIMENTAL ERRORS

In the previous chapter, we have constructed a simple multi-site model that makes use of the strand information to estimate a constant maintenance, parent de novo or daughter de novo rate across sites. Applying the multi-site model to the *FMR1* data sets gives us estimates that are comparable with estimates in previous studies (Laird et al., 2004; Genereux et al., 2005).

Assuming constant rates across sites, however, is inadequate for our purpose. One of the goals of studying the transmission process of DNA methylation is to gain understanding of variability in the rates of methylation events across sites. Genereux et al. (2005) addressed this question by estimating site-specific rates with their single-site maximum likelihood (ML) method and trying to explain the variability in rates by the variability in their estimates. This approach, however, does not recognise the distinction between these two types of variability: variability in estimates can be due to the true variability in the rate across sites, as well as imprecision of the estimation method. As explained in Section 2.2, the single-site ML method gives an estimate of 1 to de novo rates at sites with no unmethylated CpG dyads. This estimate of 1, rather an artefact of the method, can inflate drastically the variability in the estimates. The authors constructed a confidence region for each pair of these site-specific point estimates of μ and d in order to examine variability, but a confidence region reflects sampling properties, and provides no information on the *true* rate variation across sites. Therefore, we would like to have a model that can account for variability and

an estimation approach that gives an estimate of this variability.

We extend the simple multi-site model to allow for rate variability in Section 3.1. This extension implies a large increase in the dimension of the parameter space. Since there are usually not enough data, it is always desirable to apply regularisations in high-dimensional estimation problems. Here we adopt a hierarchical structure in the extended model. To assess uncertainty in the estimates and to incorporate prior knowledge, we use a Bayesian method for parameter estimation.

Temporal stationarity, an assumption crucial to the single-site ML approach and other approaches for analysing methylation patterns, provides another type of regularisation. As we will see in Section 3.2, it can greatly increase the sharpness of the likelihood surface, which is one of the reasons why the single-site ML approach is sensitive to data, as demonstrated in Chapter 2. Thus we propose a way in Section 3.2 to incorporate this assumption in a flexible manner.

In the description of the data collection procedure in Chapter 1, we also pointed out the existence of different types of experimental errors. Incorporating errors into data analysis is important, sometimes even crucial, as they may change the results qualitatively. Also some error rates in the bisulfite PCR experiment can be tricky to obtain by laboratory approaches or have to be estimated using external data. A statistical method that enables us to use the original data for error rate estimation will be much desirable. In Section 3.3 we incorporate two types of experimental errors into the Bayesian hierarchical multi-site model and provide an estimation procedure for the error rates.

3.1 Allowing for rate variability: a Bayesian hierarchical model

3.1.1 The model and distribution assumptions

The constant rate definitions (2.1.1)–(2.1.3) in Chapter 2 can be easily extended to site-specific ones:

$$\Pr(D_{ij} = 0 | P_{ij} = 1) = 1 - \mu_j \quad \text{(failure of maintenance)} \tag{3.1.1}$$

$$\Pr(Q_{ij} = 1 | P_{ij} = 0) = \delta_{p_j} \quad \text{(de novo methylation on parent)} \tag{3.1.2}$$

$$\Pr(D_{ij} = 1 | P_{ij} = 0) = \delta_{d_j} \quad \text{(de novo methylation on daughter).} \tag{3.1.3}$$

We continue to assume no active removal of methylation on the parent strand, which means that equation (2.1.4) still holds:

$$\Pr(Q_{ij} = 1 | P_{ij} = 1) = 1. \tag{3.1.4}$$

The methylation probability m_j is defined as before and it varies across sites:

$$\Pr(P_{ij} = 1) = m_j. \tag{3.1.5}$$

Conditional probabilities $h_\lambda(q_{ij}, d_{ij}; p_{ij}) = \Pr\left((Q_{ij}, D_{ij}) = (q_{ij}, d_{ij}) | P_{ij} = p_{ij}\right)$ also become site-specific (Table 3.1).

Now, parameters of interest are $\lambda = \{\mu_j, \delta_{p_j}, \delta_{dj}, j = 1, \cdots, S\}$. The probability of observing each sequence with known strand types, $d_\lambda(\mathbf{x}_i, \mathbf{y}_i)$ for example, and the likelihood function for the entire data set $L(\lambda; \{\mathbf{x}, \mathbf{y}\})$ have the same expressions as those in Section 2.1, although evaluated using site-specific rates. Here we state them again:

$$d_\lambda(\mathbf{x}_i, \mathbf{y}_i) \equiv \Pr((\mathbf{Q}_i, \mathbf{D}_i) = (\mathbf{x}_i, \mathbf{y}_i)) \tag{3.1.6}$$

$$= \prod_{j=1}^{S} \Pr((Q_{ij}, D_{ij}) = (x_{ij}, y_{ij})) \tag{3.1.7}$$

$$= \prod_{j=1}^{S} \sum_{z_{ij}=0}^{1} h_\lambda(x_{ij}, y_{ij}; z_{ij}) m_j^{z_{ij}} (1 - m_j)^{1 - z_{ij}}, \tag{3.1.8}$$

Table 3.1: Conditional probabilities of methylation events at site j, i.e., $h_\lambda(q_{ij}, d_{ij}; p_{ij}) = \Pr((Q_{ij}, D_{ij}) = (q_{ij}, d_{ij}) | P_{ij} = p_{ij})$. P_{ij}, Q_{ij} and D_{ij} are methylation states on the pre-replication parent strand, the post-replication parent strand and the daughter strand, respectively. 0 represents unmethylated and 1 methylated. This table is the site-specific version of Table 2.1 in Chapter 1.

$(Q_{ij}, D_{ij}) = (q, d)$	$P_{ij} = p$	$h_\lambda(q_{ij}, d_{ij}; p_{ij})$
(0, 0)	1	0
(0, 1)	1	0
(1, 0)	1	$1 - \mu_j$
(1, 1)	1	μ_j
(0, 0)	0	$(1 - \delta_{p_j})(1 - \delta_{dj})$
(0, 1)	0	$(1 - \delta_{p_j})\delta_{dj}$
(1, 0)	0	$\delta_{p_j}(1 - \delta_{dj})$
(1, 1)	0	$\delta_{p_j}\delta_{dj}$

and

$$L(\lambda; \{\mathbf{x}, \mathbf{y}\}) = \prod_{i=1}^{N} \Big(\Pr((\mathbf{Q}_i, \mathbf{D}_i) = (\mathbf{x}_i, \mathbf{y}_i); \lambda) + \mathbf{1}(\mathbf{x}_i \neq \mathbf{y}_i) \Pr((\mathbf{Q}_i, \mathbf{D}_i) = (\mathbf{y}_i, \mathbf{x}_i); \lambda) \Big) \tag{3.1.9}$$

$$\equiv \prod_{i=1}^{N} \Big(d_\lambda(\mathbf{x}_i, \mathbf{y}_i) + \mathbf{1}(\mathbf{x}_i \neq \mathbf{y}_i) d_\lambda(\mathbf{y}_i, \mathbf{x}_i) \Big). \tag{3.1.10}$$

Note that $h_\lambda(x_{ij}, y_{ij}; z_{ij})$ is now based on Table 3.1.

Including methylation probabilities m_j, the parameter space now has $4S$ dimensions, fairly high for the small data sets that are available. In addition, we are interested in learning about the variability across sites for each type of rate. Simply estimating site-specific rates still would not help us achieve this goal. A common technique to reduce dimensionality and to incorporate the notion of rate variability is to employ a hierarchical structure. Specifically, we assume that μ_j, δ_{p_j}, δ_{dj} and m_j each

follow a Beta(r, g) distribution with mean r and "scaled" variance g; "scaled" since the variance under this parametrisation is $gr(1-r)$. To see how this parametrisation is related to the conventional α–β parametrisation, recall that a random variable X from a Beta(α, β) distribution has density

$$f(x) \propto x^{\alpha-1}(1-x)^{\beta-1}. \tag{3.1.11}$$

Then we have:

$$r = \frac{\alpha}{\alpha+\beta}, \quad g = \frac{1}{\alpha+\beta+1}. \tag{3.1.12}$$

The r–g parametrisation is preferred in our analysis because (1) r and g are easily interpretable; and (2) both r and g take values between 0 and 1. Both properties make it easy to choose priors for r and g in the Bayesian inference.

3.1.2 The Bayesian framework

Parameters in the hierarchical model in the previous section can be estimated using maximum likelihood or Bayesian approaches. Here we adopt a Bayesian framework to assess the strength of evidence in the inference, which is particularly appealing in the case of a small sample size. Prior distributions are assigned to the top-level variables in the hierarchy, which are mean rs and scale variance gs in the beta distributions. We assign a uniform distribution defined over (0,1), i.e., Unif$(0, 1)$, as the prior to each r, and a Unif$(-4, 0)$ prior to each $\log_{10} g$.

The Unif$(-4, 0)$ prior for $\log_{10} g$ captures a wide range of variation, while assigning most probability mass to values of interest. Figure 3.1 shows distributions of a beta random variable with varying values of g. The shape and width of the distribution change drastically as $\log_{10} g$ changes from -4 to -0.001: at $\log_{10} g = -4$, the histogram is centred on 0.05 with a width of 0.02; at -1.5, the distribution shifts to 0 and spreads out to 0.2; at -0.5, the distribution is more concentrated at 0, but at the same time can take values as high as 0.9; when $\log_{10} g = -0.001$, however, the samples form two

clusters, one at 0 and the other at 1. Hence, assigning priors to g on the log scale enables us to distinguish among small values such as 0.001 and 0.01 and therefore correctly infer the level of variability in rates (see Table 3.2 for guidelines on the interpretation of g). Sometimes we exclude $(-0.5, 0)$ and use a Unif$(-4, -0.5)$ prior for $\log_{10} g$, so as to avoid spending too much time in unlikely regions of the sample space.

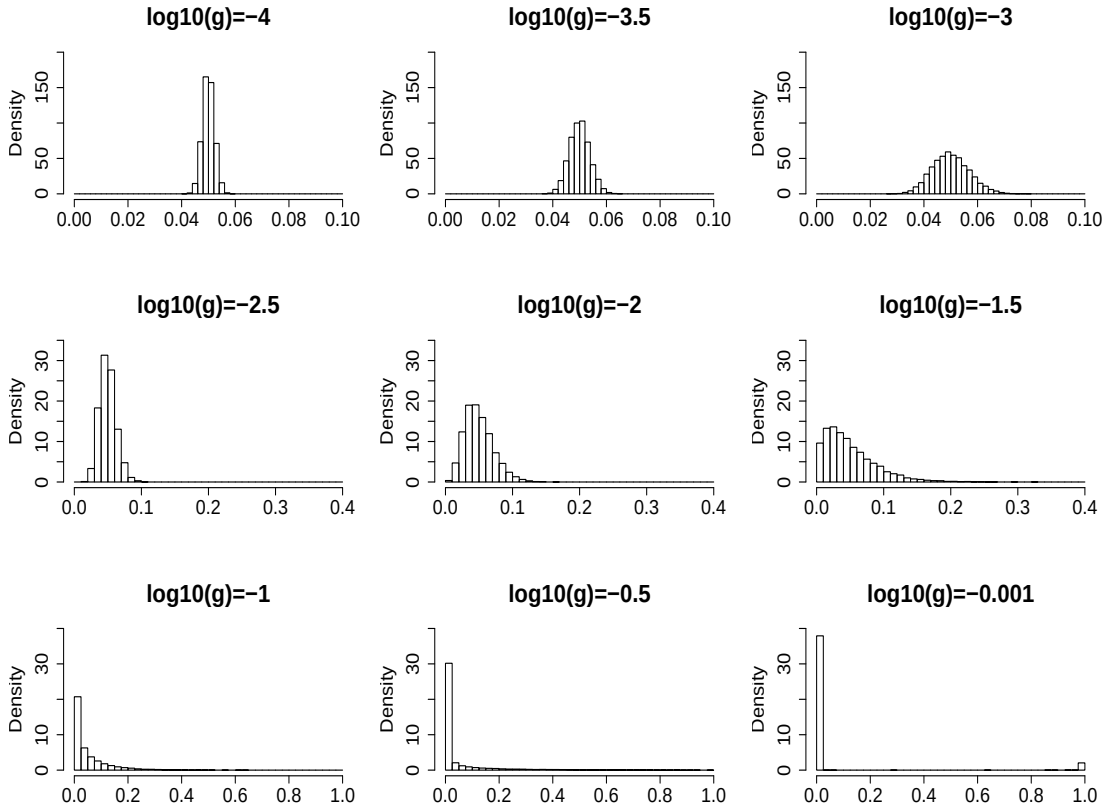

Figure 3.1: Histograms of beta distributions with mean $r = 0.05$ and different values of scaled variance g. The horizontal axis has a different range in each row, so does the vertical axis. Note that, as g increases, the histogram spreads out to the entire support of $(0, 1)$, and a second peak at 1 starts to emerge.

Table 3.2: Interpretation of scaled variance g on the $\log_{10}$ scale.

$\log_{10} g$	Variability
< -3	very low
-3 to -2	low
-2 to -1	medium
> -1	high

3.1.3 Introduction to Markov chain Monte Carlo (MCMC) methods

Markov chain Monte Carlo (MCMC) methods are sampling-based methods (see Liu, 2001; Besag, 2001). Under the Bayesian framework, the inference is generally based on the posterior distribution of parameters given the data and the prior. The posterior distribution, as the target distribution, usually does not have a closed form. The Monte Carlo strategy then is to simulate samples from the target distribution. Specially, consider random variable X and a subset B in its sample space S. X has distribution $\pi(X)$. Then evaluating a probability for X can be converted to calculating the expectation, which is just a weighted average:

$$\Pr(X \in B) = E\mathbf{1}(X \in B) = \sum_{x \in S} I(x \in B)\pi(x), \tag{3.1.13}$$

where $\mathbf{1}()$ is the indicator function. The average of independent draws $\mathbf{1}^{(t)}(x \in B)$, where $t = 1, ..., T$, under the Monte Carlo strategy,

$$\frac{1}{T}\sum_{t=1}^{T} \mathbf{1}^{(t)}(x \in B), \tag{3.1.14}$$

thus provides an unbiased estimator for the above probability $\Pr(X \in B)$.

In many cases, however, it is essentially impossible to obtain independent samples. Fortunately, dependent samples can replace the independent ones in the expression (3.1.14), if those dependent samples form an ergodic Markov chain with the same

(finite) state space S and stationary distribution π. A Markov chain is said to be ergodic if it is positive recurrent (or irreducible; being able to reach any state with a positive probability) and aperiodic. A much more stringent condition, but easier to verify, for a Markov chain to attain stationarity is detailed balance, which plays a key role in the Metropolis-Hastings algorithm and will be explained below.

The Metropolis-Hastings algorithm

Most MCMC algorithms rely on the Metropolis-Hastings (MH) algorithm (Metropolis et al., 1953; Hastings, 1970). Starting from certain initial value, the MH algorithm proposes a small move, according to a proposal distribution, around the current value, and accepts the new move with certain probability. Thus, those moves from many iterations form a Markov chain. Specifically, let the target distribution be π, and the proposal distribution be $q(\sigma)$, where σ is the standard deviation. At the t-th iteration, the current value is $X_t = x$. The MH algorithm proceeds as follows:

1. Generate a proposal $y \sim q(x; \sigma)$; that is, y is a local move around x.

2. Calculate the acceptance probability

$$A = \min\left\{1, \frac{\pi(y)q(x|y)}{\pi(x)q(y|x)}\right\}, \tag{3.1.15}$$

   where $q(x|y)$ is the probability (or density) of x given y under distribution function $q()$.

3. Generate a random number r between 0 and 1, and compare it with the acceptance probability A.

$$X_{t+1} = \begin{cases} y, & \text{if } r < A, \\ x, & \text{otherwise.} \end{cases} \tag{3.1.16}$$

Under this algorithm, the actual transition distribution is

$$p(y|x) = q(y|x) \min\left\{1, \frac{\pi(y)q(x|y)}{\pi(x)q(y|x)}\right\}, \tag{3.1.17}$$

for $x \neq y$. Detailed balance requires that

$$\pi(x)p(y|x) = \pi(y)p(x|y). \tag{3.1.18}$$

Thus,

$$\pi(x)p(y|x) = \pi(x)q(y|x) \min\left\{1, \frac{\pi(y)q(x|y)}{\pi(x)q(y|x)}\right\} \tag{3.1.19}$$

$$= \min\left\{\pi(x)q(y|x), \pi(y)q(x|y)\right\}, \tag{3.1.20}$$

and equation (3.1.18) holds because of symmetry in x and y in (3.1.20) (Liu, 2001).

In terms of practical matters, symmetric proposal distributions, such as normal, are often chosen, because $q(x|y)$ and $q(y|x)$ are then cancelled out in the acceptance probability. Standard deviation σ in the proposal distribution is fine-tuned so that the chain can explore a large sample space, but also does not have to waste too much time in places where the values are unlikely for the data. The acceptance rate of the parameter gives some indication whether the standard deviation (or the size of a local move) is large enough. Different people, however, have different opinions on what rate is satisfactory. Here we use 20–30% for this rate. One usually does a very long run or multiple shorter runs (or both), and uses samples a few moves apart, after discarding samples from the initial m iterations; those m iterations are called the burn-in period. The above measures are taken to ensure that the Markov chain explores the entire sample space with reasonable probability support from the posterior distribution; in other words, the chain *mixes* well.

Determining whether the Markov chain has mixed well is not trivial. Currently there is no standard approach to it. Since a well-mixed chain should attain the stationary distribution, one of the diagnostic tools we use is to carry out several independent runs and check whether they give similar histograms.

The Gibbs sampler

The Gibbs sampler is a special case of the Metropolis-Hastings sampler. For a joint distribution of k random variables Y_1, Y_2, ..., Y_k, the Gibbs sampler uses the full conditional distribution $\pi(Y_i|Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_k)$ as the proposal distribution for Y_i in a regular MH algorithm, and thus always accepts new moves, because the acceptance probability is always 1. Convergence is guaranteed under the Gibbs sampler, but it can take a very long time to reach convergence.

3.1.4 MCMC procedures for inference under the multi-site model

We use Markov chain Monte Carlo methods for inference. The procedure involves three main steps:

1. Sample $(\mathbf{P}, \mathbf{Q}, \mathbf{D})$ given $\{\mathbf{x}, \mathbf{y}\}$ and λ. Since sampling $(\mathbf{P}, \mathbf{Q}, \mathbf{D})$ together involves 2^{S+1} states, which is computationally impractical even for moderate values of S, we do it in two steps.

   - Sample each ordered pair $(\mathbf{Q}_i, \mathbf{D}_i)$ from $\{\mathbf{x}, \mathbf{y}\}$ and λ based on the likelihood function in (3.1.10). For example, the posterior probability of the top strand $\mathbf{x}_i$ being the post-replication parent strand $\mathbf{Q}_i$ is

     $$p((\mathbf{Q}_i, \mathbf{D}_i) = (\mathbf{x}_i, \mathbf{y}_i)|\{\mathbf{x}, \mathbf{y}\}, \lambda) = \frac{d_\lambda(\mathbf{x}_i, \mathbf{y}_i)}{d_\lambda(\mathbf{x}_i, \mathbf{y}_i) + d_\lambda(\mathbf{y}_i, \mathbf{x}_i)}, \tag{3.1.21}$$

     where $d_\lambda(\mathbf{x}_i, \mathbf{y}_i)$ is defined in (3.1.8) and $d_\lambda(\mathbf{y}_i, \mathbf{x}_i)$ is defined similarly.

   - Sample $\mathbf{P}_i$ given $(\mathbf{Q}_i, \mathbf{D}_i)$ and λ. Because methylation states at different sites are conditionally independent, sampling can be done for each P_{ij} separately.

     $$p(P_{ij}|(Q_{ij}, D_{ij}), \lambda) \propto p((Q_{ij}, D_{ij})|P_{ij}) m_j^{P_{ij}} (1 - m_j)^{1-P_{ij}}. \tag{3.1.22}$$

2. Sample m_j, μ_j, δ_{p_j} and δ_{dj} from current estimates of all other variables and parameters via a Gibbs sampler. For example, since

$$\sum_{i=1}^{N} P_{ij}|m_j \sim \text{Binomial}(N, m_j) \tag{3.1.23}$$

$$m_j \sim \text{Beta}(\alpha_m, \beta_m), \tag{3.1.24}$$

then, although the beta distribution of m_j is not a prior in the strict sense in hierarchical modelling, it can be thought of as the conjugate prior for the binomial random variable $\sum_{i=1}^{N} P_{ij}$, thus making the Gibbs sampling possible. Under the α–β parametrisation, the posterior distribution of m_j given all other variables is also beta due to conjugacy, and depends only on P_j:

$$m_j|\mathbf{P}_j \sim \text{Beta}(\alpha_m + \sum_{i=1}^{N} P_{ij}, \beta_m + N - \sum_{i=1}^{N} P_{ij}). \tag{3.1.25}$$

Likewise, we can derive the posterior distributions for μ_j, δ_{p_j} and δ_{dj} under the α–β parametrisation:

$$\mu_j|\mathbf{P}_j, \mathbf{Q}_j, \mathbf{D}_j \sim \text{Beta}\left(\alpha_\mu + \sum_{i=1}^{N} D_{ij}Q_{ij}P_{ij}, \beta_\mu + \sum_{i=1}^{N}(1 - D_{ij})Q_{ij}P_{ij}\right), \tag{3.1.26}$$

$$\delta_{p_j}|\mathbf{P}_j, \mathbf{Q}_j, \mathbf{D}_j \sim \text{Beta}\left(\alpha_p + \sum_{i=1}^{N} Q_{ij}(1 - P_{ij}), \beta_p + \sum_{i=1}^{N}(1 - Q_{ij})(1 - P_{ij})\right), \tag{3.1.27}$$

$$\delta_{dj}|\mathbf{P}_j, \mathbf{Q}_j, \mathbf{D}_j \sim \text{Beta}\left(\alpha_d + \sum_{i=1}^{N} D_{ij}(1 - P_{ij}), \beta_d + \sum_{i=1}^{N}(1 - D_{ij})(1 - P_{ij})\right). \tag{3.1.28}$$

3. We convert αs and βs in the beta distributions to rs and gs, using equations in (3.1.12), and then update rs and gs. No conjugate prior exists for beta distributions, so Metropolis-Hastings (MH) samplers are used to to update each r,

which has a Unif$(0,1)$ prior, and to update each $\log_{10} g$, which has a Unif$(-4,0)$ prior. We use r_μ and g_μ, the parameters for maintenance rate μ, below to illustrate the updating procedures.

To update r_μ, we derive its posterior density first.

$$p(r_\mu|\cdot) = p(r_\mu|g_\mu, \mu) \tag{3.1.29}$$

$$\propto p(r_\mu) \prod_{j=1}^{S} p(\mu_j|\alpha(r_\mu; g_\mu), \beta(r_\mu; g_\mu)) \tag{3.1.30}$$

$$= I(0 < r_\mu < 1) \prod_{j=1}^{S} f(\mu_j; \alpha(r_\mu; g_\mu), \beta(r_\mu; g_\mu)) \tag{3.1.31}$$

where $f()$ is the density function of a beta distribution as in expression (3.1.11) with parameters α and β, which are functions of r_μ and g_μ now. We then use the following steps to update r_μ:

(a) Generate a proposal r_μ^* from a normal distribution with the current value r'_μ as the mean, and standard deviation σ. Denote this normal transition kernel by $q(r_\mu^*|r'_\mu)$.

(b) If $0 < r_\mu^* < 1$, continue. Otherwise stop and retain the current value.

(c) Compute $\alpha(r_\mu^*; g_\mu), \beta(r_\mu^*; g_\mu)$, $\alpha(r'_\mu; g_\mu)$, and $\beta(r'_\mu; g_\mu)$ using formulas in (3.1.12).

(d) Compute the logarithm of the Hastings ratio

$$\log A \equiv \log \frac{p(r_\mu^*|\cdot) q(r'_\mu|r_\mu^*)}{p(r'_\mu|\cdot) q(r_\mu^*|r'_\mu)} \tag{3.1.32}$$

$$= \sum_{j=1}^{S} \log f(\mu_j; \alpha(r_\mu^*; g_\mu), \beta(r_\mu^*; g_\mu)) - \sum_{j=1}^{S} \log f(\mu_j; \alpha(r'_\mu; g_\mu), \beta(r'_\mu; g_\mu)). \tag{3.1.33}$$

(e) Generate a random probability p^*. Accept r_μ^* if $\log p^* < \min(0, \log A)$. Otherwise, retain r'_μ.

To update $\log_{10}(g_\mu)$, let $t_\mu = \log_{10}(g_\mu)$. Then

$$\alpha(t_\mu; r_\mu) = r_\mu \frac{1 - 10^{t_\mu}}{10^{t_\mu}}, \quad \beta(t_\mu; r_\mu) = (1 - r_\mu)\frac{1 - 10^{t_\mu}}{10^{t_\mu}}. \tag{3.1.34}$$

The posterior of t_μ is

$$p(t_\mu|\cdot) = p(t_\mu|r_\mu, \mu) \tag{3.1.35}$$

$$\propto p(t_\mu) \prod_{j=1}^{S} p(\mu_j|\alpha(t_\mu; r_\mu), \beta(t_\mu; r_\mu)) \tag{3.1.36}$$

$$= I(-4 < t_\mu < 0) \prod_{j=1}^{S} f(\mu_j|\alpha(t_\mu; r_\mu), \beta(t_\mu; r_\mu)). \tag{3.1.37}$$

We can again apply the above MH algorithm, replacing r_μ with t_μ and calculating the logarithm of the Hastings ratio as follows

$$\log A \equiv \log \frac{\pi(t^*_\mu|g_\mu, \mu)q(t'_\mu|t^*_\mu)}{\pi(t'_\mu|g_\mu, \mu)q(t^*_\mu|t'_\mu)} \tag{3.1.38}$$

$$= \sum_{j=1}^{S} \log f(\mu_j; \alpha(t^*_\mu; g_\mu), \beta(t^*_\mu; g_\mu)) - \sum_{j=1}^{S} \log f(\mu_j; \alpha(t'_\mu; g_\mu), \beta(t'_\mu; g_\mu)). \tag{3.1.39}$$

3.2 Temporal stationarity

3.2.1 Rationale and modelling

The single-site ML approach (Genereux et al., 2005) assumes that the transmission process has attained stationarity over cell division. The observation in Stöger et al. (1997) that methylation densities at the *FMR1* locus were virtually unchanged over a five-year time span in several human males provides supporting evidence. Thus we would like to incorporate this assumption into our model. Mathematically, the stationarity assumption is determined by the stationarity equations (2.2.2)–(2.2.4), and is expressed as follows:

$$m = \frac{\delta_p + \delta_d}{1 + \delta_p + \delta_d - \mu}. \tag{3.2.1}$$

Here we examine what impact this assumption has on inference and how to incorporate this assumption into the multi-site model.

Again we carry out a one-site analysis. Consider Models A and C (Section 2.5), as well as their variants Models A* and C* (Table 3.3). Models A* and C* assume temporal stationarity, whereas Models A and C do not. For each model, we simulated 100 CpG dyads for a single site, and evaluated the log likelihood without and with the temporal stationarity assumption. The log likelihood surfaces (Figure 3.2) clearly indicate that temporal stationarity is a strong assumption: it can make the log likelihood surface much sharper, and can even change the orientation of the ridge of the log likelihood surface for $1 - \mu$ and δ_p. In terms of inference, incorporating temporal stationarity when the data are not consistent with this assumption can move the peak of the log likelihood surface away from the truth, as under Models A and C, or even introduce a second peak, such as in the plot of δ_p versus δ_d under Model A. When the data are indeed in stationarity, the contour lines under this assumption condense toward the truth, as under Models A* and C*. Nonetheless, incorporating this assumption still does not resolve the multimodality issue, shown in the δ_p versus δ_d plots. This is not surprising: the two de novo rates are symmetric in equation (3.2.1).

Table 3.3: Simulation models for analysis of temporal stationarity. Note that Models A* and C* assume temporal stationarity, while A and C do not.

	$1-\mu$	δ_p	δ_d	m
A	0.05	0.001	0.15	0.8
A*	0.05	0.001	0.15	0.75
C	0.01	0.001	0.15	0.8
C*	0.01	0.001	0.15	0.94

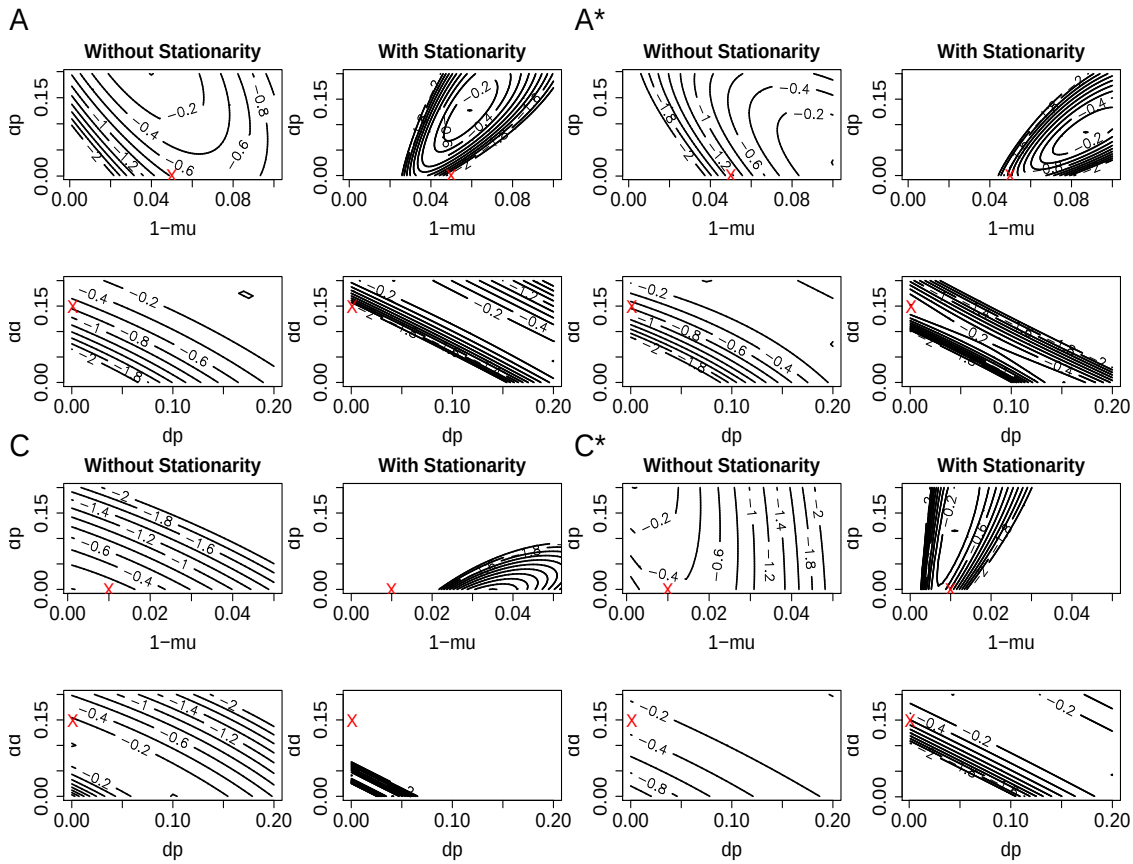

Figure 3.2: Impact of the temporal stationarity assumption on the inference for the rates of methylation events. Data were simulated at one CpG site under four models (Table 3.3): Models A and C do not assume temporal stationarity, whereas A* and C* make this assumption. Log likelihoods are evaluated without and with temporal stationarity, when other parameter values held to the truth. Red crosses indicate the simulation truth. Each surface is re-scaled so that the maximum is at 0 and the contours are shown to 2 log likelihood units below 0. These plots indicate that, when the data are not consistent with temporal stationarity, the analysis with this assumption can lead to incorrect estimates.

Based on these observations, we incorporate temporal stationarity into the multi-site model by allowing for deviation; the flexibility helps us obtain estimates that are robust to departure from stationarity. Specifically, the model for m_j is modified as

the following

$$m_j \sim Beta(r_{m,j}, g_m), \tag{3.2.2}$$

where

$$r_{m,j} = \frac{\delta_{p_j} + \delta_{dj}}{1 + \delta_{p_j} + \delta_{dj} - \mu_j}. \tag{3.2.3}$$

This distribution on m_j is centred on $r_{m,j}$ and allows for deviation, measured by g_m. The estimation procedure for g_m is the same as in Section 3.1.4: we assign a Unif$(-4, 0)$ prior to $\log_{10}(g_m)$, and update it by a Metropolis-Hastings sampler.

3.2.2 The MCMC procedure

Incorporating temporal stationarity changes the sampling procedure for rates of methylation events and the methylation probability. Under temporal stationarity, m_j is dependent on μ_j, δ_{p_j} and δ_{dj}. So we need to update rates of methylation events before updating m_j. To sample rates of methylation events, we can continue to use the full conditionals from the previous Gibbs samplers (equations 3.1.26 – 3.1.28) to generate proposals, but the acceptance probabilities need re-calculation. In other word, we now use Metropolis-Hastings samplers for the rates of methylation events (Section 3.1.3). The details are given below. Having updated μ_j, $\delta_{p,j}$, and $\delta_{d,j}$, we can still use a Gibbs sampler to update m_j, according to the posterior distribution (3.1.25). The only change is that $r_{m,j}$ is now determined by the updated μ_j, $\delta_{p,j}$, and $\delta_{d,j}$.

To illustrate how the Metropolis-Hastings sampler works for the rates of methylation events, we take the maintenance rate μ_j for example. The posterior distribution

of μ_j becomes

$$p(\mu_j|\cdot) = p(\mu_j|\mathbf{P}, \mathbf{Q}, \mathbf{D}, \mathbf{m}, \delta_\mathbf{p}, \delta_\mathbf{d}, \mathbf{r}, \mathbf{g}) \tag{3.2.4}$$

$$\propto p(\mathbf{P}_j, \mathbf{Q}_j, \mathbf{D}_j|m_j, \delta_{p,j}, \delta_{d,j}, \mu_j, r_{m,j}, g_m)p(m_j|r_{m,j}, g_m)p(\mu_j|\alpha_\mu, \beta_\mu)p(\delta_{p,j})p(\delta_{d,j}) \tag{3.2.5}$$

$$\propto p(\mathbf{P}_j, \mathbf{Q}_j, \mathbf{D}_j|m_j, \delta_{p,j}, \delta_{d,j}, \mu_j, r_{m,j}, g_m)p(m_j|r_{m,j}, g_m)p(\mu_j|\alpha_\mu, \beta_\mu) \tag{3.2.6}$$

$$= \left(\prod_{i=1}^{N} p(Q_{ij}|P_{ij}, m_j, \delta_{p,j}, \delta_{d,j})p(D_{ij}|P_{ij}, m_j, \delta_{p,j}, \delta_{d,j}, \mu_j)p(P_{ij}|m_j)\right)$$

$$\times\, p(m_j|r_{m,j}, g_m)p(\mu_j|\alpha_\mu, \beta_\mu) \tag{3.2.7}$$

$$\propto \left(\prod_{i=1}^{N} p(D_{ij}|P_{ij}, \delta_{d,j}, \mu_j)\right)p(m_j|r_{m,j}, g_m)p(\mu_j|\alpha_\mu, \beta_\mu) \tag{3.2.8}$$

$$\propto \mu_j^{\alpha_\mu + \sum_{i=1}^{N} D_{ij}P_{ij} - 1}(1-\mu_j)^{\beta_\mu + \sum_{i=1}^{N}(1-D_{ij})P_{ij} - 1}$$

$$\times\, f(m_j; \alpha_m(\mu_j, \delta_{p,j}, \delta_{d,j}), \beta_m(\mu_j, \delta_{p,j}, \delta_{d,j})), \tag{3.2.9}$$

where $\alpha_m(\mu_j, \delta_{p,j}, \delta_{d,j})$ and $\beta_m(\mu_j, \delta_{p,j}, \delta_{d,j})$ are functions of $r_{m,j}$ and g_m specified through equations (3.1.12), and $f()$ is again a beta density. We will again use variables with a * to denote new values and those with an $'$ current values. To update μ_j, new values are proposed as under the simple model using the beta distribution in (3.1.26), and accepted with probability $\min(1, A_{\mu,j})$ where

$$A_{\mu,j} = \frac{p(\mu_j^*|\cdot)q(\mu_j'|\mu_j^*)}{p(\mu_j'|\cdot)q(\mu_j^*|\mu_j')} = \frac{f(m_j; \alpha_m(\mu_j^*, \delta_{p,j}, \delta_{d,j}), \beta_m(\mu_j^*, \delta_{p,j}, \delta_{d,j}))}{f(m_j; \alpha_m(\mu_j', \delta_{p,j}, \delta_{d,j}), \beta_m(\mu_j', \delta_{p,j}, \delta_{d,j})}. \tag{3.2.10}$$

The two de novo rates are be updated in a similar fashion.

3.3 Experimental errors

3.3.1 Definitions and modelling

As mentioned in Section 1.2, bisulfite conversion is a major source of error in bisulfite PCR techniques (Genereux et al., 2008). There are two types of bisulfite conversion error here. Failure of conversion, with rate b, leaves unmethylated cytosines

unconverted and results in a reading of methylated state although the truth is unmethylated. Inappropriate conversion, the other type of error with rate c, converts methylated cytosines and leads to a reading of unmethylated states although the truth is methylated. The definitions are summarised in Table 3.4. Since the bisulfite

Table 3.4: Definitions of two types of bisulfite conversion error. b is the failure of conversion rate and c the inappropriate conversion rate.

		Observed	
		0	1
Truth	0	$1-b$	b
	1	c	$1-c$

conversion technique converts unmethylated cytosines into uracils for subsequent sequencing, the error rate b at which the process fails to convert those cytosines is a direct measure of the performance of this technique and has been well documented (see Frommer et al., 1992; Clark et al., 1994 for example). The inappropriate conversion error that converts methylated cytosines into uracils is generally assumed to be negligible. Neither type of error has been accounted for in published analyses of methylation patterns.

To incorporate bisulfite conversion error, we will continue to use Q_{ij} and D_{ij} to denote *true* methylation states on the post-replication parent and daughter strand. We introduce Q'_{ij} and D'_{ij} to represent *observed* methylation states that may be prone to error. Let $\lambda' = \{\mu_j, \delta_{p_j}, \delta_{d_j}, b_j, c_j, j = 1, \cdots, S\}$. The likelihood function under the

no error model can be easily modified as the following:

$$L(\lambda'|\{\mathbf{x},\mathbf{y}\}) = \prod_{i=1}^{N} \Pr(\{\mathbf{Q}'_i, \mathbf{D}'_i\} = \{\mathbf{x}_i, \mathbf{y}_i\}|\lambda') \tag{3.3.1}$$

$$= \prod_{i=1}^{N} \Big(d'_\lambda(\mathbf{x}_i, \mathbf{y}_i) + \mathbf{1}(\mathbf{x}_i \neq \mathbf{y}_i) d'_\lambda(\mathbf{y}_i, \mathbf{x}_i) \Big), \tag{3.3.2}$$

where

$$d'_\lambda(\mathbf{x}_i, \mathbf{y}_i) = \prod_{j=1}^{S} \sum_{Q_{ij}=0}^{1} \sum_{D_{ij}=0}^{1} \Pr((Q'_{ij}, D'_{ij}) = (x_{ij}, y_{ij})|(Q_{ij}, D_{ij}))$$
$$\times \sum_{z_{ij}=0}^{1} \Pr((Q_{ij}, D_{ij})|P_{ij} = z_{ij}) m_j^{z_{ij}} (1 - m_j)^{1-z_{ij}}. \tag{3.3.3}$$

Assuming errors occur independently across CpG sites and DNA strands, we have

$$\Pr((Q'_{ij}, D'_{ij}) = (x_{ij}, y_{ij})|(Q_{ij}, D_{ij})) = \Pr(Q'_{ij} = x_{ij}|Q_{ij}) \Pr(D'_{ij} = y_{ij}|D_{ij}), \tag{3.3.4}$$

where each term on the right hand side is a function of b_j and c_j as in Table 3.4.

Similar to the inference we have performed under the no error model, we can also assume b_j and c_j each follow a Beta(r, g) distribution and assign uniform priors to r's and $\log_{10}(g)$s.

3.3.2 The MCMC procedure

Since the presence of error impacts the distribution from which we sample $(\mathbf{Q}, \mathbf{D})$, we modify this step in the previous MCMC procedure. Again it can be broken down into two steps.

1. Sample each ordered pair with error $(\mathbf{Q}', \mathbf{D}')$ from $\{\mathbf{x}, \mathbf{y}\}$ and λ.

$$p((\mathbf{Q}'_i, \mathbf{D}'_i) = (\mathbf{x}_i, \mathbf{y}_i)|\{\mathbf{x}, \mathbf{y}\}, \lambda) = \frac{d'(\mathbf{x}_i, \mathbf{y}_i)}{d'(\mathbf{x}_i, \mathbf{y}_i) + d'(\mathbf{y}_i, \mathbf{x}_i)} \tag{3.3.5}$$

and

$$p((\mathbf{Q}'_i, \mathbf{D}'_i) = (\mathbf{y}_i, \mathbf{x}_i)|\{\mathbf{x}, \mathbf{y}\}, \lambda) = \frac{d'(\mathbf{y}_i, \mathbf{x}_i)}{d'(\mathbf{x}_i, \mathbf{y}_i) + d'(\mathbf{y}_i, \mathbf{x}_i)}. \tag{3.3.6}$$

2. Sample $(\mathbf{Q}, \mathbf{D})$ from $(\mathbf{Q}', \mathbf{D}')$ and λ. This update can be done site by site.

$$p((Q_{ij}, D_{ij})|(Q'_{ij}, D'_{ij}), \lambda) \propto p((Q'_{ij}, D'_{ij})|(Q_{ij}, D_{ij}), \lambda_j) p((Q_{ij}, D_{ij})|\lambda_j) \tag{3.3.7}$$

$$= p((Q'_{ij}, D'_{ij})|(Q_{ij}, D_{ij}), \lambda_j) \sum_{P_{ij}} h_{\lambda_j}(Q_{ij}, D_{ij}; P_{ij}) m_j^{P_{ij}}$$

$$\times (1 - m_j)^{1-P_{ij}}. \tag{3.3.8}$$

3. Update c_j via a Gibbs step. This is because

$$Q'_{ij}|Q_{ij}, b_j, c_j \sim (\text{Bernoulli}(1 - c_j))^{Q_{ij}} (\text{Bernoulli}(b_j))^{1-Q_{ij}}, \tag{3.3.9}$$

$$D'_{ij}|D_{ij}, b_j, c_j \sim (\text{Bernoulli}(1 - c_j))^{D_{ij}} (\text{Bernoulli}(b_j))^{1-D_{ij}}, \tag{3.3.10}$$

$$c_j \sim \text{Beta}(\alpha_c, \beta_c). \tag{3.3.11}$$

The posterior distribution of c_j is then

$$\text{Beta}(\alpha_c + \sum_{i=1}^{N} Q_{ij}(1 - Q'_{ij}) + \sum_{i=1}^{N} D_{ij}(1 - D'_{ij}), \beta_c + \sum_{i=1}^{N} Q_{ij}Q'_{ij} + \sum_{i=1}^{N} D_{ij}D'_{ij}). \tag{3.3.12}$$

b_j can be updated in a similar fashion.

3.4 Analysis of the *FMR1* data

In analysing the *FMR1* data, we do not try to estimate both types of error rates. This is because the failure of conversion error rate b is relatively easy to estimate using laboratory approaches; the experimental data indicate that $b = 0.003$ and that it is reasonable to assume all b_js are the same. The inappropriate conversion rate c is much harder to obtain by laboratory approaches. Hence, we fix $b_j = b = 0.003$ in our analysis and estimate c_js in addition to the failure of maintenance and de novo methylation rates. The experimental results, however, suggest that c_j is likely between 0 and 0.06. So we use a Unif(0,0.06) prior to r_c, the mean of c_j, in the following analyses.

3.4.1 Results from the Bayesian hierarchical model

We applied the Bayesian hierarchical multi-site model to the two *FMR1* data sets. For each data set, we carried out three independent runs with specifications listed in Table 3.5. Each run took about 6 days on a cluster with each CPU of approximately 2.4GHz. Trace plots (not shown) provided no obvious indication of poor mixing. The posterior density functions of $\log_{10} g$s (Figures 3.3 and 3.4) from different runs agree with one another very closely, suggesting a small MCMC variation. Hence, we pooled the MCMC samples from those independent runs together for inference.

Table 3.5: MCMC specifications of each run under the Bayesian hierarchical multi-site model for each *FMR1* data set.

	Run Length (Iterations)	Burn-in	Sampling Interval (Iterations)
Data set 1	720,000	20%	1,000
Data set 2	1080,000	20%	1,800

We summarised the average rates r in terms of 80% credible intervals in Table 3.6; 80% was chosen to include the majority of, but also not too many, MCMC samples. These credible intervals are narrow for $r_{1-\mu}$ and r_c, but much wider for r_{dp} and r_{dd}. In fact, the distribution of r_{dp} for the first data set covers a range as wide as $(0, 0.5)$ and is bimodal (Figure 3.5).

To investigate confounding among estimates of rates of methylation events, we further generate pairwise scatter plots for samples of rs (Figure 3.6). While there is almost a linear relation between r_{dp} and $1 - r_\mu$, and between r_{dd} and $1 - r_\mu$, two branches exist in the scatter plot of r_{dp} and r_{dd} in the first data set and are somewhat visible in the second data set. These observations suggest no significant confounding between the average failure of maintenance rate and the average de novo rates, but

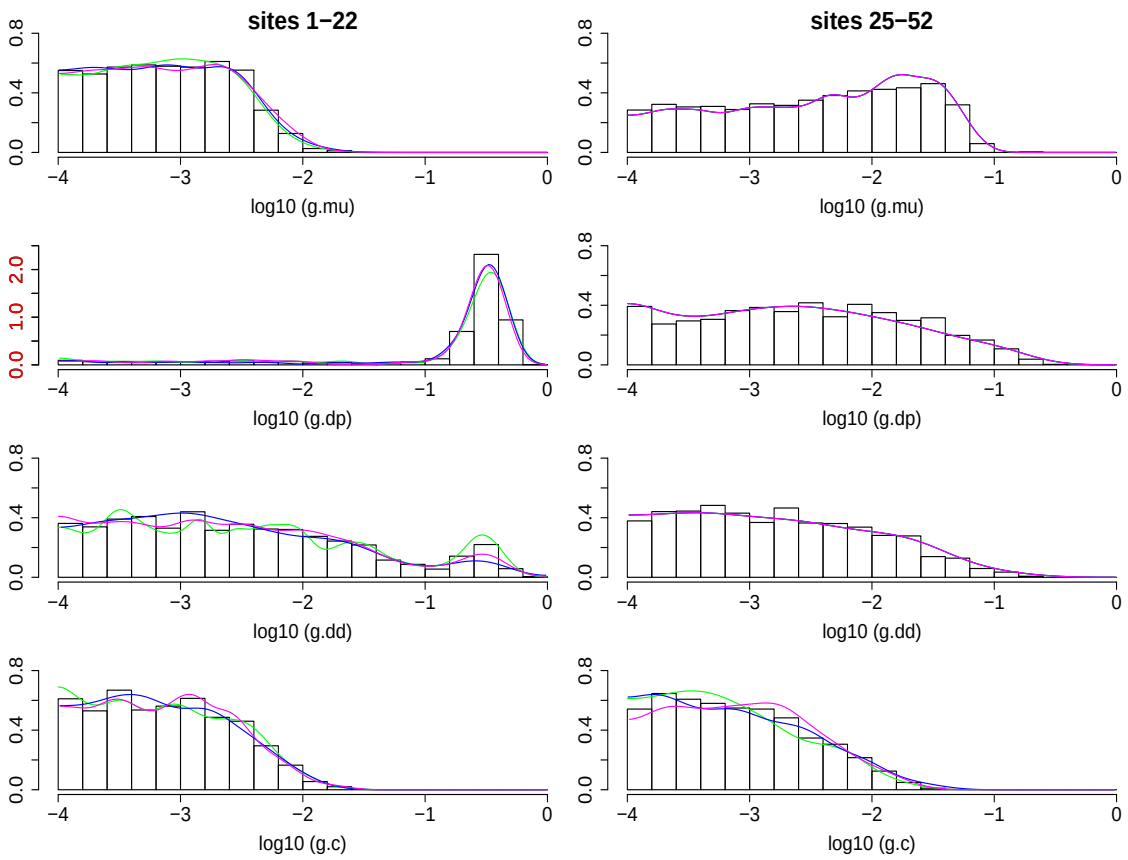

Figure 3.3: Posterior histograms (bars) and densities (curves) of scaled variance $\log_{10} g$ for both *FMR1* data sets under the Bayesian hierarchical multi-site model. The histograms pool MCMC samples from three runs together, whereas each density curve corresponds to one run. Unbiased cross validation is used to select the bandwidth for a Gaussian kernel density estimator. Note that the plot for $\log_{10}(g_{dp})$ has a vertical scale different from others.

confounding between two average de novo rates. They also seem to rule out the possibility that both de novo rates are large and point to possibly sharper estimates of the average of r_{dp} and r_{dd}. The 80% credible interval for this average (last row in Table 3.6) indeed has a width half of that of at least one of the average de novo rates.

We use the histograms shown in Figure 3.3 to infer the variability in the rates,

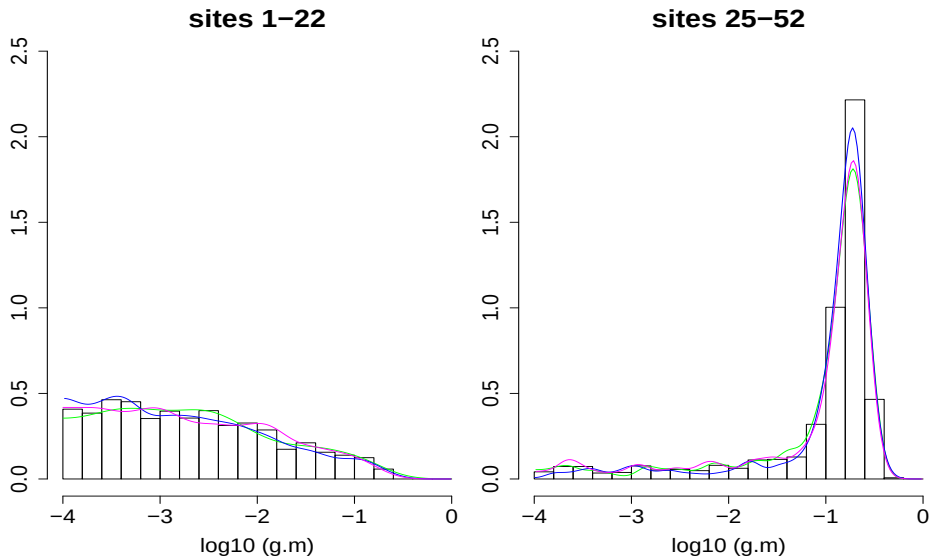

Figure 3.4: Posterior histograms (bars) and densities (curves) of measure of departure from temporal stationarity, $\log_{10}(g_m)$, for both *FMR1* data sets under the Bayesian hierarchical multi-site model. The histograms pool MCMC samples from three independent runs together, whereas each density curve corresponds to one run. Unbiased cross validation is used to select the bandwidth for a Gaussian kernel density estimator.

according to the guidelines provided in Table 3.2. Furthermore, the flatness of a density curve indicates the strength of evidence in the data: the flatter the density curve, the weaker the evidence. The results are summarised in Table 3.7.

The large variability in δ_p inferred from the density plots is intriguing, because the results from the single-site ML model also show large variability in de novo rates, although which de novo rate is variable depends on the constraint we impose to solve the stationarity equations. This result may suggest that some sites can have potentially quite different de novo rates than other sites. So we look at the posterior distributions of site-specific de novo rates (Figures 3.7–3.10). Sites 10, 14, 15 and 16 have clearly very different δ_p and somewhat different δ_d than the others among sites 1–22 (Figures 3.7 and 3.8). The daughter de novo rate at each of those four sites also

Table 3.6: 80% credible intervals of the means r of failure of maintenance rate $1-\mu$, parent de novo δ_p, daughter de novo rate δ_d, and inappropriate bisulfite conversion error rate c from the two *FMR1* data sets under the Bayesian hierarchical multi-site model, pooling results from three independent runs. The credible intervals for the average of r_{dp} and r_{dd} are also listed.

	Sites 1–22	Sites 25–52
$r_{1-\mu}$	$(0.015, 0.033)$	$(0.029, 0.054)$
r_c	$(0.007, 0.023)$	$(0.009, 0.033)$
r_{dp}	$(0.056, 0.302)$	$(0.031, 0.083)$
r_{dd}	$(0.040, 0.141)$	$(0.010, 0.067)$
$(r_{dp}+r_{dd})/2$	$(0.082, 0.193)$	$(0.031, 0.063)$

Table 3.7: Inferred variability in rates for each *FMR1* data set under the Bayesian hierarchical multi-site model. The rates are: failure of maintenance rate $1-\mu$, parent de novo rate δ_p, daughter de novo rate δ_d and inappropriate bisulfite conversion error rate c. The inference is based on the posterior distributions of $\log_{10} g$ (Figure 3.3). See guidelines in Table 3.2 for interpretation of the posterior distributions. The first data set is uninformative for variability in δ_d; the posterior distribution is essentially flat over $(-4, 0)$.

	Sites 1–22	Sites 25–52
$1-\mu$	very low	low–medium
δ_p	high	low–medium
δ_d	uninformative	low–medium
c	very low	very low

has more variation than other sites. Sites 25–52 do not show such a clear distinction (Figures 3.9 and 3.10), although site 33 seems to have a slightly more variable δ_p than other sites; these observations are consistent with the density curves of $\log_{10} g_{dp}$ and $\log_{10} g_{dp}$ (right column in Figure 3.3), which do not show evidence for large variability.

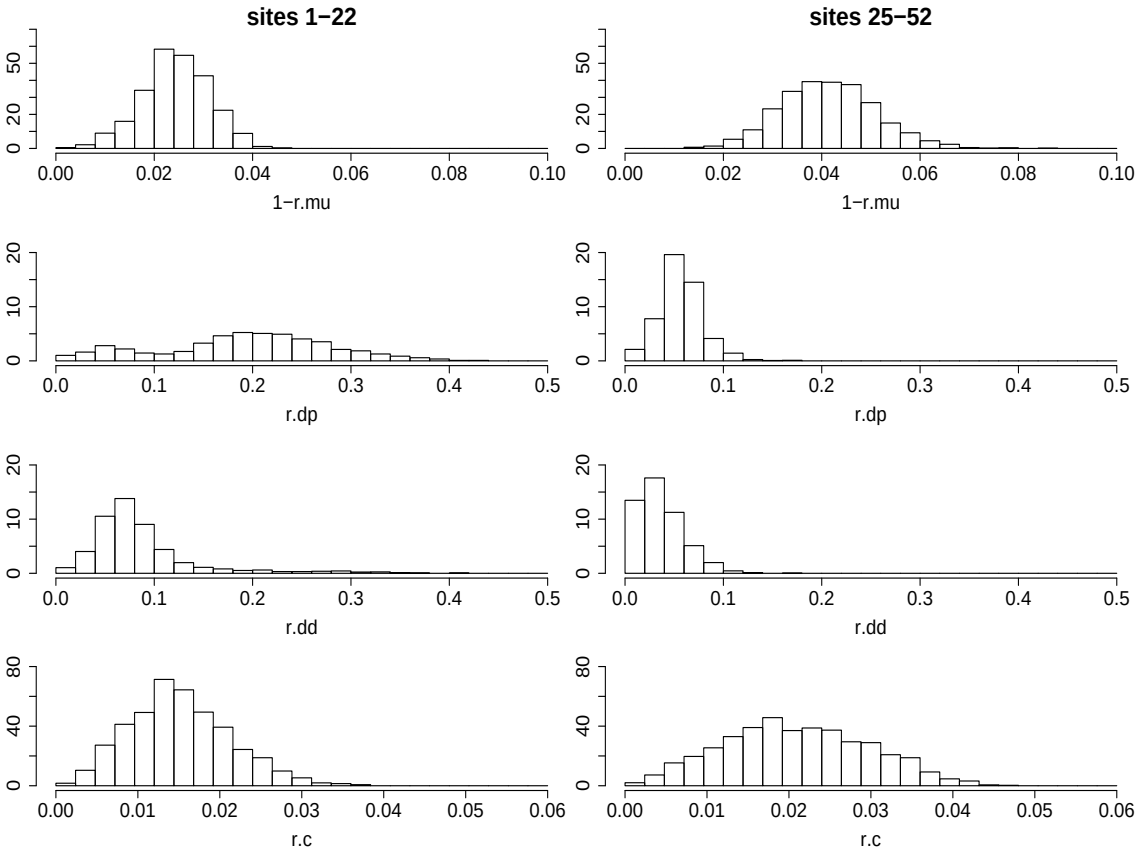

Figure 3.5: Posterior distributions of mean r for both *FMR1* data sets under the Bayesian hierarchical multi-site model, pooling results from three independent runs.

The large variability in δ_p, and maybe also δ_d, inferred under the multi-site model, may not be captured well with by a beta distribution even with a large variance (see the bottom right plot in Figure 3.1); or even if such a beta distribution could accommodate bimodality, we would still like to know *which* sites might be "outliers", in addition to the *existence* of "outliers". Therefore, we consider in the next section a mixture model for estimating de novo rates.

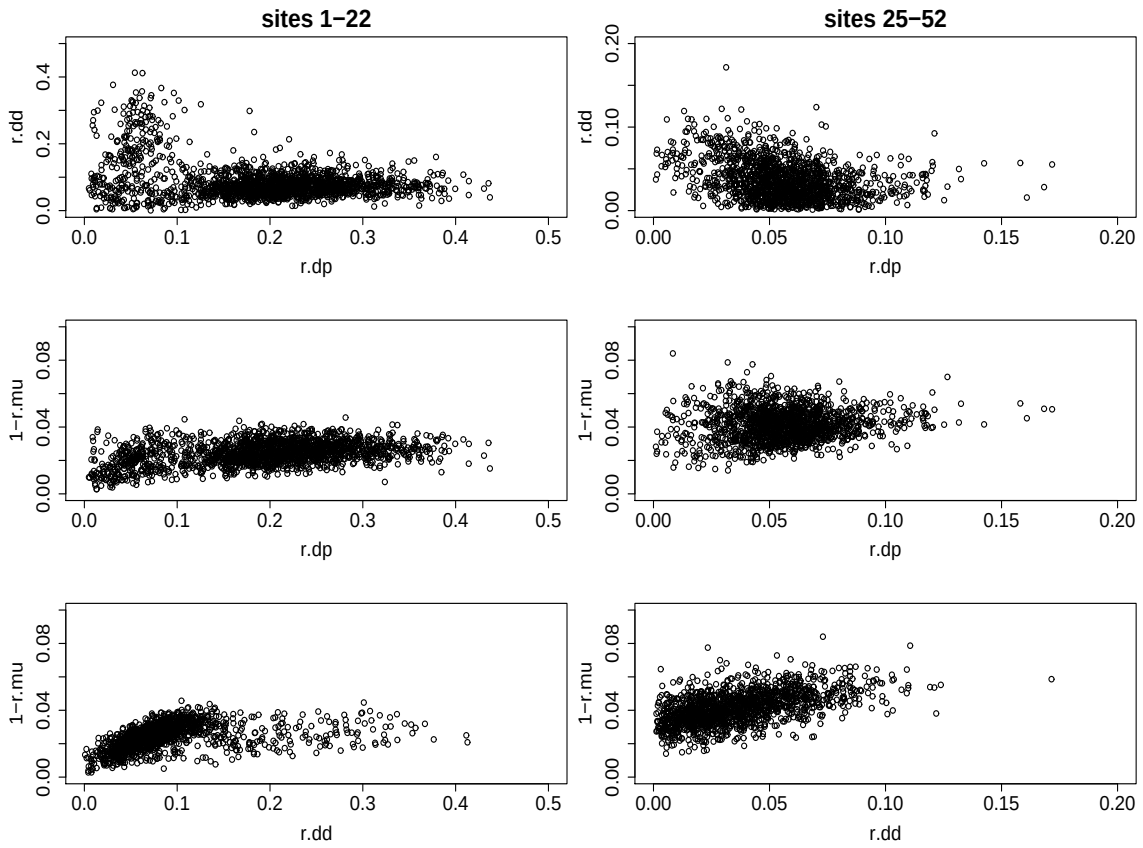

Figure 3.6: Scatter plots of MCMC samples of rs for both *FMR1* data sets under the Bayesian hierarchical multi-site model, pooling results from three independent runs.

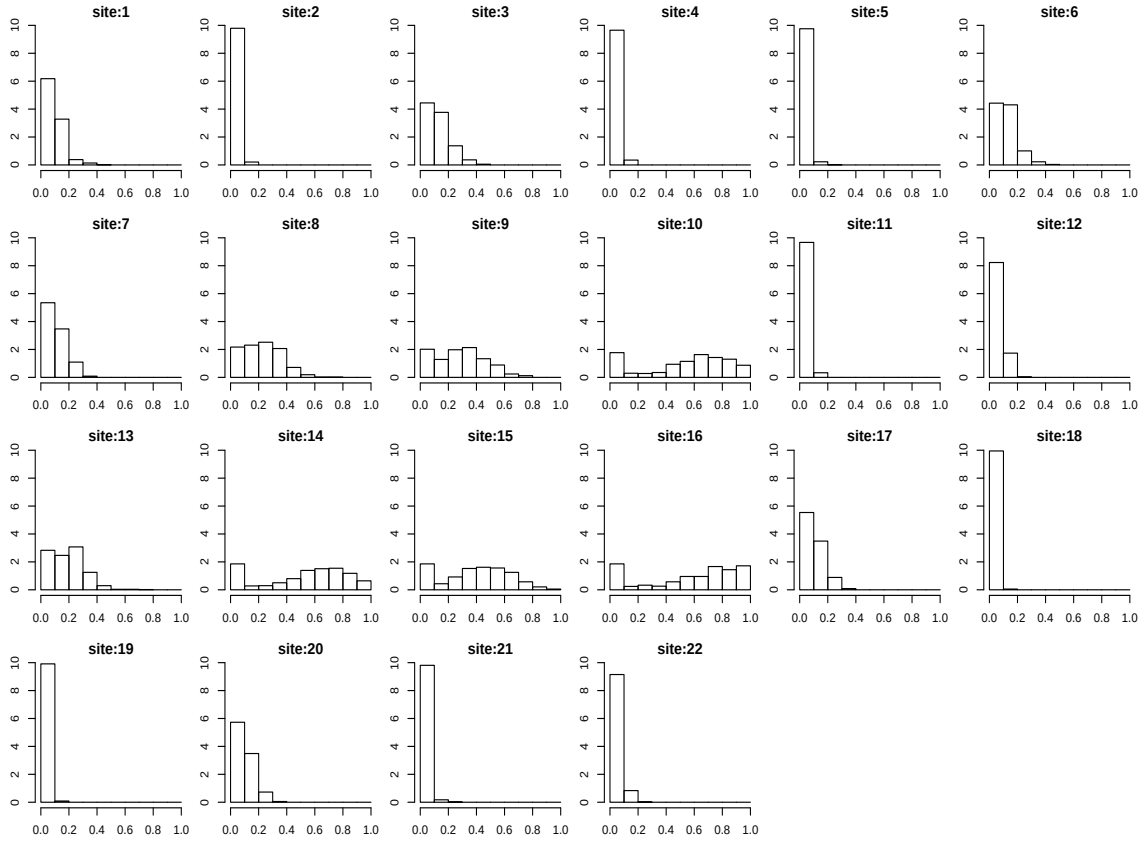

Figure 3.7: Posterior distributions of the parent de novo rate δ_p for *FMR1* data at sites 1–22 under the Bayesian hierarchical multi-site model.

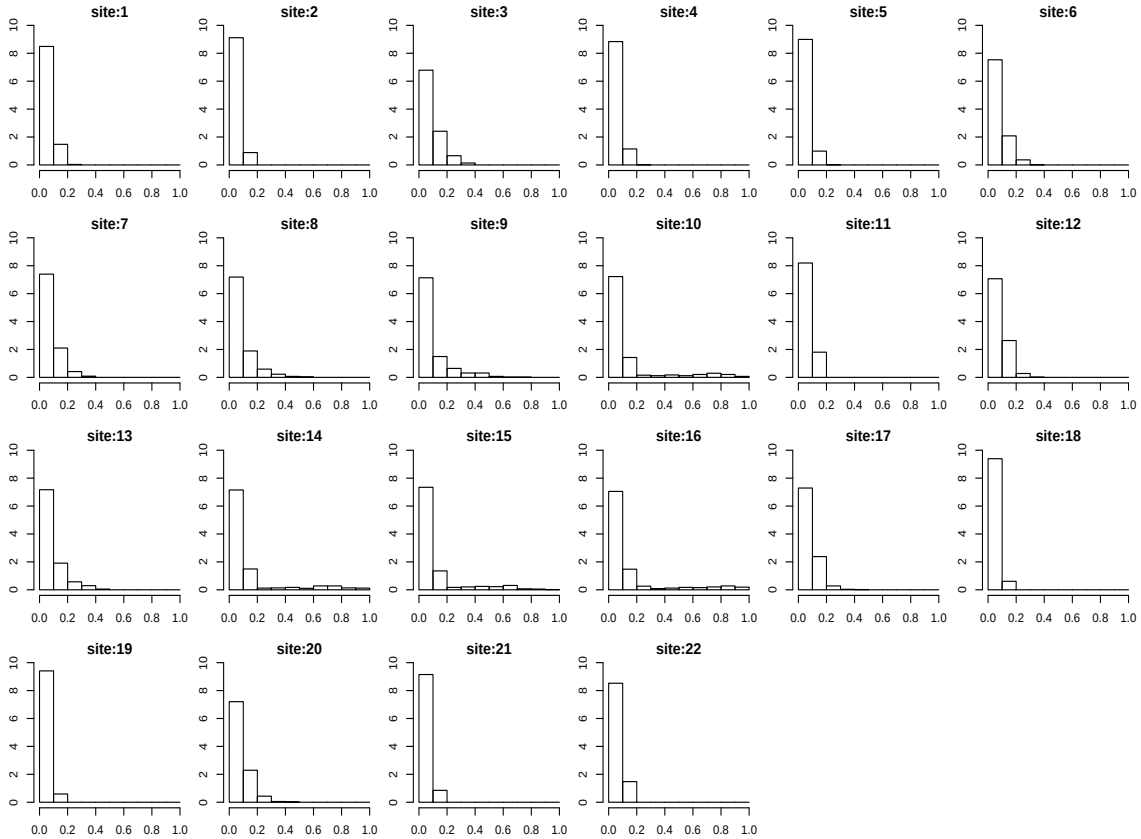

Figure 3.8: Posterior distributions of the daughter de novo rate δ_d for *FMR1* data at sites 1–22 under the Bayesian hierarchical multi-site model.

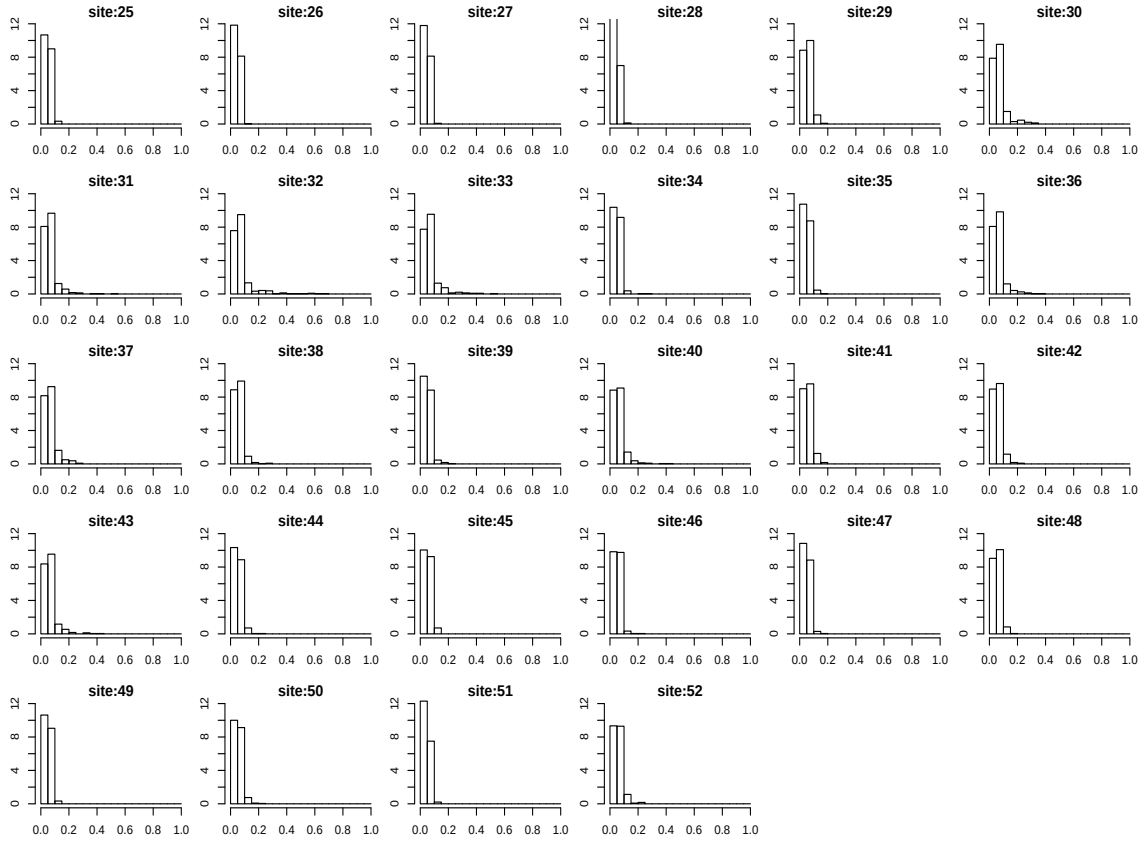

Figure 3.9: Posterior distributions of the parent de novo rate δ_p for *FMR1* data at sites 25–52 under the Bayesian hierarchical multi-site model.

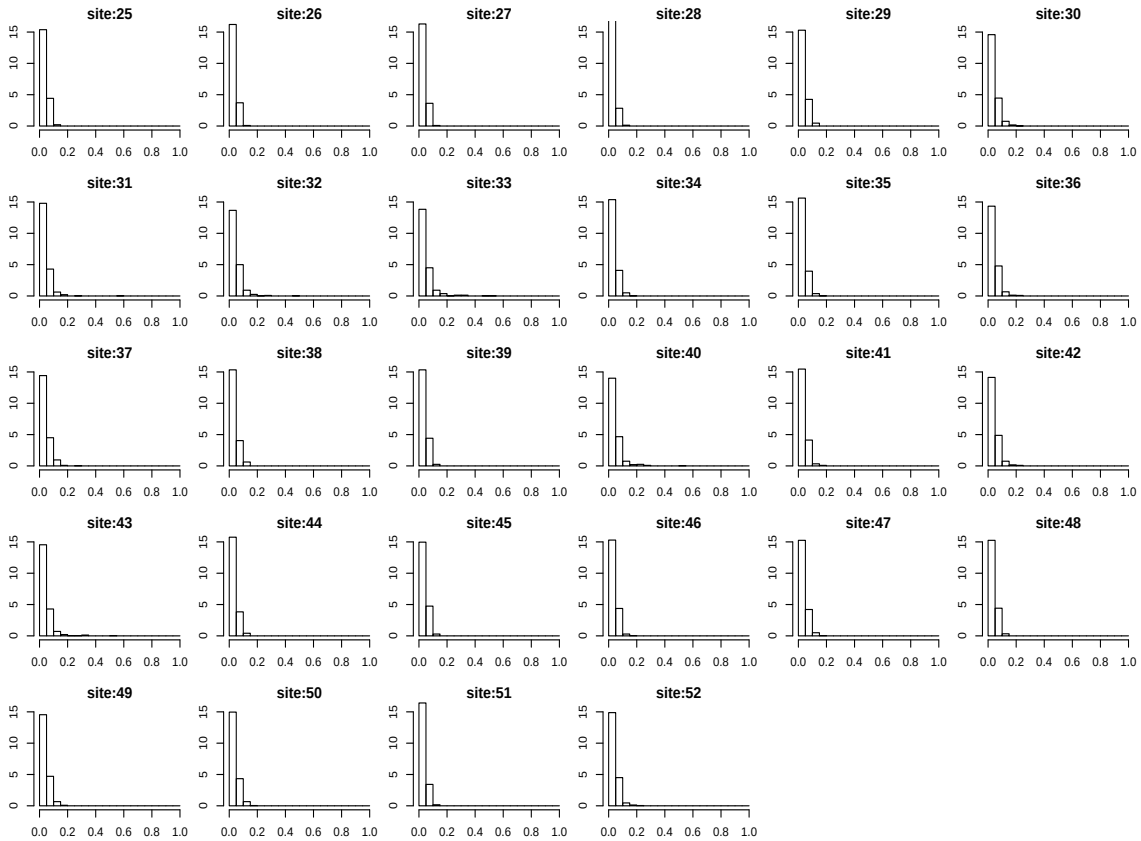

Figure 3.10: Posterior distributions of the daughter de novo rate δ_d for *FMR1* data at sites 25–52 under the Bayesian hierarchical multi-site model.

Before moving on to the mixture model, we look at the level of temporal stationarity in each data set. The posterior distributions of the measure of departure $\log_{10}(g_m)$ indicate consistency with temporal stationarity at sites 1–22, but clear departure at sites 25–52 (Figure 3.3). To identify which sites are not in stationarity, we compute the expected methylation probability m_j under temporal stationarity, defined by equation (3.2.1), using MCMC samples of the rates of methylation events. We then compare these expectations with the observed methylation density at each site (Figures 3.11 and 3.12). At sites 1–22 (Figure 3.11), the observed methylation density (the red line) lies in the middle of the distribution of the expected values at all sites, suggesting that those 22 sites are in stationarity. At sites 25–52, however, the observed methylation density is in the tail of the distribution of the expected values at a third of the sites (Figure 3.12). Some of these non-stationary sites, for example, sites 25–29 and 31–33, are physically close to each other. Furthermore, sites 25–29 have low methylation probabilities (between 20–50%), whereas sites 31–33 have relatively high methylation probabilities (above 95%). The non-random ordering of the sites may indicate spatial dependence or other covariates.

To confirm the observation that multiple sites in the second data set deviate from stationarity, we removed sites 32 and 33 from this data set and re-ran the MCMC program. These two sites not only are non-stationary, but also seem to have somewhat different estimated de novo rates (see Figure 3.9) and may be considered as "outliers". The posterior distribution of $\log_{10}(g_m)$ (not shown) shifts slightly to the left, compared with that before removing those two sites, suggesting that sites 32 and 33 contribute to the inferred departure from stationarity and that there are other non-stationary sites as well. The inference of other parameters shows essentially no change.

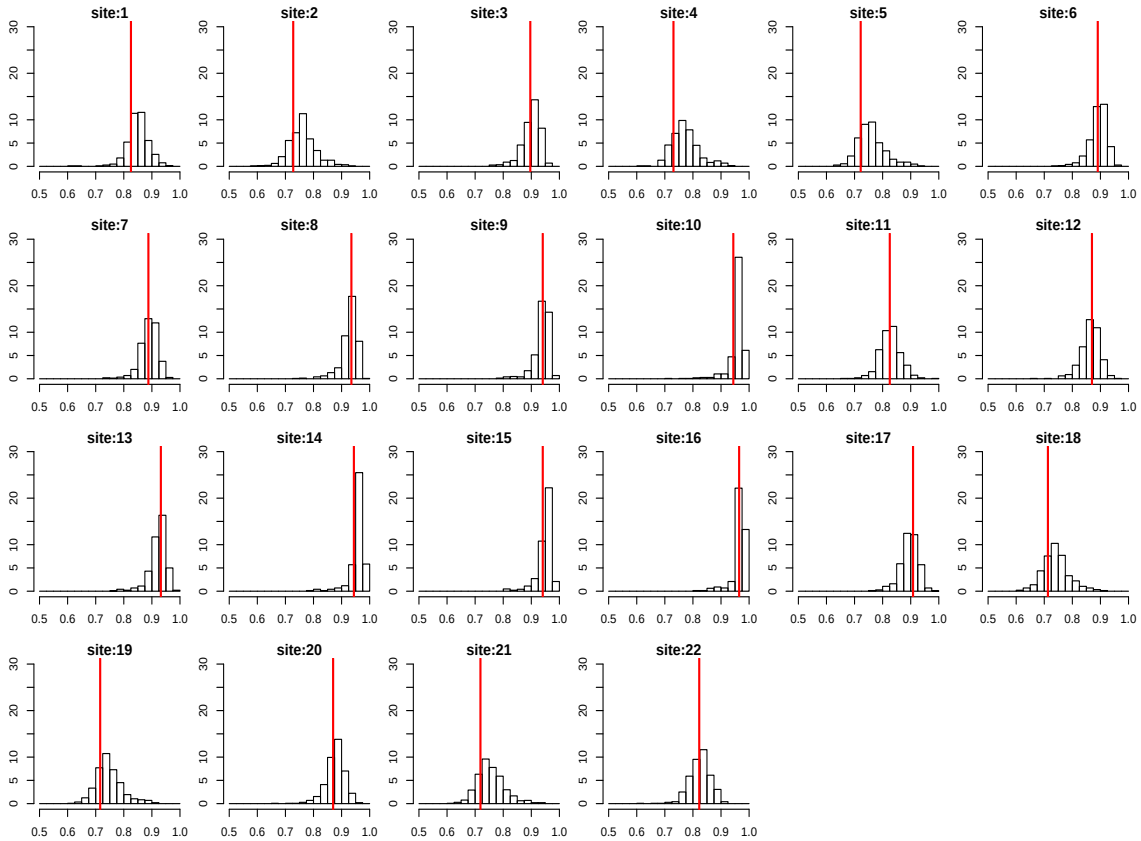

Figure 3.11: The level of departure from temporal stationarity at individual sites 1–22 under the Bayesian hierarchical multi-site model. Departure is reflected in the distance between the observed (red line) and the expected (histogram) methylation probabilities at a site. The expectations are calculated using MCMC samples of the rates of methylation events. The observed value is the empirical methylation density.

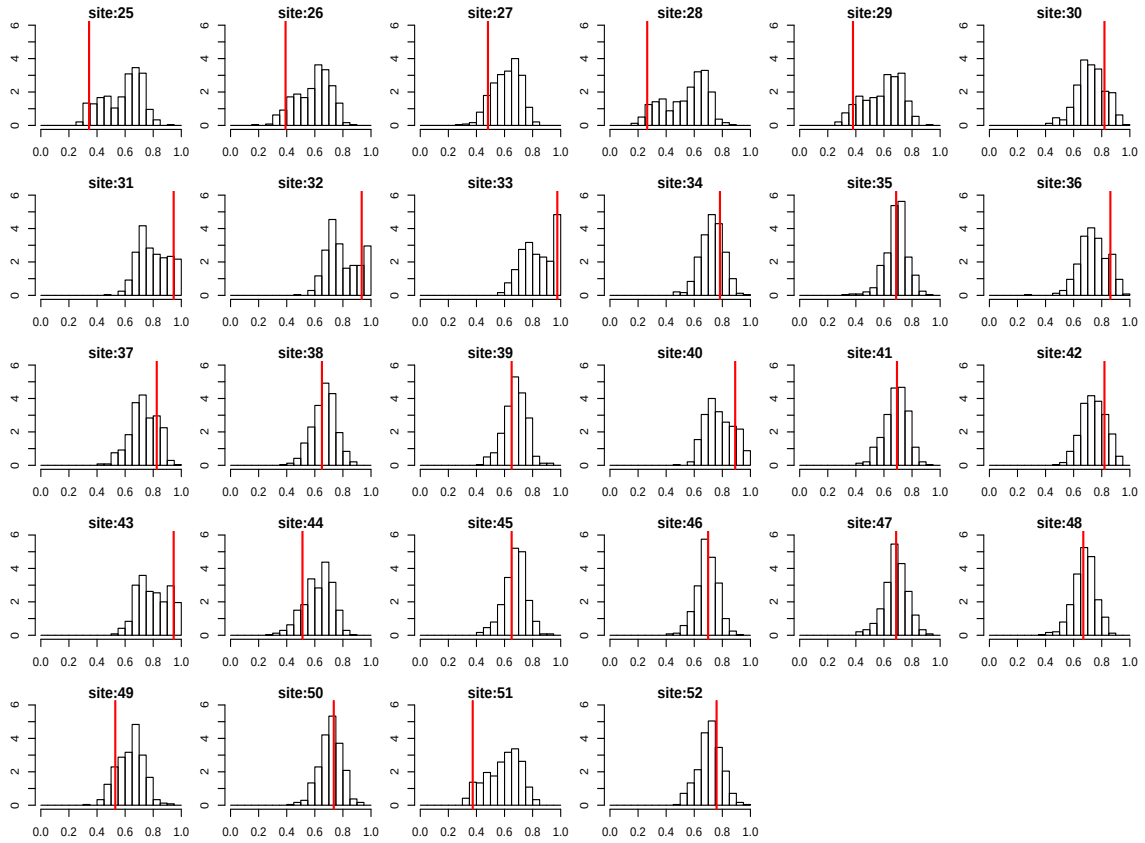

Figure 3.12: The level of departure from temporal stationarity at individual sites 25–52 under the Bayesian hierarchical multi-site model. Departure is reflected in the distance between the observed (red line) and the expected (histogram) methylation probabilities at a site. The expectations are calculated using MCMC samples of the rates of methylation events. The observed value is the empirical methylation density.

3.4.2 Identifying outlier sites in the data – a mixture model

The model and the MCMC procedure

Since the number of outlier sites is likely to be low, there may not be enough information to accurately estimate the de novo rates at those sites. A Unif$(0, 1)$ distribution, or Beta$(1, 1)$, is assigned to δ_p and δ_d at these sites. Additionally, the small data set available may not allow for separate estimation of two mixing proportions, one for each type of de novo rate. Thus, a common mixing proportion is used.

We use the binary variable Z_j to denote whether a site is an outlier, 1 for being an outlier and 0 otherwise. The probability of a site being an outlier is also the mixing proportion w. The parent and daughter de novo rates are mixtures of a Beta(r, g) and a Unif$(0, 1)$ distribution. Thus,

$$w \sim \text{Unif}\ (0, 1), \tag{3.4.1}$$

$$Z_j | w \sim \text{Bernoulli}\ (w); \tag{3.4.2}$$

and

$$p(\delta_{p_j} | Z_j, r_{dp}, g_{dp}) = f(\delta_{p_j}; r_{dp}, g_{dp})^{(1-Z_j)} h(\delta_{p_j}; 0, 1)^{Z_j}, \tag{3.4.3}$$

$$p(\delta_{dj} | Z_j, r_{dd}, g_{dd}) = f(\delta_{dj}; r_{dd}, g_{dd})^{(1-Z_j)} h(\delta_{dj}; 0, 1)^{Z_j}, \tag{3.4.4}$$

where f and h are density functions of beta and uniform, respectively.

Previously in Section 3.2.2, we employed a Metropolis-Hastings sampler for μ, δ_p and δ_d to incorporate temporal stationarity, using the full conditional distribution under the no-stationarity model to generate proposals. Under the mixture model, we condition on the outlier status Z_j and adopt a similar strategy for de novo rates. Only the full conditional distributions for de novo rates need to changed to accommodate the mixture. Take the parent de novo rate for example. Let α_p and β_p be the parameters converted from r_{dp} and g_{dp} at the non-outlier sites, and α_o and β_o the parameters at the outlier sites. Since the distribution assumption for de novo rates

at the outlier sites is Unif$(0,1)$, we have $\alpha_o = \beta_o = 1$. The full conditional for δ_p becomes

$$\text{Beta}(\alpha_p(1-Z_j) + \alpha_o Z_j + \sum_{i=1}^{N} Q_{ij}(1-P_{ij}), \beta_p(1-Z_j) + \beta_o Z_j + \sum_{i=1}^{N}(1-Q_{ij})(1-P_{ij})). \tag{3.4.5}$$

If the current value is δ_{p_j}' and the proposal $\delta_{p_j}^*$ is now generated from (3.4.5), we accept $\delta_{p_j}^*$ with probability $\min(1, A)$, where

$$A = \frac{f(m_j; \alpha_m(\mu_j, \delta_{p_j}^*, \delta_{dj}), \beta_m(\mu_j, \delta_{p_j}^*, \delta_{dj}))}{f(m_j; \alpha_m(\mu_j, \delta_{p_j}', \delta_{dj}), \beta_m(\mu_j, \delta_{p_j}', \delta_{dj}))}. \tag{3.4.6}$$

This can also be seen from the posterior distribution of δ_{p_j}, which is

$$p(\delta_{p_j}|\cdot) \propto p(\delta_{p_j}|Z_j, r_{dp}, g_{dp}) p(m_j|\mu_j, \delta_{p_j}, \delta_{dj}, g_m) \prod_{i=1}^{N} p(Q_{ij}|P_{ij}, \delta_{p_j}) \tag{3.4.7}$$

$$= f(\delta_{p_j}; r_{dp}, g_{dp})^{(1-Z_j)} f(\delta_{p_j}; \alpha_o, \beta_o)^{Z_j} f(m_j|r_{m,j}, g_m) \prod_{i=1}^{N} \delta_{p_j}^{Q_{ij}(1-P_{ij})}$$

$$\times (1-\delta_{p_j})^{(1-Q_{ij})(1-P_{ij})} \tag{3.4.8}$$

$$= \delta_{p_j}^{\alpha_p(1-Z_j)+\alpha_o Z_j + \sum_{i=1}^{N} Q_{ij}(1-P_{ij})-1} (1-\delta_{p_j})^{\beta_p(1-Z_j)+\beta_o Z_j + \sum_{i=1}^{N}(1-Q_{ij})(1-P_{ij})-1}$$

$$\times f(m_j|r_{m,j}, g_m). \tag{3.4.9}$$

Bayes factors to assess strength of evidence

We compute Bayes factors (Kass and Raftery, 1995) to assess the strength of evidence for the existence of outliers. Bayes factors have been used widely in model selection. Consider two models M_0 and M_1 for data X. The Bayes factor of M_1 over M_0 is then defined as

$$\text{BF} = \frac{p(X|M_1)}{p(X|M_0)}, \tag{3.4.10}$$

which can also be computed as the ratio of posterior odds and prior odds as follows

$$\text{BF} = \frac{p(M_1|X)}{p(M_0|X)} \Big/ \frac{p(M_1)}{p(M_0)}. \tag{3.4.11}$$

Kass and Raftery (1995) further summarised interpretation guidelines suggested by Jeffreys (1961): there is strong evidence supporting M_1 in comparison of M_0 if the value is above 3; very strong if above 10. However, Bayes factors do not take into account how likely the two models are a priori in the comparison. Hence, even if we do not think M_1 is likely to happen, the Bayes factor can still be high; the posterior probability of M_1 will be low though.

Here, we are particularly interested in whether the data contain enough information for us to reject the null hypothesis of no outlier. Specifically, we consider the event of having k outliers, denoted by E_k, versus the event of having no outliers, denoted by E_0. Then the Bayes factor for this comparison is computed as

$$\mathrm{BF}_k = \frac{p(\{\mathbf{x},\mathbf{y}\}|E_k)}{p(\{\mathbf{x},\mathbf{y}\}|E_0)} = \frac{p(E_k|\{\mathbf{x},\mathbf{y}\})}{p(E_0|\{\mathbf{x},\mathbf{y}\})} \Big/ \frac{p(E_k)}{p(E_0)}. \tag{3.4.12}$$

With a uniform prior on the mixing proportion w, the two events E_0 and E_k are equally likely *a priori*. Hence the prior odds $p(E_k)/p(E_0)$ is 1. Posterior samples of Z_j can be used to compute the terms in the posterior odds $p(E_k|\{\mathbf{x},\mathbf{y}\})/p(E_0|\{\mathbf{x},\mathbf{y}\})$:

$$p(E_k|\{\mathbf{x},\mathbf{y}\}) = \frac{1}{M}\sum_{m=1}^{M} \mathbf{1}^{(m)}(\sum_{j=1}^{S} Z_j = k) \equiv \frac{1}{M} N_k \tag{3.4.13}$$

where m indicates the m-th MCMC sample and M the total number of MCMC samples. N_k is just the total number of MCMC samples that have k outliers. The Bayes factor can then be simplified as

$$\mathrm{BF}_k = \frac{N_k}{N_0}. \tag{3.4.14}$$

A large value of BF_k would provide strong evidence for event E_k, which is the existence of k outliers.

The uniform prior on the mixing proportion w, used in our mixture model, is not a satisfactory prior for the purpose here: (i) it does not help us test the null hypothesis of no outliers; and (ii) having 1 outlier being equally likely as having 22 outliers is also counter-intuitive. To examine the impact of priors, we will consider other priors

that put a considerable amount of probability mass on the value 0, and compute the posterior probability $p(E_k|\{\mathbf{x},\mathbf{y}\})$ using the Bayes factors computed above as the weights:

$$p(E_k|\{\mathbf{x},\mathbf{y}\}) = \frac{\mathrm{BF}_k \times \tilde{p}(E_k)}{\sum_{k'} \mathrm{BF}_{k'} \times \tilde{p}(E_{k'})}, \tag{3.4.15}$$

in which $\tilde{p}(E_k)$ can be any prior.

3.4.3 Analysis of the FMR1 data under the mixture model

We carried out three independent runs of the MCMC program under the mixture model. The specifications and initial values are listed in Tables 3.8 and 3.9, respectively. The initial values range from being reasonable (Run 1) – consistent at least in magnitude with estimates from existing studies, to being nearly the opposite (Run 3).

Table 3.8: MCMC specifications of each run under the Bayesian hierarchical mixture model for each *FMR1* data set.

		Run Length (Iterations)	Burn-in	Sampling Interval (Iterations)
Data set 1	Run 1	480,000	20%	800
	Run 2	480,000	20%	800
	Run 3	480,000	20%	800
Data set 2	Run 1	200,000	20%	400
	Run 2	200,000	20%	400
	Run 3	800,000	20%	1600

Table 3.9: Initial values of mean r and scaled variance g for each of the three independent MCMC runs under the Bayesian hierarchical mixture model.

	Initial Values					
	Run 1		Run 2		Run 3	
	r	g	r	g	r	g
m	N/A	0.01	N/A	0.01	N/A	0.2
μ	0.95	1/21	0.5	0.01	0.05	0.2
δ_{dp}	0.05	1/21	0.5	0.01	0.9	0.2
δ_{dd}	0.05	1/21	0.5	0.01	0.9	0.2
c	0.03	0.01	0.05	0.01	0.005	0.2

Inference about the existence of outliers

Recall that outliers refer to sites that may have very different de novo rates from other sites. Histograms of the mixing proportion w in Figure 3.13 suggest possible existence of outliers in the first data set and maybe no outlier in the second.

We compute Bayes factors as described in Section 3.4.2 to assess the strength of evidence for the existence of k outliers. These values, listed in Table 3.11, provide strong evidence for 3-5 outliers in the first data set and very strong evidence for 4 outliers, which is consistent with the observations under the non-mixture model and the histogram of w under the uniform prior. Even when a large probability is placed on no outlier a priori, under Priors 2 and 3 for example, the posterior probability of having 4 outliers is still the highest.

Data at sites 25–52, however, provide little evidence for the existence of outliers; the highest Bayes factor is only 1.6. When the prior probability of no outlier is 0.5, the posterior probability of no outliers can be larger than 0.5.

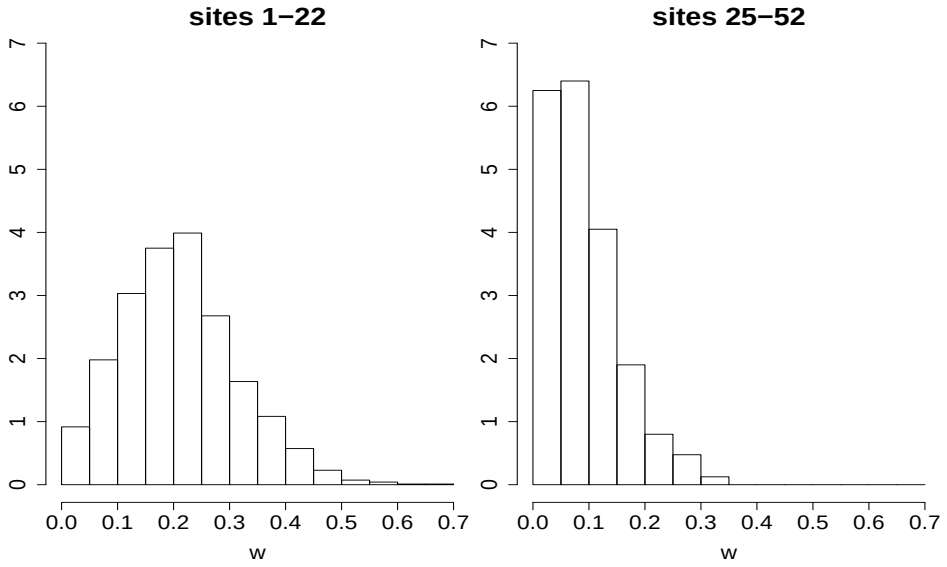

Figure 3.13: Posterior distributions of the mixing proportion w for each *FMR1* data set under the Bayesian hierarchical mixture model.

Table 3.10: Priors on having k outliers where $k = 0, \cdots, 10$. Values larger than 10 are not considered for k because the number of sites in each data set is only between 20 and 30.

Number of outliers	$p(E_k)$		
k	Prior 1	Prior 2	Prior 3
0	1/11	0.5	0.5
1	1/11	0.05	$10/55 \times 0.5$
2	1/11	0.05	$9/55 \times 0.5$
3	1/11	0.05	$8/55 \times 0.5$
4	1/11	0.05	$7/55 \times 0.5$
5	1/11	0.05	$6/55 \times 0.5$
6	1/11	0.05	$5/55 \times 0.5$
7	1/11	0.05	$4/55 \times 0.5$
8	1/11	0.05	$3/55 \times 0.5$
9	1/11	0.05	$2/55 \times 0.5$
10	1/11	0.05	$1/55 \times 0.5$

Table 3.11: Bayes factors and posterior probabilities under the Bayesian hierarchical mixture model for each *FMR1* data set. Largest values in each column are in boldface. See Section 3.4.2 for definition and computation. The priors are listed in Table 3.10.

Number of Outliers (k)	Bayes Factor (BF_k)		Posterior Probability $p(E_k \mid \{\mathbf{x}, \mathbf{y}\})$					
			Prior 1		Prior 2		Prior 3	
	1–22	25–52	1–22	25–52	1–22	25–52	1–22	25–52
0	1	1	0.02	0.18	0.20	**0.69**	0.16	**0.58**
1	1.7	1.5	0.04	0.28	0.03	0.10	0.08	0.16
2	3.1	**1.6**	0.07	**0.30**	0.06	0.11	0.08	0.15
3	7.5	1.0	0.18	0.16	0.15	0.06	0.18	0.07
4	**12.4**	0.0	**0.30**	0.06	**0.25**	0.02	**0.26**	0.02
5	9.5	0.0	0.23	0.02	0.19	< 0.01	0.17	< 0.01
6	4.4	0.0	0.10	< 0.01	0.09	< 0.01	0.07	< 0.01
7	1.5	0.0	0.03	< 0.01	0.03	< 0.01	0.02	0.00
8	0.5	0.0	0.01	0.00	< 0.01	0.00	< 0.01	0.00
9	0.2	0.0	< 0.01	0.00	< 0.01	0.00	< 0.01	0.00
10	0.0	0.0	< 0.01	0.00	< 0.01	0.00	< 0.01	0.00

Inference about rates of methylation events

Histograms of rs are plotted in Figure 3.14 and 80% credible intervals summarised in Table 3.12. Compared with the 80% credible intervals under the non-mixture model in Table 3.6, we can see remarkable consistency in estimates of $r_{1-\mu}$ and r_c between the two models. Excluding the outlier sites greatly changes the inference of r_{dp} and r_{dd} for the first data set and somewhat for the second. Specifically, the posterior distribution of r_{dp} for the first data set is much less spread out and peaks around 0.05.

Table 3.12: 80% credible intervals of the means r of failure of maintenance rate $1-\mu$, parent de novo δ_p, daughter de novo rate δ_d, and inappropriate bisulfite conversion error rate c for each *FMR1* data set under the Bayesian hierarchical mixture model, pooling results from multiple independent runs. Hence r_{dp} and r_{dd} are parameters at non-outlier sites. The credible intervals for the average of r_{dp} and r_{dd} are also listed.

	Sites 1–22	Sites 25–52
$r_{1-\mu}$	$(0.015, 0.035)$	$(0.032, 0.058)$
r_c	$(0.006, 0.024)$	$(0.008, 0.030)$
r_{dp}	$(0.027, 0.143)$	$(0.035, 0.081)$
r_{dd}	$(0.032, 0.137)$	$(0.015, 0.065)$
$(r_{dp} + r_{dd})/2$	$(0.044, 0.120)$	$(0.034, 0.061)$

Inference about variability in the rates

We plot the histograms of $\log_{10} g$s in Figure 3.15. The conclusions (summarised in Table 3.13) are almost the same as under the non-mixture model (Table 3.7), except that g_{dp} and g_{dd} now reflect variability in de novo rates across non-outlier sites. After excluding the outlier sites that may have high parent de novo rates, we do have have enough information for variation in δ_p at regular sites. This lack of information

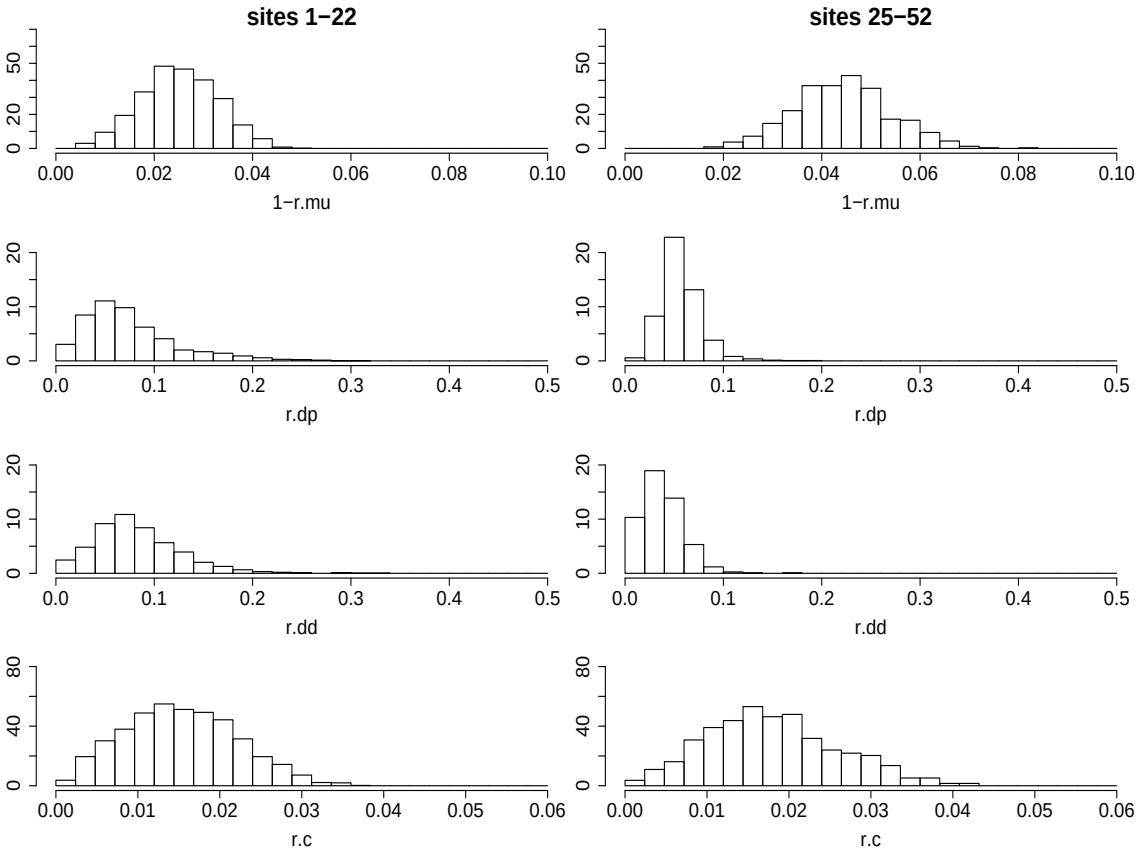

Figure 3.14: Posterior distributions of mean r for each *FMR1* data set under the Bayesian hierarchical mixture model, pooling results from multiple independent runs.

is probably due to the small sample size and the confounding between parent and daughter de novo rates.

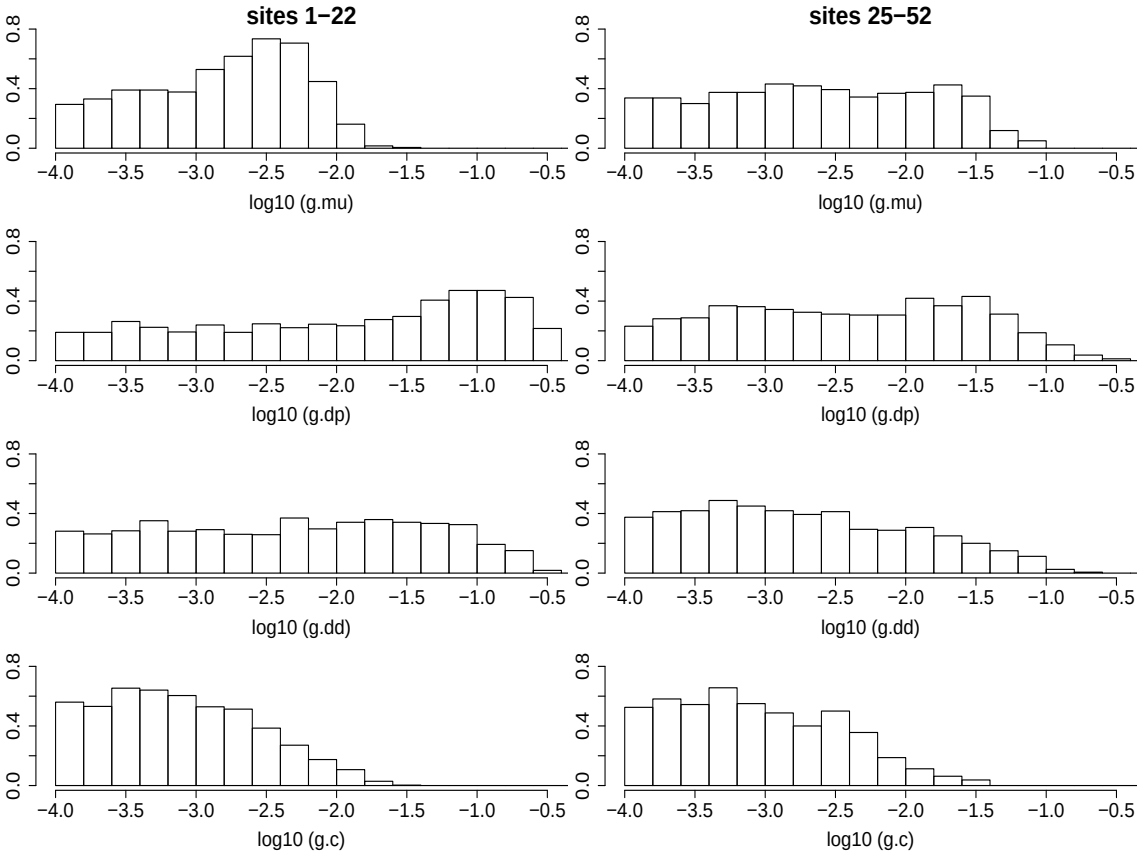

Figure 3.15: Posterior distributions of scaled variance $\log_{10} g$ for each *FMR1* data set under the Bayesian hierarchical mixture model, pooling results from multiple independent runs. Note that $\log_{10}(g_{dp})$ and $\log_{10}(g_{dp})$ refer to the non-outlier sites.

Table 3.13: Inferred variability in rates for each *FMR1* data set under the Bayesian hierarchical mixture model. The rates are: failure of maintenance rate $1 - \mu$, parent de novo rate δ_p, daughter de novo rate δ_d and inappropriate bisulfite conversion error rate c. The inference is based on the posterior distributions of $\log_{10} g$ (Figure 3.15). See guidelines in Table 3.2 for interpretation of the posterior distributions. The first data set is uninformative for variability in δ_p and δ_d; the posterior distributions are essentially flat over $(-4, 0.5)$.

	Sites 1–22	Sites 25–52
$1 - \mu$	very low	low–medium
δ_p	uninformative	low–medium
δ_d	uninformative	low–medium
c	very low	very low

Inference about temporal stationarity

Results from the mixture model shown in Figure 3.16 continue to suggest that the first data set is consistent with temporal stationarity and that the second data set has departed from this assumption.

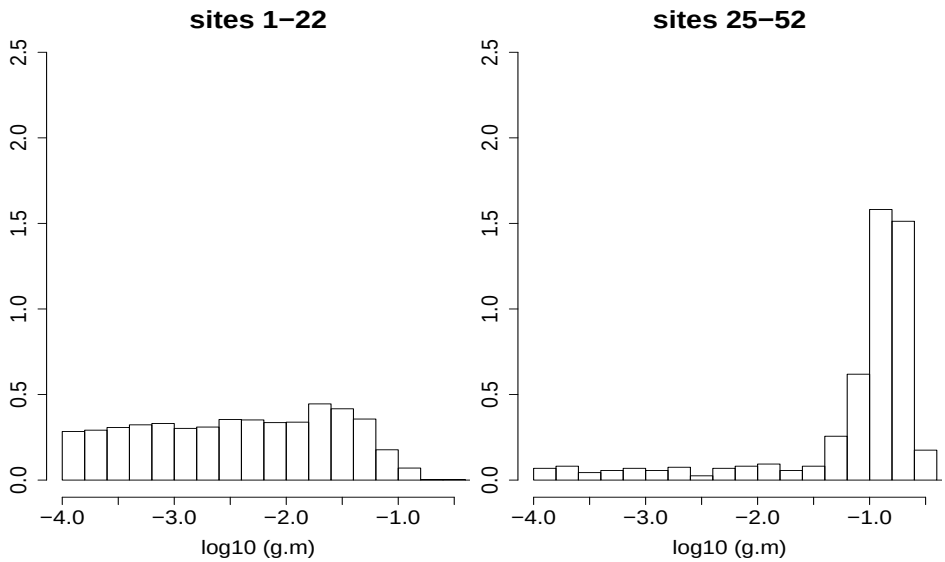

Figure 3.16: Posterior distributions of the measure of departure from temporal stationarity, $\log_{10}(g_m)$, for each *FMR1* data set under the Bayesian hierarchical mixture model from multiple independent runs.

3.5 Summary and discussion

3.5.1 Summary of the data analysis results

With the hierarchical model, we obtain sharp interval estimates for the mean of the failure of maintenance $1 - \mu$ and the mean of the inappropriate conversion error rate c. Specifically, the 80% credible interval for $r_{1-\mu}$ is $(0.015, 0.035)$ for sites 1–22 and $(0.032, 0.058)$ for sites 25–52. That for r_c is $(0.006, 0.024)$ for sites 1–22 and $(0.008, 0.030)$ for sites 25–52. These two rates also have low variability across CpG

sites. For inference of the de novo rates, data at sites 1–22 suggest that 3–5 CpG sites may have large de novo rates. The second data set does not provide much evidence for the existence of outliers though. The scatter plot of the estimated means of the two de novo rates shows confounding, hence the interval estimate of the average of these two means is much sharper than estimating them separately. The 80% credible interval of this average is $(0.044, 0.120)$ for sites 1–22 and $(0.034, 0.061)$ for sites 25–52. The first data set is not informative for variability in either de novo rate after the outlier sites are excluded, whereas the second one suggests low to medium variability in both de novo rates. Additionally, while the first data set is consistent with temporal stationarity, the second one has clearly departed from this assumption.

3.5.2 Advantages and disadvantages of the extended multi-site models

One of the main advantages of the Bayesian hierarchical multi-site models is the use of the multi-site information. Simultaneous inference enables us to infer the latent strand type and pre-replication parent strand in a double-stranded sequence. A direct benefit from this is its capability to estimate the two types of bisulfite conversion error rate, b and c, and the ease of this estimation. Neither the method in Laird et al. (2004) or the single-site ML method (Genereux et al., 2005) is able to cope with the existence of errors, whereas our multi-site models applied to the *FMR1* data give consistent estimates for c under different assumptions, in addition to estimating other parameters at the same time. These estimates for c are consistent with those obtained in the lab (Genereux et al., 2008), showing the potential of the multi-site models becoming a reliable and convenient alternative to lab approaches.

The other benefit is that the multi-site models have the ability of distinguishing between parent and daughter de novo events and hence estimating separately the two de novo rates. However, sample size limits this capability; we saw confounding in our estimates. In contrast, the two de novo rates are inherently unidentifiable in existing methods: Laird et al. (2004) and Genereux et al. (2005) imposed additional

constraints such as $\delta_p = 0$ or $\delta_p = \delta_d$.

Secondly, hierarchical modelling enables us to borrow information from other sites in inference. By assuming that rates at different sites come from the same distribution, hierarchical models can significantly reduce the dimensionality of the parameter space, which is particularly desirable for the small sample size. This approach also makes our analysis one of the first studies to incorporate and estimate the *true* variation in a rate as a parameter, rather than relying on variation in the parameter estimates.

Thirdly, the multi-site models relax the stationarity assumption that underlie most of current approaches (Otto and Walbot, 1990; Pfeifer et al., 1990; Genereux et al., 2005). By allowing for deviation from this assumption, the multi-site models are flexible; they can be applied to genomic regions where stationarity does not hold and in analysis of cancer cells whose DNA methylation densities often change dramatically over rounds of cell division (see, for instance, Foster et al., 1998).

Chapter 4

MODELS FOR INVESTIGATION OF PROCESSIVITY OF METHYLATION ENZYMES

Having studied the rates of methylation events, we now turn our attention to inference concerning the enzymes that generate those events and help shape the methylation patterns during the transmission process in somatic cells. Two types of enzymes have been identified to be associated with this process: Dnmt1, commonly known as the "maintenance" methyltransferase; and Dnmt3a and Dnmt3b, commonly known as the "de novo" methyltransferases (see Dean et al., 2005 for a review of those enzymes). Since many enzymes that bind to molecules act in a processive manner – that is, an enzyme binds to a molecule and stays on it for a distance – it is sensible to ask whether methylation enzymes (or methyltransferases) behave similarly. Processivity of one or more of the methyltransferases may explain the non-random methylation pattern across CpG sites on individual molecules.

4.1 Introduction

To understand processivity of methyltransferases, we briefly review the DNA replication process described in Section 1.1.2. During DNA replication, parts of the double-stranded molecule unwind and form replication bubbles. These bubbles keep expanding in both directions, while DNA polymerase binds to the template strands and adds nucleotides to daughter strands from the 5' end of the daughter strand to its 3' end in a processive manner.

As DNA replication proceeds on the rest of the molecule, Dnmt1 and Dnmt3a/b bind to the segments in the replication bubbles that have just completed replica-

tion. At this stage, the daughter strand is completely unmethylated. Hence the CpG dyads on the parent and daughter strands are either hemimethylated or unmethylated. Dnmt1 generally targets hemimethylated CpG dyads and methylates CpGs on the daughter strand (Pradhan et al., 1999). It is therefore usually considered as a "maintenance" methyltransferase. In comparison, Dnmt3a/b mainly bind to unmethylated CpG dyads and may be able to methylate CpGs on both the parent and daughter strand. For this reason, Dnmt3/b are generally thought to be "de novo" methyltransferases (Okano et al., 1998). However, Dnmt1 and Dnmt3a/b have overlapping functions: Dnmt1 may bind to unmethylated CpG dyads (Vilkaitis et al., 2005; Goyal et al., 2006), and Dnmt3a/b hemimethylated CpG dyads (Okano et al., 1998), although both of these events seem to happen with a low probability in somatic cells.

There is further evidence (Leonhardt et al., 1992; Schermelleh et al., 2007) that Dnmt1 may act with the replication protein complex so that it also travels along the molecule processively, moving from the 5' end of the daughter strand to its 3' end (Figure 4.1). However, studies have also demonstrated that this interaction between Dnmt1 and the replication complex is not essential and that Dnmt1 can modify hemimethylated sites efficiently without coupling with the replication complex (Vilkaitis et al., 2005; Schermelleh et al., 2007). The knowledge on the behaviour of Dnmt3a/b is more limited at the moment, but it is suggested that Dnmt3a seems to bind to and methylate CpGs randomly (Gowher and Jeltsch, 2001), whereas Dnmt3b processively (Gowher and Jeltsch, 2002). However, it is unclear whether Dnmt1 and Dnmt3a/b work concurrently or sequentially in a cell cycle. In addition, it is also plausible that an enzyme bound to a CpG may fail to methylate the cytosine, although this seems to occur with a very low probability.

Evidence supporting the processivity of methyltransferases (Vilkaitis et al., 2005; Goyal et al., 2006) is based on *in vitro* experiments and limited analytical methods. Our goal in this chapter is to develop statistical models to assess processivity using

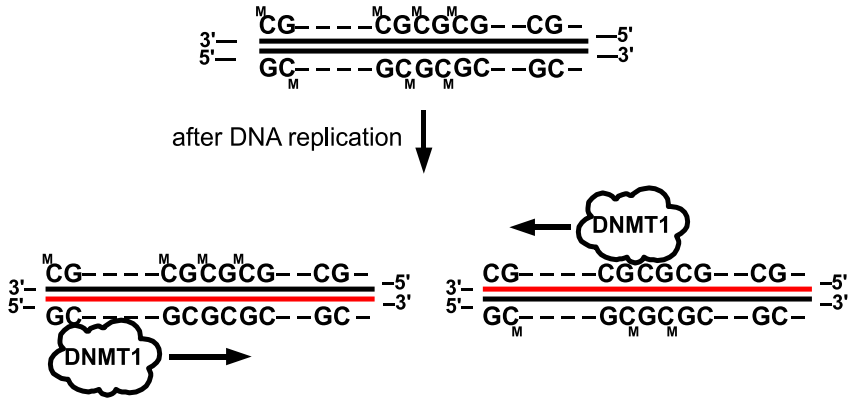

Figure 4.1: Proposed processivity of methyltransferase Dnmt1. Part of the molecule has just unwound and daughter strands (red) synthesised. Hence daughter strands are still unmethylated. As DNA replication proceeds on the rest of the molecule, Dnmt1 binds to the hemimethylated sites and moves along the molecule from the 5' end to the 3' end of the daughter strand processively.

in vivo double-stranded hairpin bisulfite PCR methylation patterns. However, as mentioned above, multiple known or unknown enzymes could have given rise to the in vivo methylation patterns. And even for the few known enzymes, their behaviours are not well understood yet. As a result, in vivo data provide information on methylation *functions*, such as maintenance and de novo, rather than specific *enzymes*. Thus, our models will describe properties of methyltransferases classified by their main function, namely, maintenance and de novo methyltransferases.

Spatial dependence among methylation events along a DNA sequence provides information of processivity of methylation enzymes. This dependence refers to the impact the occurrence of methylation events at a site has on other sites. Exploratory

data analyses below provide evidence for site-site correlation. On the one hand, we observe runs of hemimethylated CpG dyads of the same orientation – methylation occurs at consecutive sites on either the parent or the daughter strand, but not both. Table 4.1 summarises the counts of these runs. These runs can be long as the one of length four in the following sequence:

```
0,0,1,0,1,1,1,1,0,1,0,1,1,0,0,0,0,1,1,1,1,1
1,1,1,0,1,1,1,1,1,1,0,1,1,1,1,1,1,1,1,1,1,1,
```

suggesting that methylation events of the same type may have taken place at CpG sites close to one another. On the other hand, we calculate pairwise correlations in

Table 4.1: Counts of runs of hemimethylated CpG dyads of the same orientation in each *FMR1* data set. Run length is in base pairs (CpGs). Neither data set has runs of length more than 4.

Run Length (CpGs)	Sites 1–22	Sites 25–52
2	11	12
3	1	2
4	3	0

terms of hemimethylated dyads and test the null hypothesis of site independence. We re-coded the data such that hemimethylated sites are coded by 1 and the other two types, methylated and unmethylated sites, are coded by 0. We then used the re-coded data to compute pairwise correlations. Here we look at the results for each data set in turn. The pairwise correlations for the first data set are displayed in the top plot in Figure 4.2. Several blocks right off the diagonal, such as sites 6–9, 12–13 and 18–19, show higher correlations, suggesting possible dependence in a short range. Long range dependence may also exist; for example, between sites 7, 8 and sites 11, 12, and between sites 11,12 and sites 18,19. To test whether correlations at adjacent

sites are indeed higher than due to chance, or in other words, whether there exists local dependence, we carried out a permutation test, using the average correlation at adjacent sites as the test statistic and permuting the sites. The p value of this test is 0.0126. We can therefore reject the independence hypothesis at a significance level of 0.05, although the signal is not very strong. The heat map of pairwise correlations for sites 25–52 (bottom plot in Figure 4.2) shows somewhat higher correlations among the first six sites than among other sites. The p value from the permutation test is 0.03 for this data set. We can still reject the no-dependence hypothesis at the 0.05 level, but the signal across sites 25–52 is even weaker than that in sites 1–22. These signals of dependence, although not strong, motivate the development of statistical models to extract information efficiently from the data.

To quantify the level of dependence in double-stranded methylation patterns, the multi-site models developed in Chapters 2 and 3 are no longer adequate because they assume that methylation events take place independently across CpG sites. As a result, swapping data at sites, say, 8 and 19, would give the same estimates. Instead, we use hidden Markov models (HMMs) here to capture the dependence in the data and to further infer processivity of methyltransferases. We begin with a simple HMM (Section 4.2), and then extend it to incorporate bisulfite conversion errors (Section 4.2.4) and physical distance between CpG sites (Section 4.3). The extended model is applied to the *FMR1* data (Section 4.4). We compare the results with other studies and discuss the models in the remaining of the chapter.

4.2 A simple hidden Markov model (HMM)

4.2.1 Model assumptions

The description of the methylation process and methyltransferases (Section 4.1) contains several aspects, such as the multiple roles of methylation enzymes, the order in which those enzymes operate, etc. These subtleties are important, but not essential

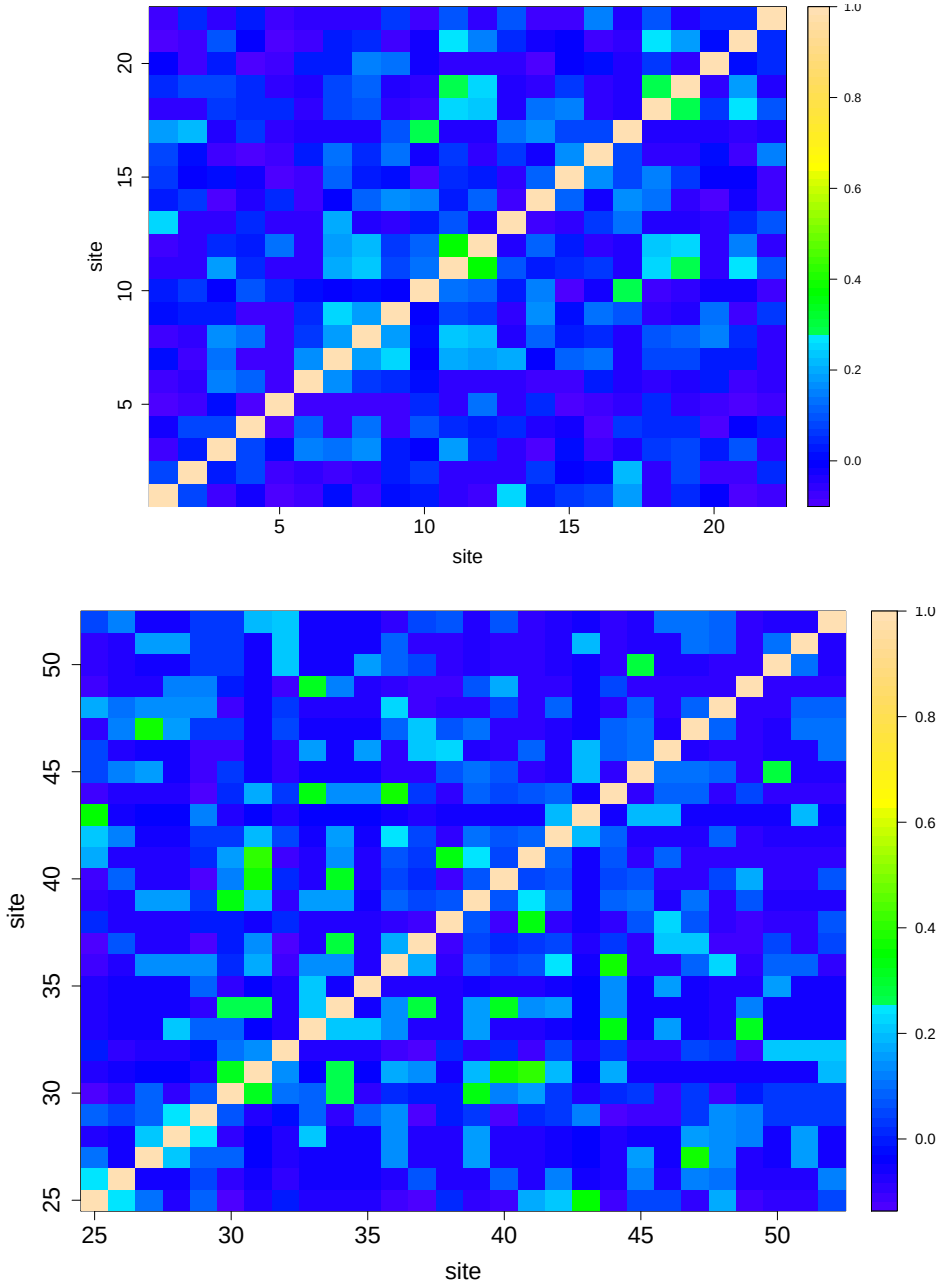

Figure 4.2: Heat maps of pairwise correlations in terms of hemimethylated CpG dyads for the *FMR1* data. Top: sites 1–22; Bottom: sites 25–52.

for modelling processivity at this early stage. Hence,for simplicity we assume that, after the formation of hemimethylated regions, maintenance methyltransferases bind to the molecule and methylate hemimethylated CpGs first, and that de novo methyltransferases come in afterwards to methylate, with some probability, CpGs that are still unmethylated. In addition, we assume, similar to the multi-site models in Chapters 2 and 3, that there is no loss of methylation on the parent strand after DNA replication.

4.2.2 The model

Whereas maintenance methyltransferases produce methylation events on the daughter strand and we can model their behaviour with one stochastic process, de novo methyltransferases operate on both the parent and daughter strand. Our results from Chapters 2 and 3 indicate that de novo events on the two strands could have different rates, so here we allow de novo enzymes to act differently on the two strands, and model parent and daughter de novo processes separately. Altogether we have three hidden Markov processes: maintenance, parent de novo and daughter de novo. Each double-stranded sequence of methylation pattern can be thought of as a mosaic of those three processes.

In statistical terms, we consider three mutually independent processes that give rise to the i-th double-stranded sequence: maintenance, $\mathbf{M}_i$; parent de novo, $\mathbf{R}_i^p$; and daughter de novo $\mathbf{R}_i^d$. Each element of the vectors takes on either 0 or 1, representing whether a methylation event has occurred at a site. For example, $M_{ij} = 1$ indicates that the j-th site is methylated due to a maintenance event underlying the i-th sequence. Assume initially that there is no bisulfite conversion error. We continue to use $(\mathbf{Q}_i, \mathbf{D}_i)$ to denote the post-replication parent and daughter strand. The formation

of the ordered parent-daughter pair $(\mathbf{Q}_i, \mathbf{D}_i)$ is then as follows:

$$Q_{ij} = P_{ij} + (1 - P_{ij})R^P_{ij}; \tag{4.2.1}$$

$$D_{ij} = P_{ij}(M_{ij} + R^D_{ij} - M_{ij}R^D_{ij}) + (1 - P_{ij})R^D_{ij}. \tag{4.2.2}$$

The three processes are independent of the pre-replication parent strand $\mathbf{P}_i$ a priori and the elements of $\mathbf{P}_i$ are independent of one another. Equation (4.2.1) says that the post-replication parent CpG either is methylated before replication or becomes methylated due to a parent de novo event. In equation (4.2.2), the first term on the right hand side says that the daughter CpG maintains the methylation pattern by either a maintenance event or a daughter de novo event – the daughter de novo event may be thought of as some sort of "repair" mechanism in case the maintenance enzyme fails. The second term is straightforward: when the parent CpG is not methylated before methylation, the methylated daughter CpG can be the result only of a daughter de novo event.

The most noticeable difference between the assumptions for the HMM and those for the hierarchical multi-site models (Chapters 2 and 3) and the two existing models (Laird et al., 2004; Genereux et al., 2005) is the role of daughter de novo events. Unlike the HMM, those models do not permit daughter de novo events to repair any failure of maintenance. This possibility can potentially change the inference of methylation events at some sites. To illustrate this point, we consider the following example:

```
1,1,1,1,1,1,1,1,1,
0,0,0,0,1,0,0,0,0.
```

Assuming that the rates of failure of maintenance and de novo methylation events are on the order of a few percent based on the results in Chapter 3, we may infer that the top strand is the parent strand. Whereas previous models may infer the methylated cytosine on the bottom strand in the middle as due to a maintenance event, we infer under the HMM that the above sequence may result from a series of

failure of maintenance events, and hence that a de novo event is more likely to have occurred in the middle on the daughter strand.

We consider each of these processes as a discretised version of a continuous-time jump process with two states, a methylation event occurring or not occurring at that site. The term "time" represents the DNA sequence. If $\{X_j\}$ is such a process with two states, then a characterisation of $\{X_j\}$ would involve, from site $j-1$ to site j, the switch probability θ_j and the probability of switch to state 1, denoted by a_j. That is,

$$a_j = \Pr(X_j = 1 | \text{jump occurs between sites } j-1 \text{ and } j) \tag{4.2.3}$$

The transition probabilities of each process can be expressed as

$$\Pr(X_j = 0 | X_{j-1} = 0) = 1 - \theta_j a_j; \tag{4.2.4}$$

$$\Pr(X_j = 1 | X_{j-1} = 0) = \theta_j a_j; \tag{4.2.5}$$

$$\Pr(X_j = 0 | X_{j-1} = 1) = \theta_j (1 - a_j); \tag{4.2.6}$$

$$\Pr(X_j = 1 | X_{j-1} = 1) = 1 - \theta_j + \theta_j a_j. \tag{4.2.7}$$

Under this parametrisation, θ_j and a_j are independent of one another a priori. The switch probability θ_j measures how much the process deviates from the scenario of all sites being independent. When $\theta_j = 1$, site j is in state 1 with probability a_j regardless of other sites. On the other hand, when $\theta_j = 0$, the state at site j is completely determined by that at site $j-1$.

Here we assume that the processes are spatially stationary along the sequence, meaning that $\theta_j = \theta$ and $a_j = a$ for each process. This also implies that a stationary probability of a site being in state 1, denoted by π, exists and that $\pi = a$. In our case, π is the rate at which a methylation event occurs, and assuming spatial stationarity is equivalent to assuming a constant rate along DNA sequence. Temporal stationarity, which means the methylation density is constant over cell division (time), is not accounted for under this model. To distinguish in notation among the three

processes we add a subscript to these parameters. For example, the parameters for the maintenance process are θ_M and π_M.

With assumptions (4.2.1) and (4.2.2), we can further compute the emission probabilities at site j, which is

$$\begin{aligned}\Pr(Q_{ij}, D_{ij}|M_{ij}, R^P_{ij}, R^D_{ij}, \lambda) = &\ \Pr(Q_{ij}, D_{ij}|P_{ij} = 0, M_{ij}, R^P_{ij}, R^D_{ij}) \Pr(P_{ij} = 0) \\ &+ \Pr(Q_{ij}, D_{ij}|P_{ij} = 1, M_{ij}, R^P_{ij}, R^D_{ij}) \Pr(P_{ij} = 1) \qquad (4.2.8)\\ = &\ \mathbf{1}(Q_{ij} = R^P_{ij}, D_{ij} = R^D_{ij})(1 - m_j) \\ &+ \mathbf{1}(Q_{ij} = 1, D_{ij} = M_{ij} + R^D_{ij} - M_{ij}R^D_{ij})m_j \qquad (4.2.9)\end{aligned}$$

where $\mathbf{1}()$ is the indicator function and m_j is the methylation probability at site j. The probabilities are evaluated in Table 4.2.

Table 4.2: Emission probabilities $\Pr(Q_{ij}, D_{ij}|M_{ij}, R^P_{ij}, R^D_{ij})$ under the hidden Markov model. We allow daughter de novo events to occur at previously methylated CpG sites, thus "repairing" failure of maintenance.

	(Q_{ij}, D_{ij})			
$(M_{ij}, R^P_{ij}, R^D_{ij})$	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
$(0,0,0)$	$1-m_j$	0	m_j	0
$(0,0,1)$	0	$1-m_j$	0	m_j
$(0,1,0)$	0	0	1	0
$(0,1,1)$	0	0	0	1
$(1,0,0)$	$1-m_j$	0	0	m_j
$(1,0,1)$	0	$1-m_j$	0	m_j
$(1,1,0)$	0	0	$1-m_j$	m_j
$(1,1,1)$	0	0	0	1

Let λ denote all the parameters including πs, θs and m_js. The likelihood of

observing the i-th double-stranded sequence given λ is

$$L(\lambda) = \Pr\left(\{\mathbf{Q}_i, \mathbf{D}_i\} = \{\mathbf{x}_i, \mathbf{y}_i\}|\lambda\right) \tag{4.2.10}$$

$$= \Pr\left((\mathbf{Q}_i, \mathbf{D}_i) = (\mathbf{x}_i, \mathbf{y}_i)|\lambda\right) + \mathbf{1}(\mathbf{x}_i \neq \mathbf{y}_i) \Pr\left((\mathbf{Q}_i, \mathbf{D}_i) = (\mathbf{y}_i, \mathbf{x}_i)|\lambda\right) \tag{4.2.11}$$

where

$$\Pr\left((\mathbf{Q}_i, \mathbf{D}_i) = (\mathbf{x}_i, \mathbf{y}_i)|\lambda\right) = \sum_{\mathbf{M}_i, \mathbf{R}_i^p, \mathbf{R}_i^d} \Pr\left((\mathbf{Q}_i, \mathbf{D}_i) = (\mathbf{x}_i, \mathbf{y}_i)|\mathbf{M}_i, \mathbf{R}_i^p, \mathbf{R}_i^d, \lambda\right) \times \Pr(\mathbf{M}_i, \mathbf{R}_i^p, \mathbf{R}_i^d|\lambda). \tag{4.2.12}$$

We use the standard forward-backward algorithm (Rabiner and Juang, 1986), described in the next section, to compute (4.2.12). Probability of the top strand being the daughter strand and the bottom the parent, $\Pr((\mathbf{Q}_i, \mathbf{D}_i) = (\mathbf{y}_i, \mathbf{x}_i)|\lambda)$, is calculated in a similar way.

4.2.3 The forward-backward algorithm to compute $\Pr\left((\mathbf{Q}_i, \mathbf{D}_i) = (\mathbf{x}_i, \mathbf{y}_i)|\lambda\right)$

In the i-th double-stranded sequence of length S, let $H_{ij} = (M_{ij}, R_{ij}^P, R_{ij}^D)$, the composite hidden variable of dimension 8 at site j, and $Z_{ij} = (x_{ij}, y_{ij})$, the ordered pair of methylation states of dimension 4. Define the transition probability to be $t_{h,h'} = \Pr(H_{i,j+1} = h'|H_{ij} = h)$ and the emission probability to be $b_h(Z_{ij}) = \Pr(Z_{ij}|H_{ij} = h)$. The stationary probabilities are defined as $\tilde{\pi}_h$, which is the product of the stationary probability on each process. Applying the forward-backward algorithm in Rabiner and Juang (1986), we have,

$$\text{forward pass:} \quad \alpha_j(h) = \Pr(\mathbf{Q}_{i,1:j} = \mathbf{x}_{i,1:j}, \mathbf{D}_{i,1:j} = \mathbf{y}_{i,1:j}, H_{ij} = h); \tag{4.2.13}$$

$$\text{backward pass:} \quad \beta_j(h) = \Pr(\mathbf{Q}_{i,j+1:S} = \mathbf{x}_{i,j+1:S}, \mathbf{D}_{i,j+1:S} = \mathbf{y}_{i,j+1:S}|H_{ij} = h). \tag{4.2.14}$$

The forward-backward algorithm calculates the forward and backward pass in each

iteration:

$$\alpha_1(h) = \tilde{\pi}_h b_h(Z_{i1}), \tag{4.2.15}$$

$$\alpha_j(h) = \sum_{h'=1}^{8} \alpha_{j-1}(h') t_{h',h} b_h(Z_{ij}), \quad \text{for } j > 1; \tag{4.2.16}$$

and

$$\beta_S(h) = 1, \tag{4.2.17}$$

$$\beta_j(h) = \sum_{h'=1}^{8} t_{h,h'} b_{h'}(Z_{i,j+1}) \beta_{j+1}(h'), \quad \text{for } j < S. \tag{4.2.18}$$

The probability of the top strand being the parent strand and the bottom the daughter strand in the i-th sequence is then

$$\Pr\left((\mathbf{Q}_i, \mathbf{D}_i) = (\mathbf{x}_i, \mathbf{y}_i)|\lambda\right) = \sum_{h=1}^{8} \alpha_S(h) = \sum_{h=1}^{8} \alpha_j(h)\beta_j(h), \quad \text{for every } j. \tag{4.2.19}$$

4.2.4 Accounting for bisulfite conversion error

Recall from Section 1.2 that there are two types of bisulfite conversion error: the failure of conversion error with rate b, and the inappropriate conversion error with rate c. Let (Q_{ij}, D_{ij}) represent the *true* states without errors on post-replication parent and daughter strand and (Q'_{ij}, D'_{ij}) the *observed* states due to inclination of errors. Then,

$$b = \Pr(Q'_{ij} = 1|Q_{ij} = 0) = \Pr(D'_{ij} = 1|D_{ij} = 0); \tag{4.2.20}$$

$$c = \Pr(Q'_{ij} = 0|Q_{ij} = 1) = \Pr(D'_{ij} = 0|D_{ij} = 1). \tag{4.2.21}$$

Just as the existence of error can add confounding to the rate estimation in Chapter 3, its existence can also reduce the level of dependence by breaking down runs of hemimethylated dyads, or produce false signals of processivity by creating longer runs. Therefore, we want to incorporate these errors into the hidden Markov model.

Let λ' denote all the parameters including πs, θs, m_js, b and c. The definition and computation of the likelihood function are similar to equations (4.2.10) through

(4.2.12), except that Q_{ij}, D_{ij} and λ are replaced by Q'_{ij}, D'_{ij} and λ', respectively. We also need to change the emission probabilities as the following to incorporate b and c into the HMM:

$$\begin{aligned}\Pr(Q'_{ij}, D'_{ij}|M_{ij}, R^P_{ij}, R^D_{ij}, m_j, b, c) &= \sum_{Q_{ij},D_{ij}} \Pr(Q'_{ij}, D'_{ij}|Q_{ij}, D_{ij}, b, c) \\ &\quad \times \Pr(Q_{ij}, D_{ij}|M_{ij}, R^P_{ij}, R^D_{ij}, m_j, b, c) \qquad (4.2.22) \\ &= \sum_{Q_{ij}} \Pr(Q'_{ij}|Q_{ij}, b, c) \sum_{D_{ij}} \Pr(D'_{ij}|D_{ij}, b, c) \\ &\quad \times \Pr(Q_{ij}, D_{ij}|M_{ij}, R^P_{ij}, R^D_{ij}, m_j) \qquad (4.2.23)\end{aligned}$$

and $\Pr(Q_{ij}, D_{ij}|M_{ij}, R^P_{ij}, R^D_{ij}, m_j)$ is the emission probability under the no-error model and is listed in Table 4.2.

4.3 Incorporating physical distance into the simple HMM

4.3.1 HMM with distance

Methylation enzymes need to stay bound to the DNA molecule in order to methylate cytosines. It is sensible to assume that the binding ability, and hence the level of processivity, decreases over molecular distance. Therefore, the farther apart the two neighbouring CpG sites are, the less dependence between them. The previous hidden Markov model is not satisfactory in this respect because it assumes equal distances between adjacent sites. We incorporate physical distance into the hidden Markov model by expressing the switch probability θ as a function of the pairwise distance.

In the previous section, we modelled each hidden Markov chain as a discretised version of a homogeneous continuous-time jump process with switch probability θ_j and the probability of switching to state 1, a_j. This continuous-time jump process is also a Poisson process with a constant jump rate ρ. Since each continuous-time jump process here has two states, the jump rate ρ on a process represents the rate at which this process jumps from one state to the other during a very short time. The relation

between the switch probability θ_j between sites $j-1$ and j, and the jump rate ρ is as the following:

$$\theta_j = \Pr(\text{jump occurs between sites } j-1 \text{ and } j) \tag{4.3.1}$$

$$= 1 - \Pr(\text{no jump occur between sites } j-1 \text{ and } j) \tag{4.3.2}$$

$$= 1 - e^{-\rho d_j}, \tag{4.3.3}$$

where d_j is the physical distance in nucleotides (nt) between sites $j-1$ and j. Thus, the incorporation of physical distance leads to varying switch probabilities along the DNA sequence, even though the jump rate per nucleotide is constant. Incidentally, the hidden Markov model in the previous section can be treated as a special case of the model here with equal distances.

Under this generalised HMM, we continue to assume the rate of any type of methylation event, π, to be constant across CpG sites and do not account for temporal stationarity. Hence, this model can also be thought of as a generalisation of the simple multi-site model in Chapter 2, with spatial dependence as well as error added in.

4.3.2 Sojourn times as a measure of processivity

The concept of sojourn time for a stochastic process in time describes how long the process stays in a certain state. For each of the underlying continuous-time jump process, we can write the infinitesimal matrix as in Table 4.3 where π is the stationary

Table 4.3: The infinitesimal matrix of the continuous-time jump process.

	0	1
0	$-\pi\rho$	$\pi\rho$
1	$(1-\pi)\rho$	$-(1-\pi)\rho$

probability of being in state 1, and ρ the jump rate. Thus, the sojourn times in

state 0 are independent and identically distributed as exponential with mean $1/(\pi\rho)$, whereas those in state 1 are independent and identically distributed as exponential with mean $1/((1-\pi)\rho)$ (Taylor and Karlin, 1998). These average sojourn times measure processivity in terms of nucleotides. For example, the maintenance events of Dnmt1 would stretch for as long as $1/((1-\pi_M)\rho_M)$ nucleotides on average, whereas the failure of maintenance events of this enzyme would stretch for $1/(\pi_M\rho_M)$ nucleotides on average. The two average sojourn times give a complete characterisation of a process. We use both of them to measure processivity.

4.3.3 Distribution assumptions and choice of priors

Let $\boldsymbol{\rho} = \{\rho_M, \rho_{RR}, \rho_{RD}\}$, $\boldsymbol{\pi} = \{\pi_M, \pi_{RP}, \pi_{RD}\}$. The parameter of interest is $\lambda'' = \{\boldsymbol{\pi}\ , \boldsymbol{\rho}\ , m_j, b, c, j = 1, \cdots, S\}$. Since temporal stationarity is not accounted for in this HMM, we assume methylation probabilities m_j to follow a single beta distribution with mean r_m and scaled variance g_m. The priors on r_m and $\log_{10}(g_m)$ are, as before, Unif$(0,1)$ and Unif$(-4,0)$, respectively. We also use a Unif$(0,1)$ prior for each of the π's. As for the error rates b and c, when analysing the *FMR1* data we fix b to be 0.3%, assume that c is constant across sites, and use a Unif (0,0.06) prior on c.

We consider two priors on the jump rate ρ: one is exponential with mean $1/\bar{d}$, where $\bar{d}$ is the average distance; and the other is exponential with mean 1. The first prior assumes one jump every $\bar{d}$ nucleotides. When all distances are equal, this prior induces a uniform (0,1) distribution on the switch probability θ, thus corresponding to making no assumption on the level of processivity. An exp(1) prior, on the other hand, assumes one jump per nucleotide on average, which is equivalent to very weak processivity. In terms of the *FMR1* data, the distances are not equal and $\bar{d}$ is about 6.7 nucleotides. The 10th- to 90th-percentile interval of Exp(mean=$1/\bar{d}$) is (0.016, 0.345), which is much more concentrated near 0 compared with that of Exp(1), which is (0.105, 2.303). Hence the Exp(mean=$1/\bar{d}$) prior implies somewhat stronger processivity.

To see why the Exp(mean=$1/d$) prior on the jump rate ρ induces a Unif$(0,1)$ on the switch probability θ when all distances between adjacent sites are the same, denoted by d, we consider Equation (4.3.3), which is the cumulative distribution function of an exponential random variable with mean $1/d$ when $d_j = d$ for all j. Thus θ as a cumulative probability would always be uniform on $(0,1)$.

4.3.4 The MCMC procedure

We use the full conditional distribution to update in turn the parameters, some of which grouped in blocks, via Metropolis-Hastings (MH) steps (Liu, 2001). We do not update all the parameters λ'' simultaneously, because the data have different amounts of information for different parameters and it would be difficult to fine-tune the standard deviation used in the normal density kernel to generate proposals for a single parameter. We update the parameters in the following order:

1. Update each of methylation event rates π and jump rates ρ.

2. Update methylation probabilities m_j jointly.

3. Update r_m, the mean of m_j in the beta distribution.

4. Update $\log_{10}(g_m)$, where g_m is the scaled variance of m_j.

5. Update the inappropriate conversion error rate c. If the failure of conversion error rate b were to be estimated as well, the MH step would be similar to that for c.

To update, say, the maintenance rate π_M, we generate the proposal π_M^* from a normal distribution with the current value π' as the mean, and standard deviation σ_M. We then compute log likelihoods $\log L(\pi_M^*, \cdot)$ and $\log L(\pi'_M, \cdot)$, where the dot $\cdot$ represents current values of other parameters. The proposal is accepted with probability

$\min(1, A)$, where

$$\log A = \log L(\pi_M^*, \cdot) - \log L(\pi_M', \cdot). \tag{4.3.4}$$

Since methylation event rate π and jump rate ρ take on values between 0 and 1, any proposal outside this range should be discarded and the current value retained. When π or ρ is close to 0 and 1, or when the standard deviation used in generating the proposal is large – due to uncertainty in the parameter – the MCMC procedure has to go through many iterations before a suitable proposal is generated, and perhaps even more iterations for a proposal to be accepted. To improve efficiency of the MCMC procedure while retaining the correct sampling distribution, we use a reflection strategy to generate proposals by reflecting proposals outside $(0, 1)$ back in the range. The rule can be summarised as finding the distance between the proposal and the nearest even number and using this distance as the "new" proposal. Suppose we are updating one of the methylation event rates π, whose current value is π'. The reflection strategy takes the following steps:

1. Generate the proposal $\check{\pi}$ from Normal(π', σ^2);

2. Find g, the integer part of $|\check{\pi}|$;

3. If g is an odd number, then the "reflected" proposal is

$$\pi^* = g + 1 - |\check{\pi}|. \tag{4.3.5}$$

   Otherwise, it is

$$\pi^* = |\check{\pi}| - g. \tag{4.3.6}$$

4.4 Analysis of the *FMR1* data

The physical distance in nucleotides between two CpG sites is measured by the number of nucleotides between the two cytosines. The distances for sites 1-22 are:

$$13, 2, 10, 11, 16, 16, 2, 4, 4, 5, 2, 2, 11, 5, 8, 4, 13, 2, 6, 2, 2$$

with mean 6.667 and median 5. The distances for sites 25-52 are:

$$2, 16, 9, 5, 21, 5, 5, 8, 14, 10, 5, 12, 11, 10, 4, 3, 5, 3, 3, 4, 2, 4, 3, 3, 6, 6, 3$$

with mean 6.741 and median 5. In Section 4.4.1, we applied the hidden Markov model with distance to the two *FMR1* data sets separately. Since the two data sets are from two regions separated only by a few base pairs and the results from analysing the two regions separately yielded similar results on processivity, we further carried out a combined analysis (Section 4.4.2), assuming that the two regions share the same jump rates ρ, rates of methylation events π, inappropriate conversion error rate c, mean r_m and scaled variance g_m of methylation probabilities m (see Section 3.1.1 for the $r - g$ parametrisation of a beta distribution). Meanwhile, we continue to assume that methylation probabilities can vary across CpG sites. In the combined analysis the likelihood of all the data is just the product of the likelihood of each data set. The MCMC procedure is similar to that in separate analyses.

4.4.1 Analysing two FMR1 data sets separately

Inference about processivity

The posterior distribution of jump rate ρ_M (Figures 4.3 and 4.4) under the prior Exp(mean=$1/\bar{d}$), where $\bar{d}$ is the average distance, changes significantly in both data sets, suggesting evidence for processivity on the maintenance process. Posterior distributions of ρ_{RP} and ρ_{RD}, however, show barely any change from the prior.

To quantify the level of processivity in the maintenance process, we use the MCMC samples of the stationary probability π_M (also the average maintenance rate), and the jump rate ρ_M to compute estimates of average sojourn times $1/(\pi_M \rho_M)$ for maintenance events and $1/((1-\pi_M)\rho_M)$ for failure of maintenance events. The 80% credible intervals of these estimates are listed in Table 4.4. The intervals of the average sojourn times, or stretch of processivity, for failure of maintenance events are narrow and almost do not change under different priors for the two data sets. In contrast,

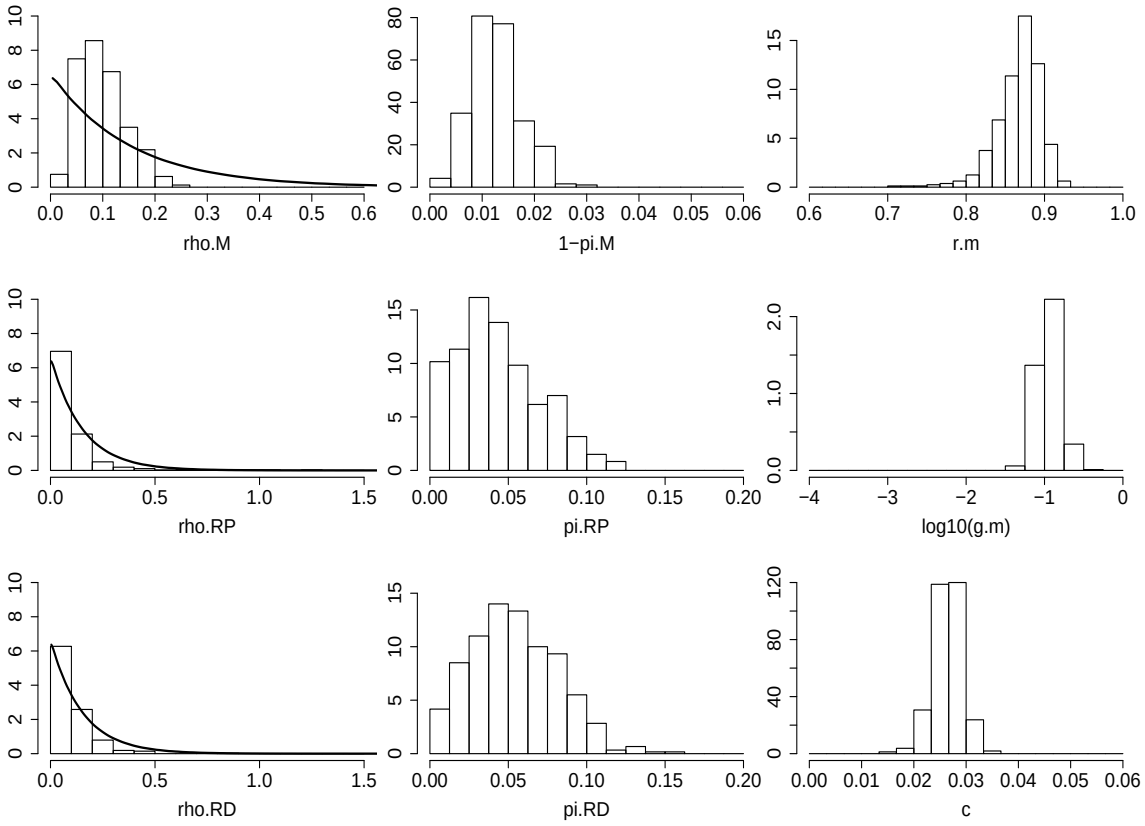

Figure 4.3: Posterior distributions of parameters for the *FMR1* data at sites 1–22 under the HMM with physical distances. Dark curves in histograms of ρs are the density of the exponential prior with mean $1/\bar{d}$, where $\bar{d} \approx 6.7$nt is the average physical distance.

the intervals for maintenance events are much wider and seem to suggest somewhat longer stretches on average at sites 1–22 than in at sites 25–52. 80% credible intervals in Figure 4.5 further visualise the difference.

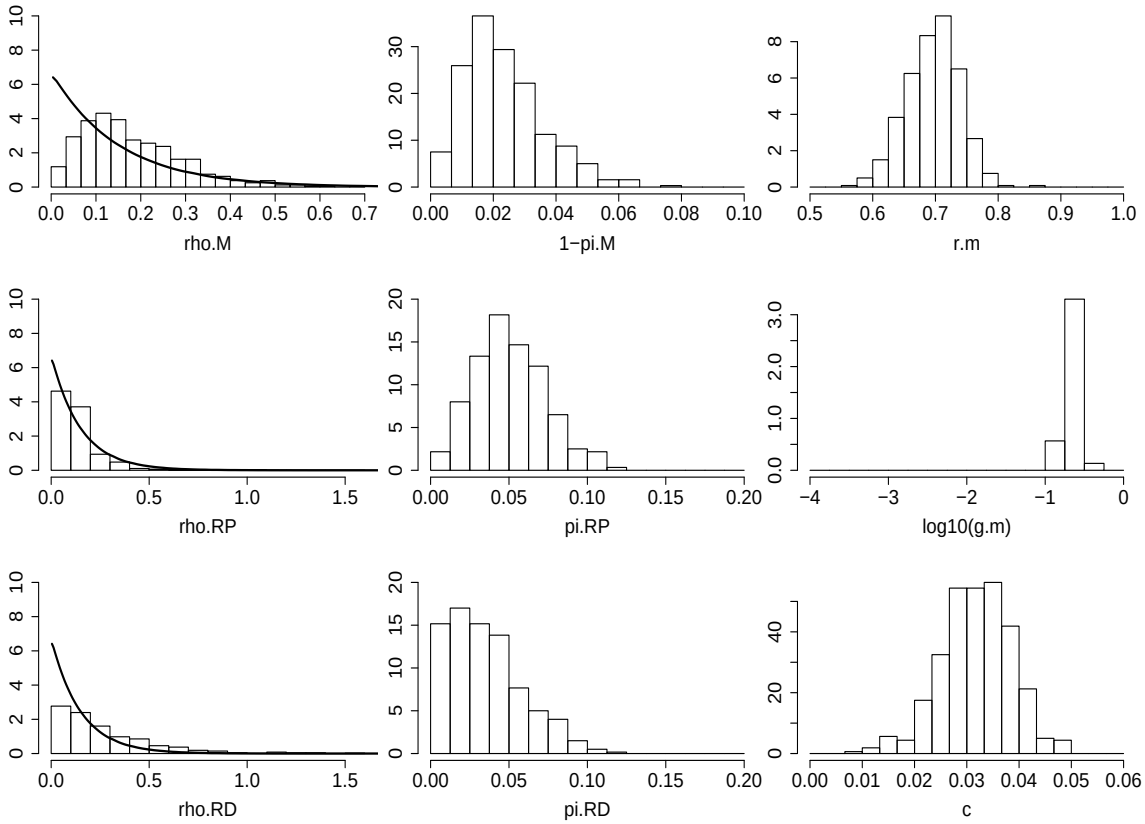

Figure 4.4: Posterior distributions of parameters for the *FMR1* data at sites 25–52 under the HMM with physical distances. Dark curves in histograms of ρs are the density of the exponential prior with mean $1/\bar{d}$, where $\bar{d} \approx 6.7$nt is the average physical distance.

Table 4.4: 80% credible intervals of processivity in nucleotides, represented by the average sojourn times in a continuous-time Markov chain, for the maintenance process, analysing two *FMR1* data sets separately and jointly. Intervals from the joint analysis are in red. For stretch of maintenance events, the average sojourn time is computed as $1/(\pi_M \rho_M)$ using estimates of average maintenance rate π_M and jump rate ρ_M. For stretch of failure of maintenance events, the average sojourn time is $1/((1-\pi_M)\rho_M)$. Two priors were used on jump rate ρ. One is exponential with mean $1/\bar{d}$, where $\bar{d} \approx 6.7$nt is the average physical distance; this prior corresponds to some processivity. The other prior is Exp(1), corresponding to much weaker processivity.

	Exp(mean=$1/\bar{d}$)		Exp(1)	
Event	Sites 1–22	Sites 25–52	Sites 1–22	Sites 25–52
Maintenance	$(413, 2204)$	$(301, 2067)$	$(77, 1413)$	$(34, 946)$
	$(347, 1716)$		$(304, 1514)$	
Failure of maintenance	$(6, 20)$	$(5, 19)$	$(3, 18)$	$(2, 14)$
	$(6, 17)$		$(5, 16)$	

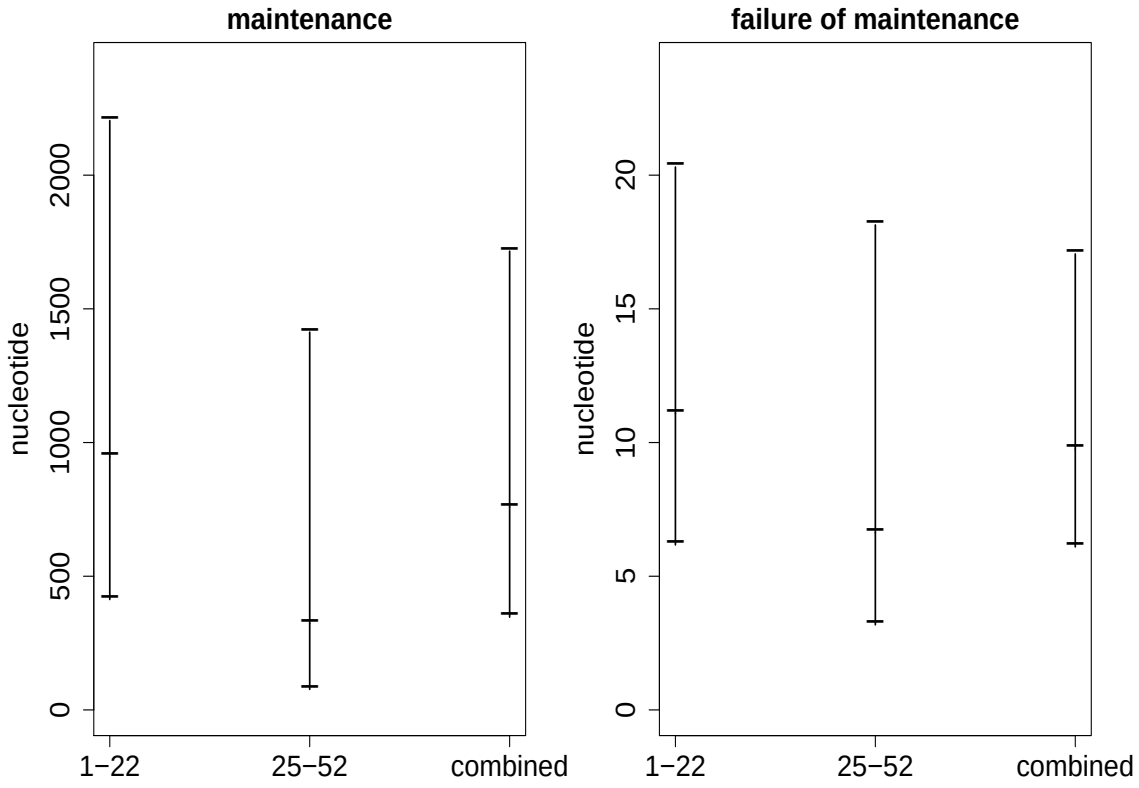

Figure 4.5: Posterior medians and 80% credible intervals of average sojourn times (in nucleotides) for maintenance and failure of maintenance events of maintenance methyltransferases for the *FMR1* data. Each plot compares separate analysis of each data set with the joint analysis.

Inference about the rates

We summarise the estimates of the rates in terms of their 80% credible intervals in Table 4.5. The stationary probabilities π are rates of three types of methylation events. Not surprisingly, their 80% credible intervals are robust to the choice of priors on jump rates. The average failure of maintenance rates are a little lower than under the multi-site models, whereas the average inappropriate bisulfite conversion error rates are a little higher than under the multi-site models (Tables 3.6 and 3.12). The differences might be due to the lack of temporal stationarity in the hidden Markov model. Furthermore, we cannot rule out the possibility that no de novo events occur on the parent strand at sites 1–22, or on the daughter strand at sites 25–52. This result is consistent with that under the multi-site mixture model (Table 3.12).

4.4.2 Analysing two FMR1 data sets jointly

Since our inference for processivity from separate analyses yielded similar levels of processivity in the two regions, we carried out a joint analysis, writing the likelihood of the combined data set as the product of the likelihood of each data set. We plot the posterior distributions of the parameters under the prior Exp(mean=$1/\bar{d}$) in Figure 4.6. Again the posterior distribution of the jump rate ρ_M changes significantly from the prior, whereas this is not the case for ρ_{RP} and ρ_{RD}. The 80% credible intervals of stretch of processivity and those of the parameter estimates are listed in red in Table 4.4 and 4.5. Overall, the joint analysis produces sharper interval estimates, which lie in between those from the separate analysis on either data set.

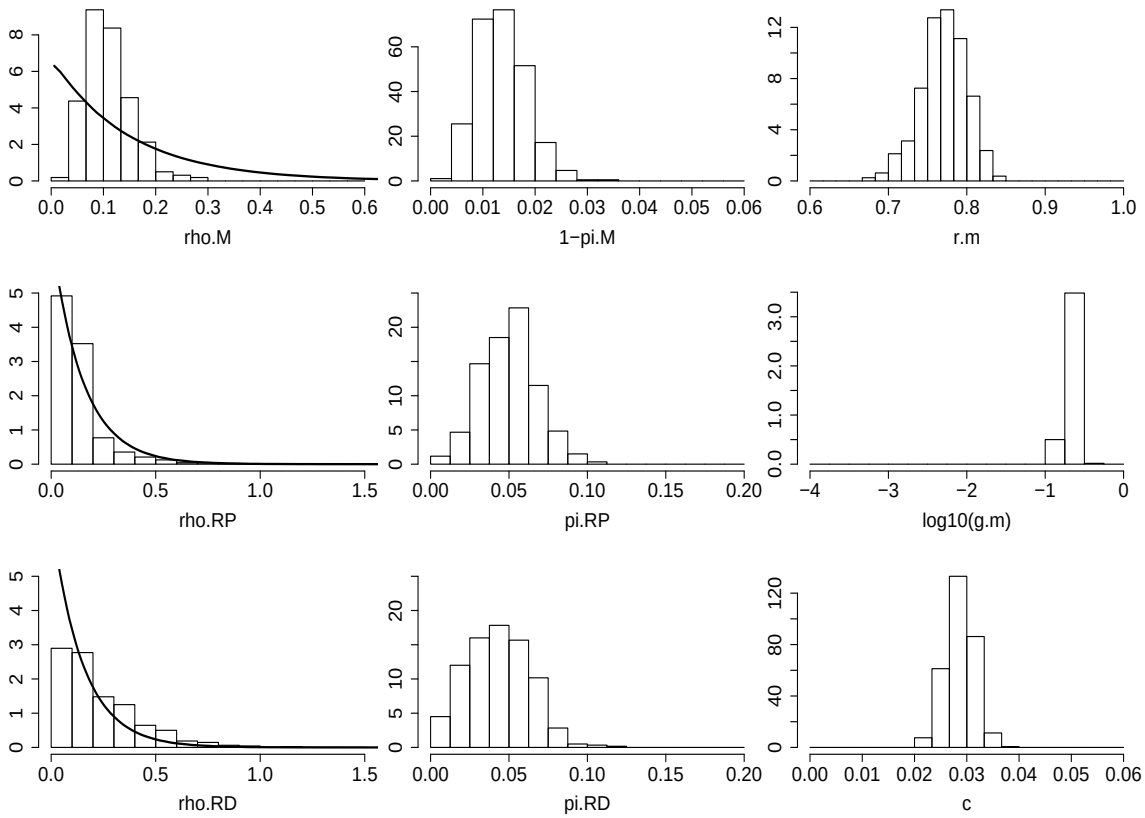

Figure 4.6: Histograms of parameters under the HMM with physical distances, analysing the two *FMR1* data sets jointly. Dark curves in histograms of ρs are the density of the exponential prior with mean $1/\bar{d}$, where $\bar{d} \approx 6.7$nt is the average physical distance.

Table 4.5: 80% credible intervals of parameter estimates under the hidden Markov model with physical distance from the *FMR1* data under the separate and joint analyses. Intervals from the joint analysis are in red. Two priors were used on jump rate ρ. One is exponential with mean $1/\bar{d}$, where $\bar{d} \approx 6.7$nt is the average physical distance; this prior corresponds to some processivity. The other prior is Exp(1), corresponding to much weaker processivity. The 10th- and 90th-percentile of each exponential distribution are listed in boldface.

	Exp(mean=$1/\bar{d}$)		Exp(1)	
	Sites 1–22	Sites 25–52	Sites 1–22	Sites 25–52
Prior	$(\mathbf{0.016}, \mathbf{0.345})$		$(\mathbf{0.105}, \mathbf{2.303})$	
ρ_M	$(0.050, 0.165)$	$(0.057, 0.329)$	$(0.053, 0.195)$	$(0.075, 0.676)$
	$(0.059, 0.167)$		$(0.065, 0.189)$	
ρ_{RP}	$(0.019, 0.195)$	$(0.050, 0.254)$	$(0.028, 1.585)$	$(0.065, 1.798)$
	$(0.049, 0.275)$		$(0.047, 1.341)$	
ρ_{RD}	$(0.023, 0.212)$	$(0.045, 0.580)$	$(0.035, 0.820)$	$(0.116, 2.561)$
	$(0.061, 0.484)$		$(0.083, 1.999)$	
$1 - \pi_M$	$(0.007, 0.019)$	$(0.001, 0.042)$	$(0.007, 0.021)$	$(0.011, 0.048)$
	$(0.008, 0.020)$		$(0.009, 0.021)$	
π_{RP}	$(0.011, 0.083)$	$(0.023, 0.082)$	$(0.007, 0.080)$	$(0.022, 0.083)$
	$(0.027, 0.074)$		$(0.027, 0.076)$	
π_{RD}	$(0.019, 0.092)$	$(0.008, 0.070)$	$(0.017, 0.092)$	$(0.015, 0.078)$
	$(0.019, 0.068)$		$(0.024, 0.072)$	
$(\pi_{RP} + \pi_{RD})/2$	$(0.029, 0.069)$	$(0.029, 0.057)$	$(0.026, 0.068)$	$(0.033, 0.063)$
	$(0.035, 0.059)$		$(0.038, 0.062)$	
r_m	$(0.831, 0.898)$	$(0.638, 0.748)$	$(0.835, 0.898)$	$(0.635, 0.740)$
	$(0.733, 0.805)$		$(0.731, 0.801)$	
c	$(0.023, 0.030)$	$(0.023, 0.040)$	$(0.022, 0.030)$	$(0.015, 0.035)$
	$(0.025, 0.032)$		$(0.024, 0.031)$	

4.5 Summary and discussion

Our analyses of the in vivo *FMR1* data provide strong evidence for processivity in maintenance enzymes, but little evidence for processivity in de novo enzymes. This conclusion about processivity of maintenance methyltransferases is consistent with that from in vitro experiments in literature, although those experiments (see for example Vilkaitis et al., 2005; Goyal et al., 2006) were carried out under quite different conditions from that for the *FMR1* experiment and were aimed specifically at the enzyme Dnmt1. We compare our analysis with these published studies in Table 4.6.

The *FMR1* experiments are unique compared with other studies in that the *FMR1* experiments are in vivo and therefore involve not only maintenance enzymes, but also other types of methyltransferases. On the one hand, analysing multiple enzymes simultaneously presents challenges for statistical modelling. On the other, the experiments provide us with an opportunity to draw conclusions regarding in vivo properties of methyltransferases.

By comparison, the other two published studies (Vilkaitis et al., 2005; Goyal et al., 2006), in simple terms, synthesised hemimethylated and unmethylated substrates (sequences) and subjected them to purified Dnmt1 in vitro. An advantage of these experiment-based analyses is their ability to study a specific enzyme. Nevertheless, since DNA replication is not involved in these experiments, the way in which Dnmt1 binds to molecules and modifies cytosines in these in vitro experiments might differ from the way in which Dnmt1 and other methylation enzymes gave rise to the in vivo *FMR1* data. For example, the hypothesis that the maintenance process couples with the replication complex does not apply in the in vitro experiments. Perhaps due to this difference, methylation densities in the DNA products from those two studies are significantly lower than that in the *FMR1* data, and runs of hemimethylated CpG dyads can also be much longer than those in the *FMR1* data. These differences could have led to different quantitative characterisations of processivity as we see in Table

Table 4.6: Comparison of studies on processivity in methyltransferases. Our analysis of the *FMR1* data is currently the only in vivo study to our knowledge. Vilkaitis et al. (2005) and Goyal et al. (2006) are two representative in vitro studies.

	In Vivo	In Vitro	
Origin of data	*FMR1* experiments (Miner and Stöger, unpublished data)	Vilkaitis et al. (2005)	Goyal et al. (2006)
Enzymes studied	maintenance and de novo	Dnmt1	Dnmt1
DNA source	human in vivo	mouse in vitro	mouse in vitro
Length of sequence	350bp	520bp	634bp; 566bp
No. of CpG sites	50	30	54
Analysis method	hidden Markov models		random walk for maintenance activity
Results on maintenance activity	stretch of maintenance: 347–1716 nt (52–256 CpG sites)	"Dnmt1 remains associated with one strand"..."in stretches as long as 520 bp [30 CpG sites]".	estimated diffusion length: 6000 nt (566 CpG sites)[1]
Results on failure of maintenance activity	stretch of failure of maintenance: 6–17 nt (1–3 CpG sites)	not determined	not determined

[a]Estimate is based on the average number of CpG sites in one binding event of Dnmt1 in the data, which is 14–16 (Goyal et al., 2006).

4.6.

The three studies under comparisons also take very different approaches to quantifying processivity. Vilkaitis et al. (2005) did not have a statistical model for processivity. Instead, they simply stated, as quoted in Table 4.6, that results from their

experiments indicate that Dnmt1 can be associated with DNA molecules in length as long as 520 bp, which is also the length of substrates used in the experiments. Goyal et al. (2006), on the other hand, employed a random walk model for Dnmt1 and estimated the diffusion length to be around 6000 bp for processivity. In comparison, our results suggest that maintenance methyltransferases, whether they are Dnmt1 or other enzymes, can processively methylate hemimethylated sites for 347–1716 nt on average, although our estimates seem to be smaller than that from Vilkaitis et al. (2005).

As mentioned in Section 4.2.1 the HMMs developed in this chapter are based on a simplified description of the methyltransferases and the methylation process. It will be worthwhile in our future work to relax those assumptions to make the models more realistic. Two assumptions are particularly worth reconsidering. One relates to the de novo methylation function of maintenance methyltransferases, Dnmt1 in particular (Vilkaitis et al., 2005; Goyal et al., 2006). We can allow for this function by introducing a de novo rate for maintenance methyltransferases into the model. This extension will enable us to infer the preference ratio of hemimethylated versus unmethylated CpGs for maintenance methyltransferases; this ratio has been measured under different experimental conditions (see, for example, Okano et al., 1998; Vilkaitis et al., 2005; Goyal et al., 2006) and used to determine whether an enzyme is a maintenance or de novo methyltransferase (Okano et al., 1998). Nonetheless, different experimental conditions have led to a wide range of the estimated ratio, such as 2–50 as reported in Goyal et al. (2006). The other future direction is to allow for spontaneous failure of methylation in maintenance methyltransferases: the enzymes may bind to a CpG but fail to methylate it regardless of the methylation state of the parent strand. Hence the extension allows maintenance methyltransferases to act imperfectly, and hence is more biologically realistic. We are currently working on this relaxed model to see what more insight it could provide for the study of processivity.

Another possible extension of the HMMs introduced in this chapter is to allow

for variability in the rates of methylation events (they are assumed to be constant across CpG sites now), by letting the underlying continuous-time Markov processes be heterogeneous. However, we think that this direction may not be necessary or even desirable: not only our main interest here is the dependence rather than the rates of methylation events, but also the suggested extension increases dimensionality of the parameter space to nearly four times of the current dimension, and is even more demanding for the limited data currently available.

Chapter 5

CONCLUSIONS AND DISCUSSION

5.1 Conclusions and discussion

DNA methylation as an epigenetic mechanism has drawn a considerable amount of attention in recent years; we are gaining more understanding and appreciation of its role in gene regulation, which has a large impact on normal development in organisms and can lead to diseases, cancer in particular (see, for example, Feinberg and Vogelstein, 1987; Jones and Baylin, 2002; Laird, 2003; Chen and Riggs, 2005).

In this thesis we have focused on the transmission process of methylation patterns over cell division in mammals and investigated the following two issues: (1) inference for rates of methylation events, which occur during the transmission process, and for the variation in those rates across CpG sites. Specifically, we have considered failure of maintenance, parent de novo and daughter de novo events; and (2) the possibility of processivity in methylation enzymes, namely, maintenance and de novo methyltransferases. Processivity leads to dependence across sites in the observed methylation patterns.

Different types of methylation data are available for studying the transmission process. Here we analysed double-stranded sequence methylation data collected from the *FMR1* locus (data courtesy of Brooks Miner and Reinhard Stöger) using the hairpin-bisulfite PCR technique (Laird et al., 2004). These double-stranded data provide information on both parent and daughter strands, thus enabling us to study more directly methylation events, which involve two generations. However, current technologies cannot determine which strand in each sequence is the parent strand, and which is the daughter strand.

To estimate rates of methylation events and to assess variation in those rates across CpG sites, we have formulated the question as a latent variable problem and developed multi-site models to exploit the multi-site information. The latent variables are: (1) methylation states on the pre-replication parent strand, **P**; and (2) the strand type, i.e. which of the two strands in a sequence is the post-replication parent strand **Q** and which is the daughter strand **D**. Using the methylation density and multi-site information in the data, the multi-site models are capable of distinguishing among different types of methylation events, as demonstrated in the simulation studies in Chapter 2, although cases also exist where those events are unidentifiable.

We have further imposed a hierarchical structure to specifically incorporate variability in a rate as a parameter into the model. This allows us to estimate this quantity from the data, rather than relying on variability in the estimates; variability in the estimates can be due to sampling or the estimation method. Hierarchical modelling also allows us to borrow information from other sites, which helps reduce dimensionality of the parameter space. We have adopted a Bayesian framework and used Markov chain Monte Carlo methods for inference and to assess the strength of evidence, which is quite desirable especially when high-throughput double-stranded data are still difficult to get.

In addition to these standard techniques, we have added several features to the models so that they are more suitable for this particular set of scientific questions. A prominent feature is the ease with which they incorporate and estimate rates of bisulfite conversion errors, a type of experimental error. Those errors have a significant impact on the inference for the failure of maintenance and de novo rates. By contrast, most existing methods ignore errors in their analyses of methylation patterns. The second feature is to allow for departure from temporal stationarity and to assess the level of departure. Temporal stationarity assumes that methylation densities are stable over cell division; it is an assumption underlying many existing methods. Allowing for deviation from this assumption makes the models more flexible and

applicable to genomic regions where stationarity does not hold. We applied our models to the *FMR1* data at sites 25–52, of which several sites seem to have departed from stationarity (Figure 3.12). The third feature is their ability to identify outlier CpG sites that may have much higher de novo rates than others do.

Applying the multi-site models to the *FMR1* data, we conclude that,

1. The average inappropriate bisulfite conversion rate is 1–3% with little variation across CpG sites. This result is consistent under multi-site models with different assumptions. It is also consistent with estimates obtained using laboratory approaches (Genereux et al., 2008). Ignoring this error could lead to significant overestimation of the failure of maintenance rate and bias in estimates of the de novo rates.

2. Under the mixture model, the average failure of maintenance rate is 2–4% at sites 1–22 and 3–6% at sites 25–52, with little variation in either region. These estimates do not change much under other multi-site models.

3. The mixture model suggests that, de novo events may have occurred at a high rate at sites 10, 14, 15 and 16 on the parent strand. De novo events have also occurred at some of the other sites among sites 1–22 on either the parent or the daughter strand. The *FMR1* data, however, do not provide enough information for separate estimation of parent and daughter de novo rates at most sites.

4. Data at sites 25–52 indicate the occurrence of de novo events on the parent strand, and that the parent de novo rate does not vary much across CpG sites. In comparison, we cannot rule out the possibility of no de novo events on the daughter strand.

5. Data at sites 1–22 are consistent with the temporal stationarity assumption, whereas data at sites 25–52 indicate that at least a third of the sites have devi-

ated from temporal stationarity (Figure 3.12). The deviation, if not due to sampling, can be quite interesting: there is the suggestion that aging can increase or decrease methylation (Richardson, 2003); those sites that show deviation from stationarity might be an indication of aging of the molecules.

To model processivity for methyltransferases, the hierarchical multi-site models are no longer adequate, because they assume that methylation events happen independently across CpG sites. We have developed hidden Markov models (HMMs) for this purpose and hence provided the first (to our knowledge) statistical analysis of the in vivo processivity in methyltransferases (see Table 4.6 for comparisons with two in vitro studies). Briefly, we model possible behaviour of methyltransferases as stationary continuous-time Markov chains that produce maintenance, parent de novo and daughter de novo events. We have incorporated physical distances between CpG sites into the HMM. Sojourn times, times (or distances in this case) during which an enzyme remains in a state (active or inactive), are then natural measures of processivity. Results from applying the HMM to the *FMR1* data provide strong evidence for processivity in maintenance methyltransferases, but little evidence for processivity in de novo methyltransferases. The average sojourn time for maintenance events are 347–1716 nucleotides (or about 52–256 CpG sites), and the average sojourn time for failure of maintenance events are 6–17 nucleotides (or about 1–3 CpG sites). The strong processivity of maintenance methyltransferases might provide an intriguing explanation for the clustering of methylated CpGs in some CpG islands in the genome.

5.2 Future work

As mentioned in Section 4.5 in Chapter 4, we are currently extending the hidden Markov models for processivity of methyltransferases to more realistic versions. The extensions are mainly for maintenance methyltransferases, for which the *FMR1* data contain plenty of information. One extension is to allow these enzymes to methylate

with certain probability CpG sites unmethylated before replication, and the other extension is to allow this enzyme, when bound to a CpG, hemimethylated or unmethylated, fails to methylate the cytosine with certain probability.

Another possible future direction is to go beyond double-stranded data and make use of the more widely available single-stranded data, extending methods developed in this thesis. The experiment to obtain single-stranded data is much easier than hairpin-bisulfite PCR, and it is also easier to obtain data from longer strands, which would provide methylation patterns from more CpG sites in general. Existing methods for estimating rates of methylation rates (details in Appendix B) generally ignore the multi-site information and reduce the data to methylation densities at each site. Failure of maintenance and de novo rates are not identifiable under those methods without additional assumptions. There is no doubt that single-stranded-ness presents great challenges to separate estimation of failure of maintenance and de novo rates. On the other hand, similar to runs of hemimethylated dyads in double-stranded data, runs of unmethylated sites in single-stranded data, from hypermethylated regions in particular, may contain information that is helpful to distinguish failure of maintenance and de novo events. A multi-site approach that exploits this information will be quite valuable.

BIBLIOGRAPHY

Barlow, D., R. Stöger, B. Hermann, K. Saito, and N. Schweifer (1991). The mouse insulin-like growth factor type-2 receptor is imprinted and closely linked to the *Tme* locus. *Science 349*, 84–87.

Besag, J. (2001). Markov chain Monte Carlo for statistical inference. *University of Washington, Center for Statistics and Social Sciences. Working paper*, No. 9.

Bird, A. (1978). Use of restriction enzymes to study eukaryotic DNA methylation: II. the symmetry of methylated sites supports semi-conservative copying of the methylation pattern. *Journal of Molecular Biology 118*, 49–60.

Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes and Development 16*, 6–21.

Burden, A., N. Manley, A. Clark, S. Gartler, C. Laird, and R. Hansen (2005). Hemimethylation and non-CpG methylation levels in a promoter region of human LINE-1 (L1) repeated elements. *Journal of Biological Chemistry 280(15)*, 14413–14419.

Chen, Z. and A. Riggs (2005). Maintenance and regulation of DNA methylation patterns in mammals. *Biochemistry and Cell Biology 83*, 438–448.

Clark, S., J. Harrison, C. Paul, and M. Frommer (1994). High sensitivity mapping of methylated cytosines. *Nucleic Acid Research 22*(15), 2990–2997.

Clark, S. and J. Melki (2002). DNA methylation and gene silencing in cancer: which is the guilty party? *Oncogene 21*, 5380–5387.

Cubas, P., C. Vincent, and E. Coen (1999). An epigenetic mutation responsible for natural variation in floral symmetry. *Nature 401*(6749), 157–161.

Dean, W., D. Lucifero, and F. Santos (2005). DNA methylation in mammalian development and disease. *Birth Defects Research (Part C) 75*, 98–111.

Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological) 39*(1), 1–38.

Dittrich, B., W. Robinson, H. Knoblauch, K. Buiting, K. Schmidt, G. Gillessen-Kaesbach, and B. Horsthemke (1992). Molecular diagnosis of the Prader-Willi and Angelman syndromes by detection of parent-of-origin specific DNA methylation in 15q11-13. *Human Genetics 90*, 313–315.

Feinberg, A. and B. Vogelstein (1987). Alterations in DNA methylation in human colon neoplasia. *Seminars in Surgical Oncology 3*(3), 149–151.

Foster, S., D. Wong, M. Barrett, and D. Galloway (1998). Inactivation of p16 in human mammary epithelial cells by CpG island methylation. *Molecular and Cellular Biology 18*(4), 1793–1801.

Frommer, M., L. McDonald, D. Millar, C. Collis, F. Watt, G. Grigg, P. Molloy, and C. Paul (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proceedings of the National Academy of Sciences of the United States of America 89*(5), 1827–1831.

Genereux, D., W. Johnson, A. Burden, R. Stöger, and C. Laird (2008). Errors in the bisulfite conversion of DNA: modulating inappropriate- and failed-conversion frequencies. *submitted*.

Genereux, D., B. Miner, C. Bergstrom, and C. Laird (2005). A population-epigenetic model to infer site-specific methylation rates from double-stranded DNA methylation patterns. *Proceedings of the National Academy of Sciences of the United States of America 102*, 5802–5807.

Gowher, H. and A. Jeltsch (2001). Enzymatic properties of recombinant Dnmt3a DNA methyltransferase from mouse: the enzyme modifies DNA in a non-processive manner and also methylates non-CpG sites. *Journal of Molecular Biology 309*, 1201–1208.

Gowher, H. and A. Jeltsch (2002). Molecular enzymology of the catalytic domains of the Dnmt3a and Dnmt3b DNA methyltransferases. *Journal of Biological Chemistry 277*(23), 20409–20414.

Goyal, R., R. Reinhardt, and A. Jeltsch (2006). Accuracy of DNA methylation pattern perservation by the Dnmt1 methyltransferase. *Nucleic Acid Research 34*(4), 1182–1188.

Grunau, C., S. Clark, and A. Rosenthal (2001). Bisulfite genomic sequencing: systematic investigation of critical experimental parameters. *Nucleic Acid Research 29*(13), e65.

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika 57*(1), 97–109.

Jaenisch, R. and A. Bird (2003). Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature Genetics 33*, 245–254.

Jeffreys, H. (1961). *Theory of Probability* (Third ed.). Oxford University Press.

Jones, P. and S. Baylin (2002). The fundamental role of epigenetic events in cancer. *Nature Reviews Genetics 3*, 415–428.

Kangaspeska, S., B. Stride, R. Métivier, M. Polycarpou-Schwarz, D. Ibberson, R. Carmouche, V. Benes, F. Gannon, and G. Reid (2008). Transient cyclical methylation of promoter DNA. *Nature 452*(6), 112–116.

Kass, R. and A. Raftery (1995). Bayes factors. *Journal of the American Statistical Association 90*(430), 773–795.

Laird, C. (1987). Proposed mechanism of inheritance and expression of the human fragile-X syndrome of mental retardation. *Genetics 117*, 587–599.

Laird, C., N. Pleasant, A. Clark, J. Sneeden, K. Hassan, N. Manley, J. Vary, T. Morgan, R. Hansen, and R. Stöger (2004). Hairpin-bisulfite PCR: assessing epigenetic methylation patterns on complementary strands of individual DNA molecules. *Proceedings of the National Academy of Sciences of the United States of America 101*, 204–209.

Laird, P. (2003). The power and the promise of DNA methylation markers. *Nature Reviews Cancer 3*, 253–266.

Leonhardt, H., A. Page, H. Weier, and T. Bestor (1992). A targeting sequence directs DNA methyltransferase to sites of DNA replication in mammalian nuclei. *Cell 71*(5), 865–873.

Liu, J. (2001). *Monte Carlo Strategies in Scientific Computing* (First ed.). Springer.

Lyon, M. (1972). X-chromosome inactivation and developmental patterns in mammals. *Biological Reviews 47*(1), 1–35.

Métivier, R., R. Gallais, C. Tiffoche, C. Le Péron, R. Jurkowska, R. Carmouche, D. Ibberson, P. Barath, F. Demay, G. Reid, V. Benes, A. Jeltsch, F. Gannon, and G. Salbert (2008). Cyclical DNA methylation of a transcriptionally active promoter. *Nature 452*(6), 45–52.

Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics 21*(6), 1087–1092.

Miner, B., R. Stöger, A. Burden, C. Laird, and R. Hansen (2004). Molecular barcodes detect redundancy and contamination in hairpin-bisulfite PCR. *Nucleic Acid Research 32*, e135.

Murphy, S. and A. van der Vaart (2000). On profile likelihood. *Journal of the American Statistical Association 95*(450), 449–465.

Okano, M., S. Xie, and E. Li (1998). Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. *Nature Genetics 19*, 219–220.

Ordway, J., J. Bedell, R. Citek, A. Nunberg, A. Garrido, R. Kendall, J. Stevens, R. Cao, R. Doerge, Y. Korshunova, H. Holemon, J. McPherson, N. Lakey, J. Leon, R. Martienssen, and J. Jeddeloh (2006). Comprehensive DNA methylation profiling in a human cancer genome identifies novel epigenetic targets. *Carcinogenesis 27*(12), 2409–2423.

Otto, S. and V. Walbot (1990). DNA methylation in eukaryotes: kinetics of demethylation and de novo methylation during the life cycle. *Genetics 124*, 429–437.

Peterson, S. and N. Reich (2006). GATC flanking sequences regulate Dam activity: evidence for how Dam specificity may influence *pap* expression. *Journal of Molecular Biology 355*(3), 459–472.

Pfeifer, G., S. Steigerwald, R. Hansen, S. Gartler, and A. Riggs (1990). Polymerase chain reaction-aided genomic sequencing of an X chromosome-linked CpG island: methylation patterns suggest clonal inheritance, CpG site autonomy, and an explanation of activity state stability. *Proceedings of the National Academy of Sciences of the United States of America 87*, 8252–8256.

Pradhan, S., A. Bacolla, R. Wells, and R. Roberts (1999). Recombinant human DNA (cytosine-5) methyltransferase. I. expression, purification, and comparison of de novo and maintenance methylation. *Journal of Biological Chemistry 274*(46), 33002–33010.

Rabiner, L. and B. Juang (1986). An introduction to hidden Markov models. *ASSP Magazine, IEEE 3*(1), 4–16.

Reik, W., W. Dean, and J. Walter (2001). Epigenetic reprogramming in mammalian development. *Science 293*(5532), 1089–1093.

Richardson, B. (2003). Impact of aging on DNA methylation. *Ageing Research Reviews 2*(3), 245–261.

Riggs, A. and Z. Xiong (2004). Methylation and epigenetic fidelity. *Proceedings of the National Academy of Sciences of the United States of America 101*, 4–5.

Schermelleh, L., A. Haemmer, F. Spada, N. Rösing, D. Meilinger, U. Rothbauer, M. Cardoso, and H. Leonhardt (2007). Dynamics of Dnmt1 interaction with the replication machinery and its role in postreplicative maintenance of DNA methylation. *Nucleic Acids Research 35*(13), 4301–4312.

Self, S. and K. Laing (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association 82*(398), 605–610.

Severini, T. (2000). *Likelihood Methods in Statistics*. Oxford University Press.

Shiraishi, M. and H. Hayatsu (2004). High-speed conversion of cytosine to uracil in bisulfite genomic sequencing analysis of DNA methylation. *DNA Research 11*(6), 409–415.

Stöger, R., T. Kajimura, W. Brown, and C. Laird (1997). Epigenetic variation illustrated by DNA methylation patterns of the fragile-X gene *FMR1*. *Human Molecular Genetics 6*, 1791–1801.

Strachan, T. and A. Read (2004). *Human Molecular Genetics* (Third ed.). Garland Science.

Taylor, H. and S. Karlin (1998). *An Introduction to Stochastic Modeling* (Third ed.). Academic Press.

Ushijima, T., N. Watanabe, E. Okochi, A. Kaneda, T. Sugimura, and K. Miyamoto (2003). Fidelity of the methylation pattern and its variation in the genome. *Genome Research 13*(5), 868–874.

Vilkaitis, G., I. Suetake, S. Klimašauskas, and S. Tajima (2005). Processive methylation of hemimethylated CpG sites by mouse Dnmt1 DNA methyltransferase. *Journal of Biological Chemistry 280*(1), 64–72.

Appendix A

ISSUES AND SOLUTIONS RELATED TO DESIGN OF PCR EXPERIMENTS

A main goal in collecting DNA methylation patterns is to ensure that the final data set contains a representative sample of the methylation patterns from each individual.

One challenge is to avoid PCR contamination and redundancy. These are common and serious problems for the analysis of PCR-amplified DNA. PCR contamination occurs when a PCR reaction intended to amplify a given sample becomes contaminated by one or more molecules from another PCR reaction. PCR redundancy, actually a defining characteristic of PCR, refers to a PCR product that contains multiple copies of a single molecule from the original DNA sample. These problems are more severe in bisulfite PCR, because the bisulfite conversion process introduces nicks into the backbone of DNA molecules, rendering most of them unamplifiable. To address these issues, the Laird Lab use batchstamps and barcodes (Miner et al., 2004). A unique batchstamp is used to label DNA from each individual and thus enables identification and removal of contaminant sequences; a unique barcode is used to label each molecule collected from a given individual and thus identifies redundant sequences. Contaminant and redundant sequences have been removed prior to assembly of the sequences of the *FMR1* data we have analysed here.

Another challenge is to ensure that the data reflect the true distribution of methylation densities of a population of molecules. Two factors may contribute to sampling bias. First, a particular PCR protocol may favour amplifying either densely methylated or unmethylated sequences, sometimes for unknown reasons. Second, bias toward less densely methylated sequences may occur at the bacterial subcloning step:

one bisulfite-converted, PCR-amplified molecule is transformed into one bacterial cell, making it possible to sequence each molecule individually. Less densely methylated molecules have more unmethylated cytosines that are converted by bisulfite. These sequences yield adenine/thymine (A/T) rich sequences, which may be favoured by some bacteria (Clark et al., 1994). The experiment to collect the data from the *FMR1* locus on the X chromosome, however, has a built-in control for density-based sampling bias: each cell from a female has a densely methylated X chromosome and a sparsely methylated one. The observed numbers of hyper- and hypomethylated sequences do not differ by a statistically significant amount, indicating no substantial bias in this data set.

Appendix B

EXISTING APPROACHES TO ANALYSING SINGLE-STRANDED METHYLATION DATA

We review the methods developed by Otto and Walbot (1990) and Pfeifer et al. (1990), the theoretical framework of which is closely related to the single-site maximum likelihood approach in Genereux et al. (2005), described in Section 2.2. Otto and Walbot (1990) in fact started with parent-daughter pairs of strands, and derived recursive equations for frequencies of methylated (M), hemimethylated (H) and unmethylated (U) CpG dyads, which we denote by (p_M, p_H, p_U), over generations. More specifically, they aimed at estimating maintenance rate μ (parameter α in their notation) defined the same way as we do, and a single de novo rate ν (parameter β in their notation) defined as

$$\nu = \Pr(Q = 1, D = 1 | P = 1), \tag{B.0.1}$$

where P, Q and D, as before, refer to the methylation states at a single CpG site on pre-replication parent, post-replication parent and daughter strands. Otto and Walbot (1990) assumed that de novo events occur simultaneously on parent and daughter strands, which is different from our models. That is,

$$1 - \nu = \Pr(Q = 0, D = 0 | P = 1). \tag{B.0.2}$$

Equations (B.0.1) and (B.0.2) together imply that de novo methylation events do not produce hemimethylated CpG dyads under this model and that those dyads must be attributed to failure of maintenance events. They derived the following recursive

equations:

$$p_M^{n+1} = \frac{2\mu p_M^n + (\mu+\nu)p_H^n + 2\nu p_U^n}{2}, \tag{B.0.3}$$

$$p_H^{n+1} = \frac{2(1-\mu)p_M^n + (1-\mu)p_H^n}{2}, \tag{B.0.4}$$

$$p_U^{n+1} = \frac{(1-\nu)p_H^n + 2(1-\nu)p_U^n}{2}. \tag{B.0.5}$$

Assuming that the transmission process has attained temporal stationarity, they obtained the following expressions:

$$p_M = \frac{\nu(1+\mu)}{1-\mu+2\nu}, \tag{B.0.6}$$

$$p_H = \frac{2\nu(1-\mu)}{1-\mu+2\nu}, \tag{B.0.7}$$

$$p_U = 1 - p_M - p_H = \frac{(1-\mu)(1-\nu)}{1-\mu+2\nu}. \tag{B.0.8}$$

For the single-stranded sequence data, the above equations are reduced to the following one:

$$m = \frac{2\nu}{1-\mu+2\nu}. \tag{B.0.9}$$

They then used the empirical methylation density $\hat{m}$ to estimate relative efficiency μ/ν.

Since the single-stranded methylation data provide information only on m rather than (p_M, p_H, p_U), Pfeifer et al. (1990) worked directly with counts of methylated and unmethylated CpG dinucleotides, denoted by M and U. They obtained differential equations over time for maintenance rate E_m (μ in our notation) and de novo rate E_d, which is equivalent to 2ν in the Otto and Walbot method:

$$\partial M/\partial t = aE_m M + aE_d U, \tag{B.0.10}$$

$$\partial U/\partial t = b(1-E_d)U + b(1-E_m)M, \tag{B.0.11}$$

where a and b are cell-growth rate constants. At temporal stationarity,

$$\frac{\partial m}{\partial t} = \frac{\partial(M/(M+U))}{\partial t} = 0, \tag{B.0.12}$$

which is

$$\frac{\partial M}{\partial t} - m\left(\frac{\partial M}{\partial t} + \frac{\partial U}{\partial t}\right) = 0. \tag{B.0.13}$$

Let $a = b$ as in Pfeifer et al. (1990). Equations (B.0.10), (B.0.11) and (B.0.13) then give rise to the following equation that is similar to equation (B.0.9) under the Otto and Walbot method:

$$m = \frac{E_d}{1 - E_m + E_d}. \tag{B.0.14}$$

The Pfeifer et al. approach does not need to assume parent and daughter de novo events to occur at the same time, even though there is still no distinction between parent and daughter de novo rates. Furthermore, it also allows for the possibility that methylation can be lost on the template strand.

Both methods (Otto and Walbot, 1990; Pfeifer et al., 1990) assume that CpG sites are independent and that the process has attained stationarity. They also both need additional constraints to separately estimate maintenance and de novo rates, which are some weighted averages of the parent and daughter de novo rates.

VITA

Qiuyan Fu earned a Bachelor of Science degree in Management Science from Fudan University, Shanghai, China, in 1999, a Master of Science degree in Statistics from the University of British Columbia in 2002, and a Doctor of Philosophy in Statistics (Statistical Genetics Track) from the University of Washington in 2008.