Analyzing the Loss of Allele-Specific Methylation in Human Colorectal Adenoma with Bisulfite Sequencing

Master Thesis

Author(s):
Machlab, Dania

Publication date:
2015

Permanent link:
https://doi.org/10.3929/ethz-a-010489852

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Analyzing the Loss of Allele-Specific Methylation in Human Colorectal Adenoma with Bisulfite Sequencing

Master Thesis

Dania Machlab

July 7, 2015

Supervisor: Prof. Dr. Mark D. Robinson
Statistical Bioinformatics Group UZH

CBB Masters at Department of Computer Science, ETH Zürich

Abstract

DNA methylation, the addition of a methyl group to a Cytosine, can alter gene expression. It has served as an explanation for the progression of several diseases and cancers. In particular, there are a number of regions that lose their normal state of showing allele specific methylation whereby one parental allele exhibits methylation and the other one does not. Tools like `amrFinder` and `bsseq` search for regions that show allele-specific methylation and differential methylation, respectively. We have applied these tools on BS-seq data of normal and adenoma colorectal lesions in an effort to look for loss of allele-specific methylation and imprinting. We found a multitude of regions that displayed loss of allele-specificity in all adenoma samples. To assess the regions predicted to be allele-specific, we developed a scoring function of our own. This function performed well compared with `allelicmeth` but was too strict with methylation imbalances.

Acknowledgements

I wish to express my sincere thanks to my supervisor, Prof. Dr. Mark Robinson, for letting me be part of a wonderful learning environment in his group, for his support and for the encouragement. I also extend my thanks to every member of the statistical bioinformatics group of Dr. Mark Robinson for their helpful insights.

I want to thank Mirco Menigatti for the wet lab work on the BS-seq reads and to Abdullah Kahraman for his help in aligning our reads with Bismark and running amrFinder and BisSNP.

Special thanks to Helen Lindsay and Mark for reviewing this thesis and for the great suggestions.

Contents

Chapter 1

Introduction

Over the past decades and especially the last one, epigenetics has grown more and more as a prominent field in biology. The authors of [19] define epigenetics as ”the study of the mechanisms of inheritance and control of gene expression that do not involve permanent changes in the DNA sequence. Such changes occur during somatic cell division and sometimes can be transmitted transgenerationally through the germline”. The notion of only the DNA sequence sufficing as an explanation to changes in gene expression has been pushed aside with the revelations made in epigenetics. Instead, changes in DNA methylation and histone modification have served as explanations for some of the shifts in gene expression and progression of diseases and cancers. ”Epigenome” simply means above the genome – explanations beyond the actual DNA sequence. In an effort to look at allele-specific methylation and its loss in pre-colorectal cancer, we used available tools on Bisulfite sequencing (BS-seq) reads to identify such regions and developed a score that aims to further assess the allele-specificity of these regions.

In this chapter we introduce the epigenetic mechanism of DNA methylation, genomic imprinting and its importance in the study of disease. We also introduce the available technologies to look at DNA methylation profiles on a genome scale, the alignment tools, and the nature of the given data set and on which all subsequent analysis was based on.

1.1 Biology Background

The study of DNA methylation within the concept of epigenetics began as early as the 1980s when correlations between the level of Cytosine methylation at CpG DNA sequences and the level of gene transcription were discovered [19].

1.1.1 Methylation Mechanism

DNA methylation consists of the covalent attachment of a methyl group to a Cytosine residue at position C-5 In mammals, his happens predominantly to Cytosines that are followed by a Guanine and are said to fall in a 5'-CpG-3' (Cytosine phosphate Guanine) context [19].. This is a mitotically heritable epigenetic modification.

Enzymes called methyltransferases regulate DNA methylation. There are two types of methyltransferases in mammals: *de novo methyltransferases* (DNA methyltransferase 3), which establish methylation, and *maintenance methyltransferases* (DNA methyltransferase 1), which maintain methylation. The DNA methyltransferases (DNMTs) have 10 conserved motifs [19]. On the other hand, there are two types of enzymes involved in de-methylation: activation-induced cytosine deaminase (AID) and apolipoprotein B RNA-editing catalytic component 1 (APOBEC1). AID deaminates 5-methyl Cytosines and results in T:G mismatches [28]. However AID-dependent de-methylation is probably not the main de-methylation process in mammals and other processes might be involved.

1.1.2 CpGs and CpG Islands

Changes in DNA methylation and histone modifications have been shown to alter gene expression levels. Cytosines that occur followed by a guanine (CpG context) tend to occur in clusters on the genome called CpG islands (CGIs). These islands typically occur near promoter regions of genes. 70 to 80 % of cytosines that fall in a CpG context are methylated in mammals [36]. The human genome has an average GC content of about 42% but the frequency of CpG dinucleotides is less than 1% [19].

CGIs are defined as regions that have at least 200 base pairs and an observed to expected ratio of greater than 60% [19], with that ratio calculated as follows:

$$observed = (number\ of\ CpGs) * (length\ of\ the\ sequence)$$

$$expected = (number\ of\ Cs) * (number\ of\ Gs)$$

Usually, CpG sites in the CGIs of promoters are unmethylated allowing the expression of the genes. Gene silencing has been observed once these sites become methylated. Methylation can interfere with the binding of transcription factors or alter chromatin structure. However, the opposite can also happen, where methylation may induce gene expression. In cancers, for example, tumor suppressor genes are no longer expressed after methylation. Figure 1.1 illustrates an example of this.

In mammals, CGIs have been found in or near promoter regions about 40% of the time [19]. CGI shores are CpG regions that occur 2000 bp away from the CGIs. Their methylation has also been linked to gene expression changes in cell differentiation and cancers. In some regions methylation changes at CpG shores played the key role in gene expression rather than methylation at the CGIs [16].

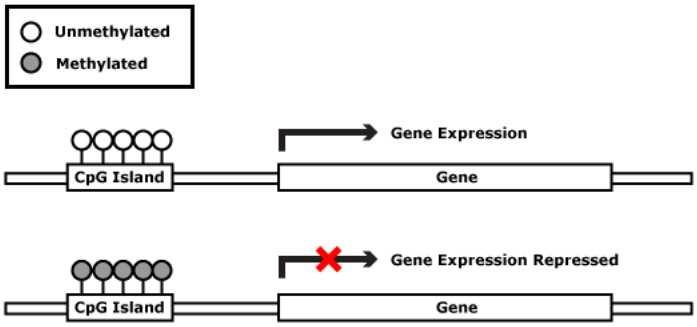


Figure 1.1: Gene silencing from methylation on the CpG sites [27]

.

1.1.3 Genomic Imprinting

Imprinted genes are regulated by epigenetics and have been the focus of many studies, especially in development and disease progression.

Genomic imprinting has been established in fungi, plants, and animals and results from epigenetic processes involving DNA methylation and histone modification. It is the circumstance whereby alleles and thereby genes are expressed in a parent of origin manner. Humans inherit two copies of every autosomal gene: one from the father and one from the mother and usually both are expressed. However, with some genes, one copy is turned off in a parent-of-origin dependent manner [18]. For example, if the allele from the mother is imprinted, then the paternal allele is expressed and the maternal one is not.

There have been some suggestions for the evolution of imprinting in mammals. If we consider a pregnant female, the more nutrients the embryo gets, the bigger it gets and the more likely it is to survive after birth. However, a greater nutrient demand from the pregnancy may have costs on the mother's potential future reproduction. There is thus a conflict of interest, because the mother's future offspring may have a different father [26]. That is why paternally imprinted genes tend to be growth promoting (greater fitness for the offspring at the expense of the mother) and maternally imprinted genes growth limiting.

Imprinting has been described in mammalian developmental processes, especially

during embryonic development. Several diseases have been associated with loss of imprinting. Some examples are Angelman and Prader-Willi Syndromes, Alzheimer's disease, diabetes, obesity, and schizophrenia, as well as a number of cancers: bladder, breast, cervical, colorectal, esophageal, hepatocellular, lung, mesothelioma, ovarian, prostate, testicular, and leukemia and more [18].

1.2 Technologies

We present some of the current methods for DNA methylation profiling. There are various tools that profile DNA methylation and they differ in their capture specificities, target regions and costs.

1.2.1 Introducing Bisulfite Treatment

Bisulfite treatment of DNA fragments followed by PCR amplification has become widely popular in the examination and study of DNA methylation regions. Illumina sequencing has enabled the sequencing of a vast amount of these bisulfite treated DNA fragments in an assay commonly called BS-seq.

DNA fragments are treated with Bisulfite which transforms unmethylated Cytosine into a Uracil that is later converted to Thymine during PCR amplification. The methylated Cytosines remain unchanged and stay as Cytosines in the amplification process [22]. The method was first introduced by Frommer et al [8].

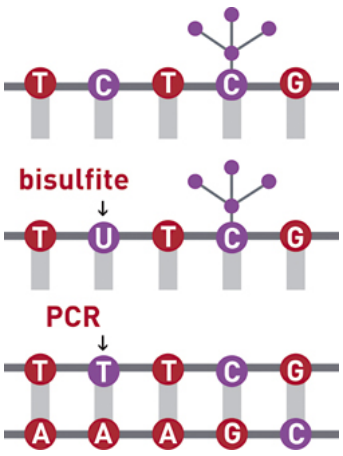


Figure 1.2: Bisulfite treatment [5]

1.2.2 Tools for DNA Methylation Profiling

We give a brief summary on the existing technologies to profile methylated DNA.

Methylation microarrays

Bisulfite-converted DNA can be hybridized to a microarray. There are two bead types for each CpG site per locus. Attached to the beads are oligonucleotide sequences that differ only at the free ends. One of the beads corresponds to the methylated Cytosine locus, and the other bead to the unmethylated one. Figure 1.3 exemplifies this. The amplified DNA fragments hybridize to the appropriate oligonucleotide via allele-specific annealing. After hybridization, the oligonucleotides are extended by a single base (using labeled nucleotides). The level of methylation per locus is determined by the ratio of the fluorescent signals from the methylated vs unmethylated sites [12].

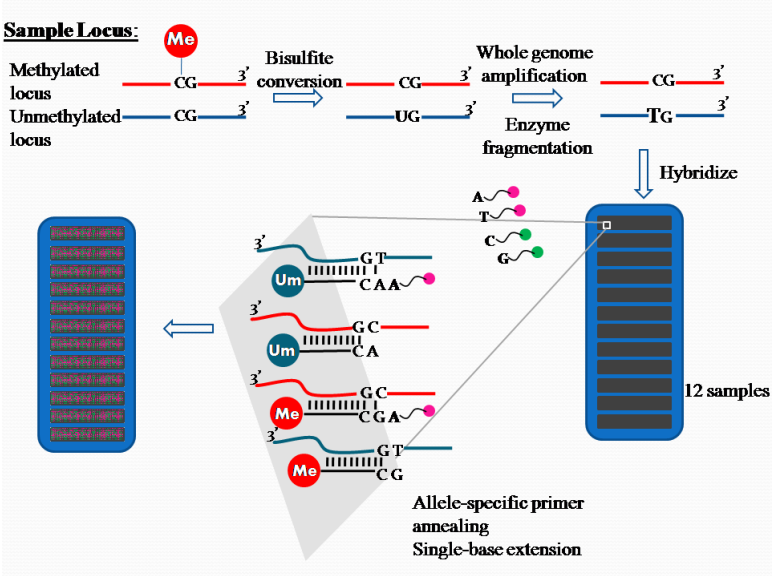


Figure 1.3: Methylation microarrays [34]

Whole genome bisulfite sequencing (WGBS)

In this method, genomic DNA is bisulfite treated before the addition of tags. The DNA fragments are then sequenced with random primer extension. For large sample numbers, this technique is time consuming and costly with costs ranging from $5000 to $6000 per sample [14].

Reduced representation bisulfite sequencing (RRBS)

This method is used to reduce the portion of the genome to be analyzed as well as sequencing cost. A methylation insensitive enzyme is used to digest the DNA. The

enzyme cuts at the CCGG sequence, and thus the enzyme targets CpG sites. After end-repair is done to the sticky ends of the DNA fragments, adapters are ligated. The various fragments are then separated by gel electrophoresis and the desired sizes are extracted. These are then treated with bisulfite. The fragments then undego PCR amplification followed by next generation sequencing [25]. The reads are aligned to a reference genome using one of the existing alignment tools for BS-seq data. This method is biased towards repeats and CpG rich sequences.

MeDip-Seq

Methylated DNA immunoprecipitation (MeDip) uses antibodies that target 5-methyl cytosine to enrich for methylated DNA. DNA is extracted and fragmented using sonication. The fragments are ideally 100 to 300 bp long. The DNA fragments are then denatured and stored with the antibody. Immunoprecipitation follows [14]. The antibody is more likely to bind the more methylated cytosines there are. The method is thus biased towards repeat and CpG-rich sequences. MeDip-seq couples MeDip with next generation sequencing to produce a large number of fragments that are then aligned to a reference genome.

1.2.3 Paired End vs Single End Reads

During PCR, the DNA polymerases recognise the primers on DNA fragments and initiate replication. When for each DNA fragment another fragment is produced, we refer to these reads as SE reads. With paired end (PE) reads two reads are produced per fragment. One is produced in the forward direction (called R1) from the start of the fragment. The other (called R2) is produced from the 3′ end of the original fragment. A data set that consists of PE reads rather than SE reads is generally of higher quality, since the PE reads are more likely to map to the reference genome, especially when it comes to repeats. Figure 1.4 illustrates this. PE reads convey more information on the position of the fragment than SE reads do.

1.2.4 Alignment Tools for BS-seq data

Table 1.1 gives a comprehensive summary on some of the popular alignment tools that are available for BS-seq reads as presented by the authors of [33] who show that `bismark` performs the best on real data followed by BiSS, BSMAP, and finally BRAT-BW and BS-Seeker with very similar performance. `Bismark` is thus a good choice for a mapping program if CPU time is not a constraint. Other available tools include ERNE-bs5, BatMeth, RMAP, MAQ, PASH, Novo-align, Methyl-coder, GSNAP, BFAST and Segemehl. We used `bismark` to align our BS-seq reads.

1.2.5 Our Chosen Technology: SureSelect

SureSelect was the technology that was used to produce BS-seq data because it is more cost effective and allows for multiple samples to be analyzed. SureSelect targets

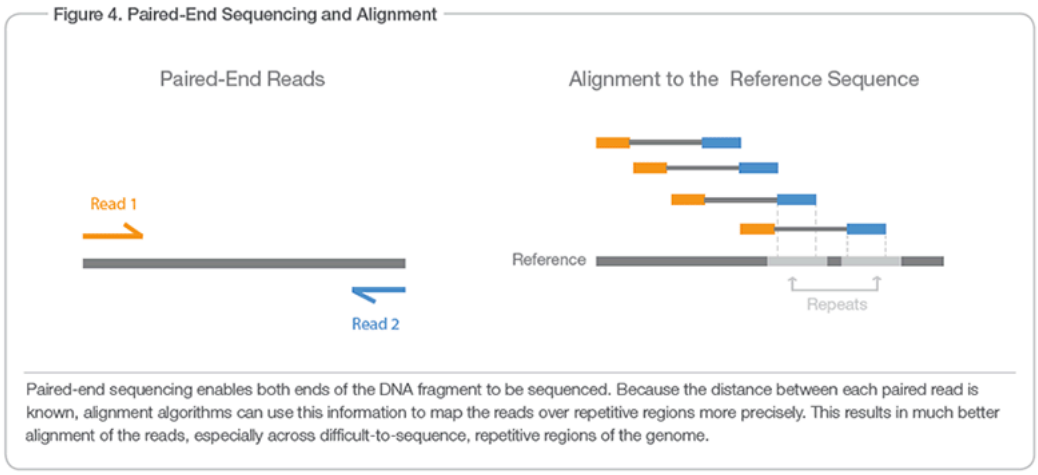


Figure 1.4: PE reads and DNA repeats [13]

Table 1.1: BS-seq alignment tool overview

program	year	algorithmic technique used	aligner	min/max read length	mis-matches	indels	gaps	single/ paired-end	multi-threaded
BSMAP	2009	hashing of reference genome and bitwise masking tries all possible T to C combinations for reads	SOAP	up to 144 bp	up to 15 in a read		up to 3 bp	both	yes
bismark	2011	FM-Index enumerates all possible T to C conversion	bowtie/ bowtie2	bowtie: up to 1000 bp bowtie2: unlimited	0 or 1 in a seed	yes	yes	both	yes
BS-Seeker	2010	FM-Index, enumerates all possible T to C conversion, converts the genome to 3 letters, and uses Bowtie to align reads	bowtie	50–250 bp	up to 3 per read	yes	no	single	no
BS-Seeker2	2013	FM-Index enumerates all possible T to C conversion	Bowtie2/Bowtie/ SOAP/RMAP	50-500 bp	up to 4 per read	yes	yes	single	no
BiSS	2012	Reference genome hashing, local Smith-Waterman alignment	none	up to 4096 bp	($-i$ from 0 to 1) in a read default $i = 65\%$	yes	yes	yes	yes
BRAT-BW	2012	Converts a TA reference and CG reference; two FM indices are built on the positive strand of the reference genome		32bp-unlimited	unlimited	no	no	both	yes

assigned regions of the genome and allows over 3.7 million CpG sites to be analyzed. CpG islands as well as CpG shores and shelves which are found around 4000 bp away on either side of the islands are targeted. It also captures regions that are known to be differentially methylated in cancer. The targets in SureSelect are captured regardless of their methylation state [31]. Figure 1.5 shows the workflow in SureSelect. Genomic DNA is fragmented and the library is prepared (addition of adapters to the fragments). The samples are then hybridized with biotinylated RNA library baits. Target regions are selected with the magnetic streptavidin beads and amplification and sequencing follow [30].

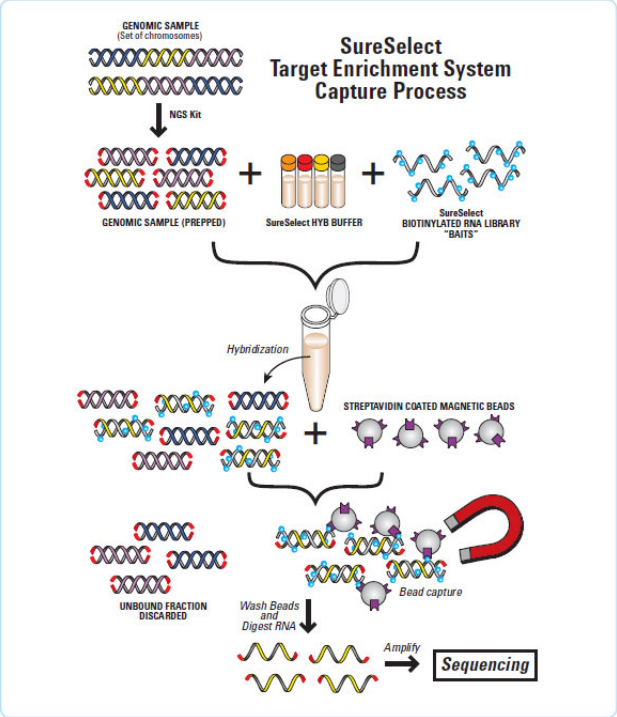


Figure 1.5: SureSelect [31]

1.3 Data Set

In this section we introduce our data set. We have paired end BS-seq reads from 13 patients, 3 of which are normal samples from healthy individuals and ten of which are adenoma samples from individuals who have pre-colorectal cancer.

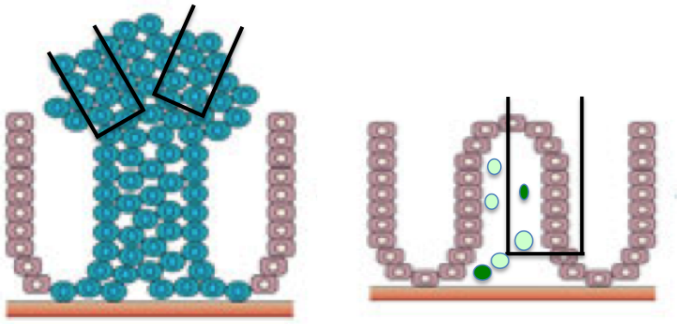
1.3.1 Origin

The reads came from lesions made on the colons of normal crypts and adenoma ones as shown in figure 1.6. In the adenomas, stromal contamination was accounted for. The epithelial cell content – the adenoma is at the level of the epithelial cells (mucosa) – was estimated to be around 90% by qPCR evaluation of vimentin expression, which is a stromal marker that is abundant in the colon lamina propria specimens.

Table 1.2 presents the samples we have and the assigned genders. All normal samples were those of females.

1.3.2 Overview on Colorectal Cancer

We give a short summary on colorectal cancer since this is the condition that some of our data stems from. Colorectal cancer involves the development of cancer in the



(a) Adenoma sample (b) Normal sample

Figure 1.6: Lesions made at the colon

Table 1.2: Sample information

sample number	condition	gender	age
2852	normal crypt	Female	42
5222	normal crypt	Female	44
5223	adenoma	Male	72
5224	adenoma	Male	77
5225	adenoma	Male	NA
5226	adenoma	Female	78
5227	adenoma	Female	85
5228	adenoma	Male	74
5229	adenoma	Female	70
5230	adenoma	Male	61
5231	adenoma	Male	73
5232	normal crypt	Female	79
5233	adenoma	Female	73

colon or rectum and some of the risk factors are older age, lifestyle, inherited genetic disorders and a family history of the disease [15].

Around 1 million individuals are diagnosed with colorectal cancer every year in the world. Around 85% of the cases have chromosomal instability, allelic imbalance and chromosome amplification and translocation. The other 15% are mainly due to microsatellite instabilities and mutations that occur in tandem repeats [4].

Our data was taken at the adenoma stage. Adenomas usually grow on a stalk, resembling small mushrooms and they grow slowly over a period of time (a decade or more). The adenoma is more likely to become a cancer the bigger it is and the longer it has been growing. Almost all colorectal cancers start in the mucosa, i.e the innermost lining, of the large intestine [3].

1.4 Goals

The aim is to look at differential methylation and loss of allele-specific methylation (ASM) from the normal state in the DNA of the colon to the adenoma state. The adenoma stage is a pre-cancerous one. Finding regions that lose their ASM in the adenoma stage may be an indicator or marker for the progression of colorectal cancer that can be further investigated. Ultimately, we are interested in loss of ASM and imprinting in the adenoma stage and how much we can trust the predicted ASMs that tools like `amrFinder` generate.

Chapter 2

Data Quality Control

In this chapter we do some some checks to evaluate the quality of our reads. Tools like `fastQC` and `TEQC` were used to assess the quality at the single base level of the reads to be sure that the coverage is sufficient and to look for other indicators that can be informative on our BS-seq reads.

2.1 FastQC

Quality control was done on the raw PE BS-seq reads to be sure that the data was of sufficient quality and that there were no particular biases. This is step was important to decide if trimming was necessary if we had low quality read ends. The reads were subsequently trimmed at the ends before they were mapped to the reference genome: the human genome hg 19. Figure 2.1 shows some of the outputs that resulted from `fastQC` for the R1 reads of the adenoma sample number 5227. The rest of the samples showed similar results.

Figure 2.1a shows the read length distribution. We see a peak at a read length of 101 base pairs, signifying that there wasn't so much variability in the sequence lengths.

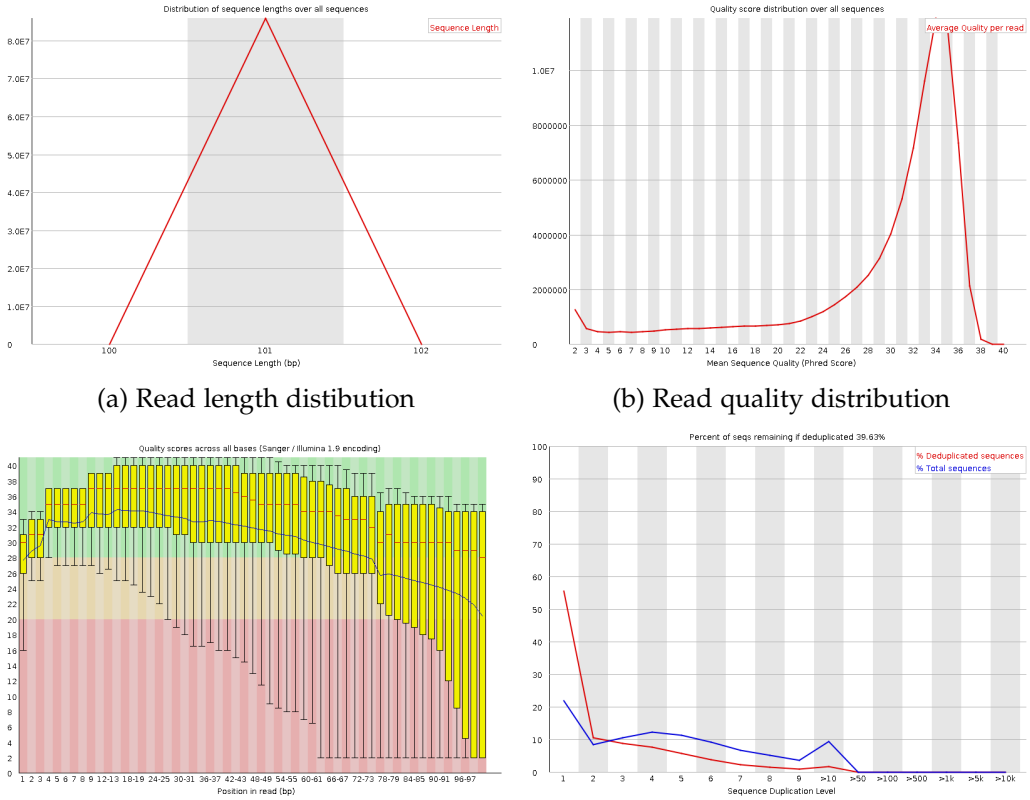
Figure 2.1b shows the quality score distribution. Most of the reads had good scores as we see a peak on the right hand side. The shape of the curve isn't as narrow as it could have been. For high numbers of good quality reads this distribution is more narrow at the high scores. We already see an increase in the frequency of reads that show score in the 20s or lower since the red line is above the x-axis. We also notice the bump at the beginning indicating a number of reads that were of poor quality.

Figure 2.1c shows the read quality distribution at each position of the reads. The reads are each 101 base pairs long. At each position we see the distribution of quality scores across all reads at that position. The red line is the median value and the blue line is the mean value. The yellow box plots represent the inter-quartile range (25% to 75%) and the upper and lower whiskers represent the 10% and 90% points. A score below 20 is considered to be bad and usually discarded. We see that the right hand side shows some bad scores, well below 20. The mean value drops at the end of read R1.

Figure 2.1d indicates duplication levels. Only the first 200,000 sequences were taken into consideration for this plot, to serve as an estimate. For this sample, most of the reads show no duplication. However, we do observe larger numbers of duplicates on the x-axis – as high as 5 and 10 thousand. There is a small peak at duplication levels of more than 10.

2.1.1 Trimmomatic

Figure 2.1c indicates that a certain amount of trimming needed to be done at the end of the reads, to remove bad quality base pairs. This allows `bismark` [20] to map the reads more effectively to the reference genome. We used `trimmomatic` [2] to remove the leading and trailing bases that had a quality score of less than 20 for each of the forward and reverse reads of the PE reads. Four output files were produced: one where both reads survived, two where one read survived but not the other, and one where both reads did not survive. For the analysis steps that followed (mapping to the genome etc.), we used the output files of `trimmomatic` called "paired" where both reads had survived trimming. Table 2.1 shows the number of reads that were kept



(a) Read length distibution

(b) Read quality distribution

(c) Per base quality

(d) Duplication levels

Figure 2.1: `FastQC` figures on R1 reads of adenoma sample 5227

after `trimmomatic`. These were subsequently mapped to the reference genome with `bismark`.

Table 2.1: `Trimmomatic` Results

sample number	input read pairs	both surviving	% both surviving	% forward only surviving	% reverse only surviving	% dropped
2852	36,314,707	35,849,794	98.72	0.98	0.26	0.05
5222	70,620,501	67,936,943	96.20	2.37	0.86	0.57
5223	87,564,110	82,513,943	94.23	3.65	1.30	0.81
5224	109,561,326	103,815,645	94.76	3.36	1.12	0.77
5225	122,814,306	116,783,014	95.09	3.46	0.87	0.58
5226	59,113,922	55,629,810	94.11	4.36	0.92	0.62
5227	85,814,449	83,245,843	97.01	1.74	0.77	0.48
5228	108,214,132	102,181,678	94.43	3.55	1.14	0.88
5229	94,294,386	89,619,863	95.04	2.87	1.20	0.89
5230	74,479,955	71,785,892	96.38	2.08	0.92	0.61
5231	41,567,074	36,604,826	88.06	10.78	0.49	0.66
5232	113,665,000	98,524,229	86.68	11.85	0.57	0.90
5233	107,644,931	93,805,265	87.14	11.39	0.59	0.88

Figure 2.2 shows the quality score distributions of the bases on the R1 reads for normal 5222 before and after `trimmomatic` was used. We see an improvement in the score distribution after having removed the low quality bases. The x-axis shows the position on the read, and the y-axis the quality distribution in the reads at that position.

Box 2.1 shows how `trimmomatic` was run.

```
$ java -Xmx2G -cp ~/bin/Trimmomatic-0.32/trimmomatic-0.32.jar org.usadellab.
    trimmomatic.TrimmomaticPE -threads 2 -phred33 NGS-5222_R1.fastq.gz NGS-5222
    _R2.fastq.gz NGS-5222_R1_t20l20_paired.fastq.gz NGS-5222_R1_t20l20_single.
    fastq.gz NGS-5222_R2_t20l20_paired.fastq.gz NGS-5222_R2_t20l20_single.fastq
    .gz LEADING:20 TRAILING:20 >& NGS-5222_t20l20_trim.out
```

Box 2.1: Commands for `trimmomatic`

2.2 TEQC

Bioconductor's Target Enrichment Quality Control (`TEQC`) [11] was used to assess how well the mapped BS-seq reads covered the target regions (Sureselect target regions). This was done for all 13 samples and Figure 2.3 shows some example figures that were generated for the adenoma sample number 5225. The interpretations are similar for the figures generated for the rest of the samples.

Sample 5225 had a total of 90,854,576 read pairs and a target size of 350,537 target regions (from SureSelect). Figure 2.3c gives an understanding of how much of the genome these targets cover. The yellow bars represent the fractions of the target bases on the genome (per chromosome).

Figure 2.3a shows the distribution of the read pair insert sizes. `FastQC` indicated that the individual reads had a length of around 101 bp. We see that the PE reads cover sizes of 198 on average. The majority of DNA fragments are between between 137 and 260 base pairs long.

Figure 2.3b is the coverage histogram for sample 5225. We see the fraction of the target bases that have the coverages shown on the x-axis. 90% of the target bases have a coverage of at least 8.

The sample had a capture specificity of 89.17%. This value was measured by looking at the fraction of reads that overlapped with the targets. A read pair was considered on-target if at least one of the reads overlapped with a target region by at least 1 bp. An enrichment value of 33 was obtained. The enrichment score is defined as (# on-target read pairs / # aligned reads) / (target size / genome size).

Table 2.2 summarizes on the specificity (fraction of reads on target) and enrichment values of all samples, keeping in mind that an even an overlap of just 1 bp is accepted as covering the target.

Table 2.2: Specificity values for PE reads on their targets across the samples

sample	condition	capture specificity (in %)	enrichment
2852	normal crypt	89.85	33
5222	normal crypt	87.68	33
5223	adenoma	95.49	35
5224	adenoma	87.59	33
5225	adenoma	89.17	33
5226	adenoma	94.64	35
5227	adenoma	91.93	34
5228	adenoma	94.63	35
5229	adenoma	93.88	35
5230	adenoma	94.62	35
5231	adenoma	94.32	35
5232	normal crypt	88.01	33
5233	adenoma	88.56	33

Figure 2.3c depicts the fraction of reads pairs and targets, respectively, that fall on each chromosome. For the read pairs this fraction is within the total number of read pairs. As for the targets, the fractions of targeted bases on each chromosome were calculated. We would like the amount of reads (green) to correspond more or less to the amount of targets (yellow). This was the case with all our samples.

2.2.1 Overall Coverage

The boxplots of coverage per sample are shown in Figure 2.5. The lowest median coverage across the samples is 37 whilst the lowest average coverage is around 35 reads.

2.3 Mapping Efficiency

The reads were mapped against the reference genome hg19 using `bismark`. Table 2.3 summarizes some of the mapping results in terms of the total read counts and the percentage of reads mapped uniquely. Note that the duplicate reads had not been removed at this stage.

2.4 Methylation Bias

`Bismark` outputted methylation bias files based on which we decided what conditions to follow for the `methtuple` tool, which is further explained in the Methods section. Figure 2.6 reflects the methylation bias plots that `bismark` generated. The plots were the same for all samples. We see the plot for the forward and reverse read (since we have PE reads). The R1 reads (forward read) showed a methylation bias in the first 10 and last 5 base pairs, roughly. The R2 reads (reverse reads) showed a methylation

Table 2.3: Bismark Mapping Efficiency

sample	condition	total reads	unique best hit	mapping efficiency	Cs methylated in CpG context
2852	normal crypt	35,849,794	28,022,127	78.2%	52.0%
5222	normal crypt	67,936,943	44,979,341	66.2%	55.0%
5223	adenoma	82,513,943	64,272,402	77.9%	47.7%
5224	adenoma	103,815,645	80,847,770	77.9%	45.1%
5225	adenoma	116,783,014	90,903,073	77.8%	39.4%
5226	adenoma	55,629,810	41,624,166	74.8%	47.4%
5227	adenoma	83,245,843	42,172,072	50.7%	49.6%
5228	adenoma	102,181,678	81,953,788	80.2%	43.9%
5229	adenoma	89,619,863	70,690,094	78.9%	46.2%
5230	adenoma	71,785,892	55,216,430	76.9%	44.4%
5231	adenoma	36,604,826	24,435,574	66.8%	48.2%
5232	normal crypt	98,524,229	70,274,766	71.3%	53.3%
5233	adenoma	93,805,265	67,546,101	72.0%	42.7%

bias in the first 10 base pairs (the start being the direction in which this read was synthesized: the opposite direction of the R1 read).

Based on these results, we chose to omit the first 10 and last 5 bases of the R1 reads, as well as the first 10 bases of the R2 reads to generate the tuples when we used `methtuple` as explained in section 3.4.

2.5 Duplicates

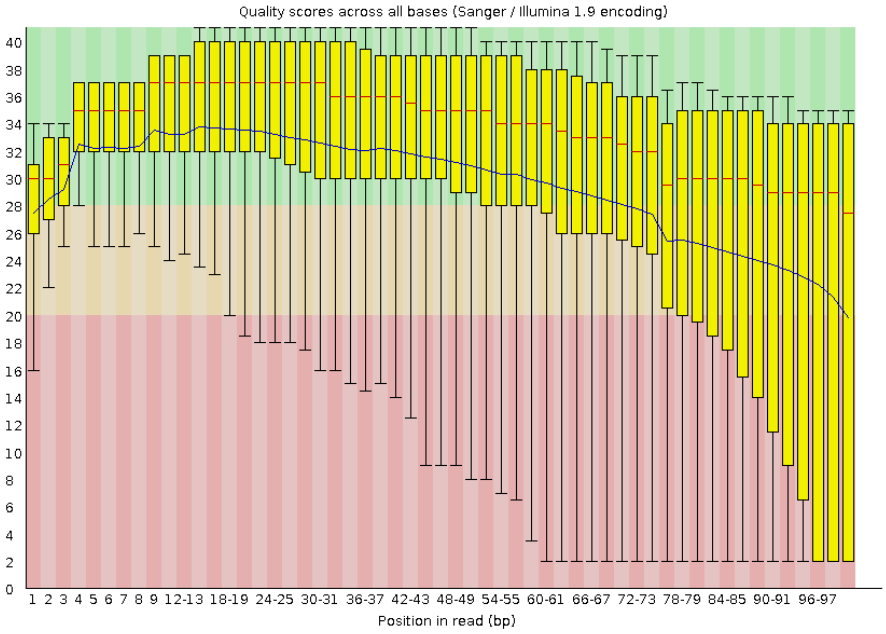
Some of the samples had relatively high amounts of duplicates as will be discussed in the results section. `FastQC` and `TEQC` generated figures that gave some insight into the duplicate situation.

We had relatively large amounts of duplicate reads that were removed. Figure 2.7 shows the percentage of reads that were duplicated across the samples. The numbers range from as small as about 7% to as big as 80%. The removal of the duplicate reads was done with the "duplicate-remover" command found in `bismark`. The duplicates were also independently removed at a certain step in the analysis done with the methpipe tools to use `amrFinder` and get the AMRs. Both methods removed the same amounts of duplicates.

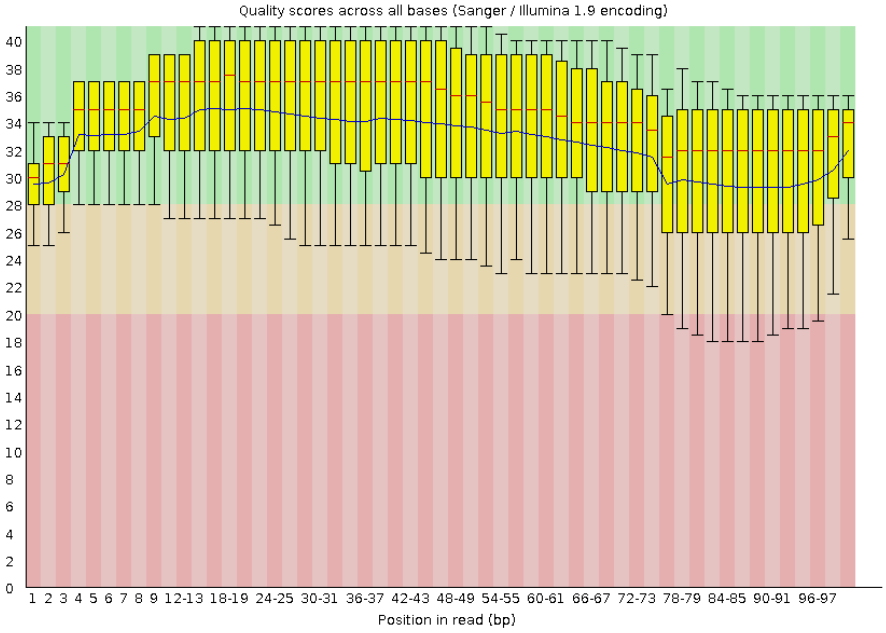
In the early stages of Illumina sequencing, DNA libraries are amplified and then inserted into flow cells for next generation sequencing. If identical fragments from the same original DNA enter different flow cells that particular fragment is amplified much more and this is how duplicates arise. This can misrepresent certain fragments in terms of coverage – giving them a higher read count than they actually have.

Reads are considered duplicated if the first positions of their forward reads (R1) and the first positions of their reverse reads (R2) map to the same position on the genome – as in the PE reads map to the exact same start and end positions on the genome. Ideally, we would want to see a high number of reads that are unique (and have a

read multiplicity of 1). Yet Figure 2.4 shows us that there is a high multiplicity in the reads and a considerable fraction of reads that show multiplicity.

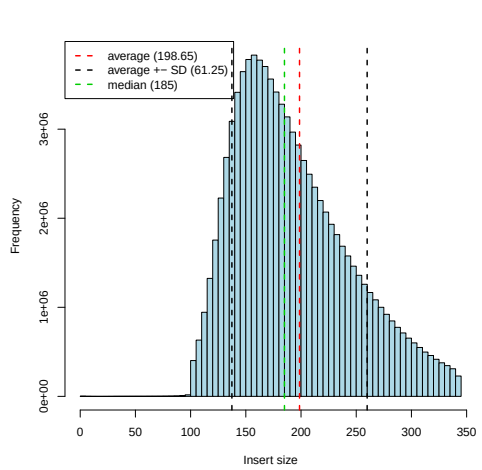


(a) Read quality of the raw reads

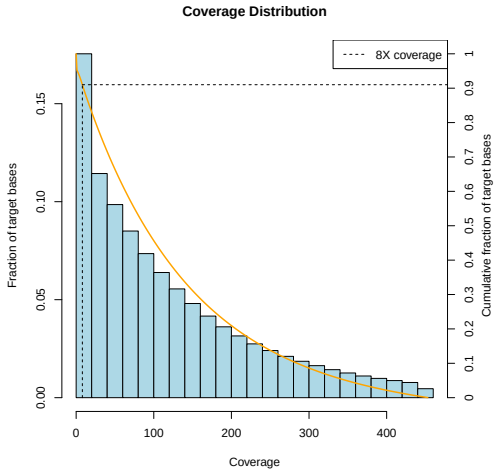


(b) Read quality after Trimmomatic

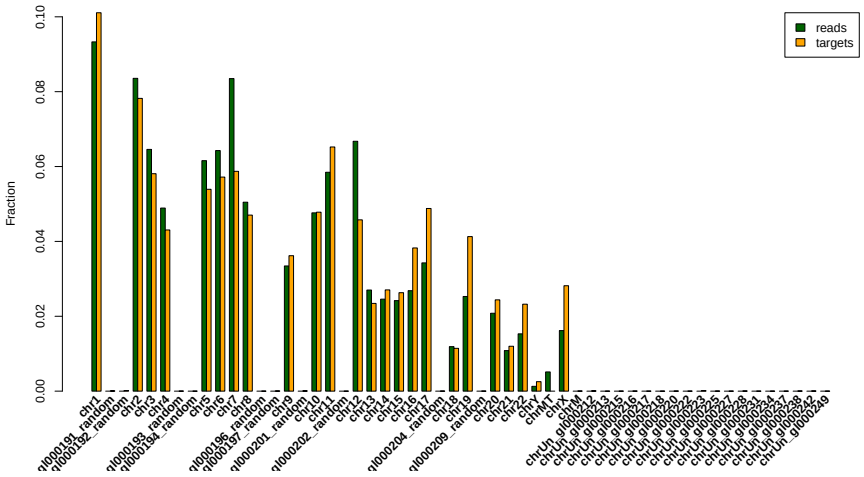
Figure 2.2: `Trimmomatic` results on R1 reads (forward reads of PE reads) of normal 5222: (a) is prior and (b) is post trimming. The red line is the median value. The blue line is the mean. The yellow boxes are the interquantile ranges (25-75%)



(a) Insert-size histogram



(b) Coverage histogram



(c) Read pairs and targets per chromosome barplot

Figure 2.3: Some TEQC plots for adenoma sample number 5225

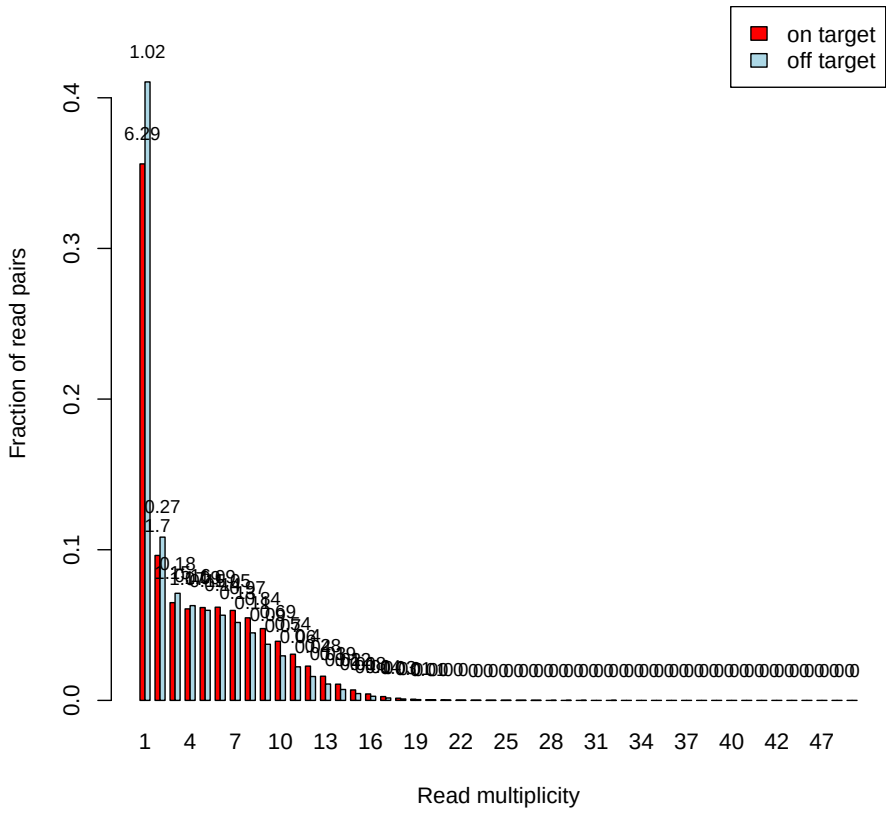


Figure 2.4: TEQC figure on adenoma 5225

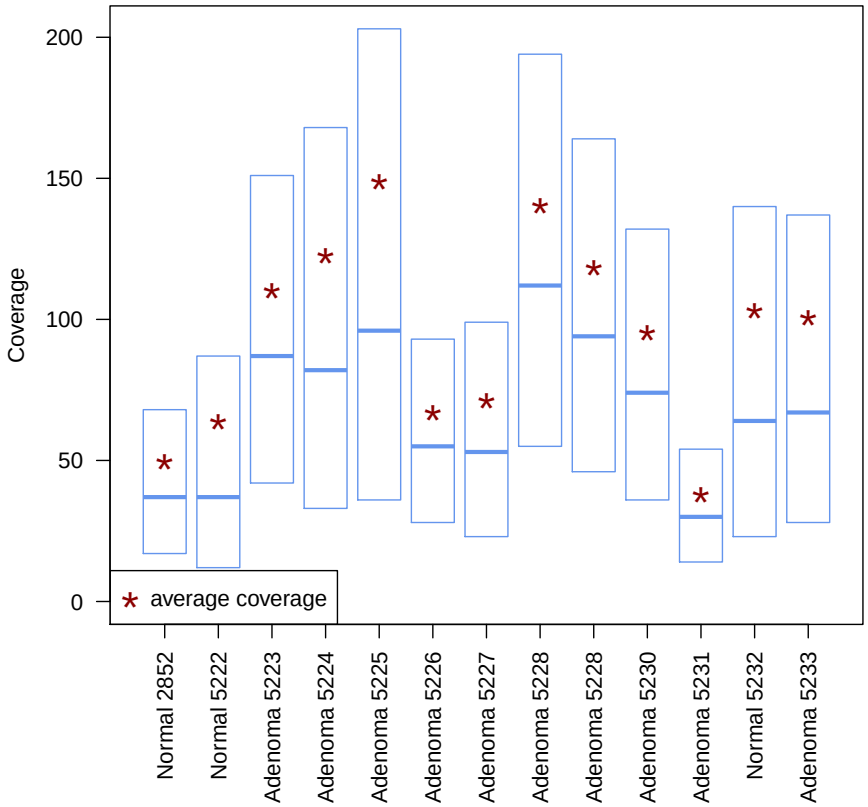


Figure 2.5: Coverage Across Samples

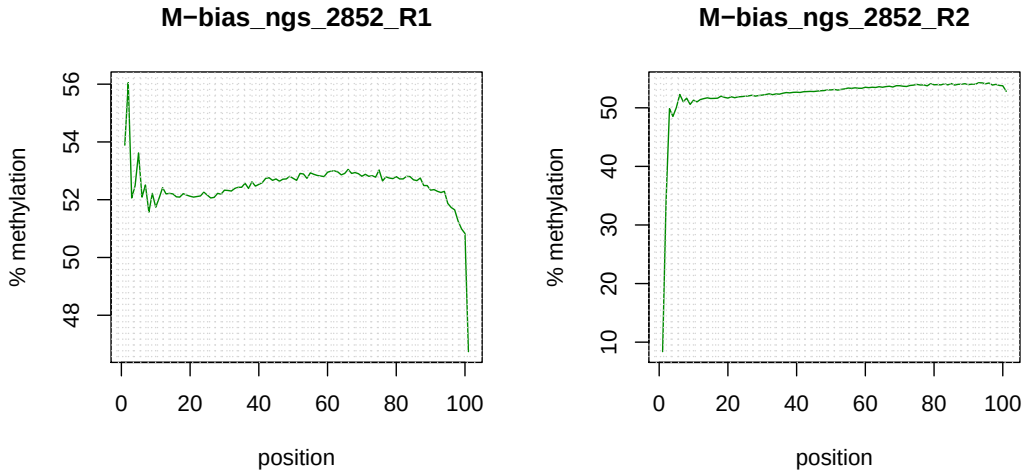


Figure 2.6: M-bias for normal sample 2852: the left shows R1 read and the right the R2 read

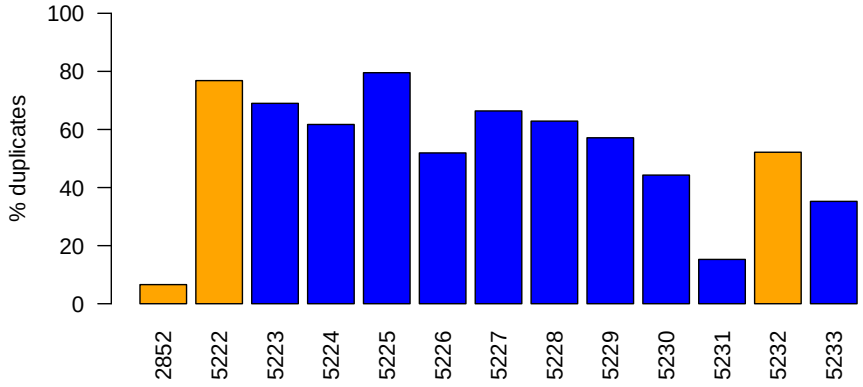


Figure 2.7: Percentage duplicates removed across the samples. The x-axis shows the sample numbers. The orange bars correspond to the normal samples, and the blue ones to the adenoma samples.

Chapter 3

Methods

In this chapter we summarize the spectrum of methods that were applied to our data. These range from mapping the BS-seq reads to the genome using `bismark` – the alignment tool, and go on with exploring the tools that were used to look for ASM regions and differentially methylated regions. We looked at the regions that overlap between the tools and searched for ASM regions that were lost from the normal to the adenoma state. We also explain a method of our own that was developed to evaluate some of these regions that the tools predicted to show ASM.

3.1 Bismark

`Bismark` [20] was used to map the reads to the human genome hg19. Box 3.1 shows the commands that were used for a single sample. The main command used in `bismark` was the actual mapping of the BS-seq reads to hg19. The output was a SAM [21] (or BAM) file of the aligned reads. `Bismark` then used the SAM file to generate coverage and percentage methylation information on the individual C sites (in a CpG context). `Bismark` was also used for deduplication.

```
# Align reads
$ bismark --bowtie2 -p 4 -o NGS-5222_bismark_bt2_pe.sam /home/Shared/data/
    annotation/_Archive/Human/genome/GRCH37 -1 NGS-5222_R1_t20l20_paired.fastq.
    gz -2 NGS-5222_R2_t20l20_paired.fastq.gz

# Remove duplicates
$ deduplicate_bismark  -p NGS-5222_bismark_bt2_pe.sam.gz

# Run methylation extractor
$ bismark_methylation_extractor -p --comprehensive NGS-5222_bismark_bt2_pe.
    deduplicated.sam

# Run bisamrk_to_bedGraph
$ bismark2bedGraph --counts -o  CpG_context_NGS-5222_bismark_bt2_pe.
    deduplicated.bedGraph  CpG_context_NGS-5222_bismark_bt2_pe.deduplicated.txt
```

Box 3.1: `Bismark` commands with *NGS 5222* as an example

3.2 AmrFinder

Methpipe's `amrFinder` [6] was used to get a list of allelically methylated regions (AMRs).

Box 3.2 displays the commands that were used from methpipe [7].

```
# convert SAM files to mr files
$ to-mr -m bismark -o NGS-5222_paired.fastq.gz_bismark_bt2_pe.mr NGS-5222_
    bismark_bt2_pe.sam.gz -v

# sort mr files
$ export LC_ALL=C; sort -k1,1 -k2,2n -k3,3n -k6,6 NGS-5222_paired.fastq.gz_
    bismark_bt2_pe.mr | grep -a "^chr" > NGS-5222_paired.fastq.gz_bismark_bt2_
    pe.mr.sorted

# remove duplicates
$ duplicate-remover -S NGS-5222_paired.fastq.gz_bismark_bt2_pe_dremove_stat.txt
     -o NGS-5222_paired.fastq.gz_bismark_bt2_pe.mr.sorted.dremove NGS-5222_
    paired.fastq.gz_bismark_bt2_pe.mr.sorted

# convert to epiread files
$ methstates -c /home/Shared/data/seq/bisulphite_mirco/FASTQ/genome -o NGS-5222
    _paired.fastq.gz_bismark_bt2_pe.epiread NGS-5222_paired.fastq.gz_bismark_
    bt2_pe.mr.sorted.dremove

# use amrFinder
$ amrfinder -o NGS-5222_paired.fastq.gz_bismark_bt2_pe.amr -c /home/Shared/data
    /seq/bisulphite_mirco/FASTQ/genome NGS-5222_paired.fastq.gz_bismark_bt2_pe.
    epiread

# use allelicmeth
$ allelicmeth -c /home/Shared/data/seq/bisulphite_mirco/FASTQ/genome -o NGS
    -5222_paired.fastq.gz_bismark_bt2_pe.allelicmeth NGS-5222_paired.fastq.gz_
    bismark_bt2_pe.epiread -v
```

Box 3.2: `amrFinder` on NGS 5222 sample as an example: our SAM files had to be converted to epiread files (a different format containing the same information) to run `amrFinder` on our data. The commands used are from the methpipe manual [7].

3.2.1 AmrFinder Methodology

`AmrFinder` works by going through the aligned reads on a sliding window of a certain size which is defined as a number of consecutive CpG sites. The tool looks at the methylation status of the Cytosines in the window. To decide if the region is an AMR, two models are fitted: a ”site-specific” and an ”allele-specific” model. The highest scoring model determines if the methylation is allele-specific (a putative AMR) or not.

The site-specific methylation model (one-allele model) in a single allele is defined as

$$\Theta = (\theta_1, ..., \theta_n),$$

where θ_i is the probability that the Cytosine is methylated at position i. The *likelihood* of the model becomes:

$$L_1(\Theta|R) = Pr(R|\Theta) \propto \prod_{i=1}^{n} \theta_i^{m(R,i)} (1-\theta_i)^{u(R,i)}, \tag{3.1}$$

where n is the number of CpGs in the genomic interval, R is the set of reads, and $m(R,i)$ and $u(R,i)$ are the number of methylated and unmethylated observations from reads mapping onto the i^{th} interval, respectively. Estimates for θ_i are obtained assuming a binomial distribution for methylation states $m(R,i)$ [6].

The allele-specific methylation model (two-allele model) is presented as follows:

$$\Theta = ((\theta_{11,12}), ..., (\theta_{n1}, \theta_{n2}),$$

where θ_{i1} and θ_{i2} are the methylation probabilities at position i on allele 1 and allele 2, respectively. The *likelihood* of the model is:

$$L_2(\Theta|R,\gamma) = \begin{pmatrix} |R| \\ |\gamma_1| \end{pmatrix} 0.5^{|R|} \prod_{i=1}^{n} \prod_{j=1}^{2} \theta_{ij}^{m(\gamma_j,i)} (1-\theta_{ij})^{u(\gamma_j,i)}, \tag{3.2}$$

where $\gamma = \{\gamma_1, \gamma_2\}$ represents the allele of origin (allele 1 or allele 2). The probability that a read r originates from an allele is 0.5 since this is a diploid organism. This model is fitted using the Expectation Maximization (EM) algorithm [6].

3.2.2 Our amrFinder Conditions

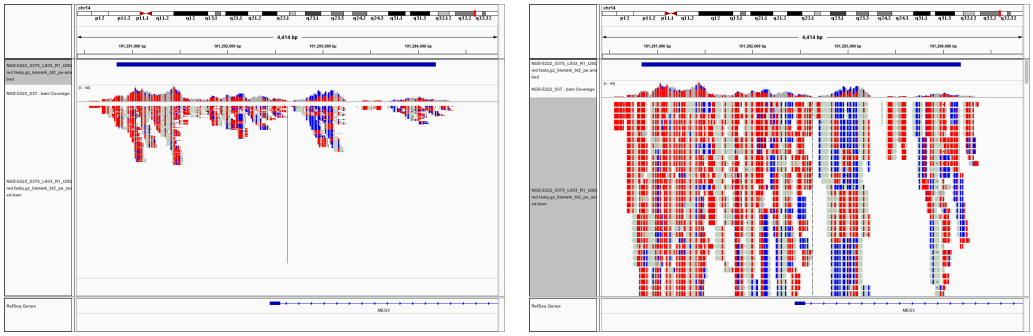
We used the default options of `amrFinder`. This means that a sliding window size of 10 CpGs was used, and the minimum coverage per site was 4.

Two resulting AMRs that are a specific maximum distance apart are merged as one, and this default maximum is 1000 base pairs. At the end, the AMRs whose size is less than this gap size are also eliminated [7].

Figure 3.1 shows an example AMR that `amrFinder` had predicted for sample 5222 (normal sample), as seen in IGV. This is a known imprinted gene called MEG3. The blue block is the region that was predicted to show allele-specific methylation. We added the BAM file of sample 5222 (normal crypt) that shows the individual reads that map to that region. The blue indices indicate an A base and the red ones a G base. So the blue positions indicate Cytosines that had been unmethylated and were converted to Uracils after bisulfite treatment and then to Thymines after PCR amplification. The Adenines are thus complementary to unmethylated Cytosines whereas the Guanines are complementary to methylated Cytosines, as these positions were not transformed during bisulfite treatment.

3.2.3 Allelicmeth

`Allelicmeth` is another tool in Methpipe [7] that gives us information on the AMRs. Instead of looking at whole regions, however, it spits out a p-value for allele-specific methylation (ASM) for every two consecutive CpG sites. This will serve as a means of comparing a scoring function of our own which is further explained in the Results and Discussion chapter.



(a) MEG 3 AMR and aligned reads (b) Zoom in on AMR

Figure 3.1: MEG3 predicted as an AMR for normal sample 5222 – seen in IGV [32]

3.3 Bsseq

`Bsseq` was used to call on differentially methylated regions (DMRs) [9]. It allows for coverages as low as 4x with the methods it applies and makes use of the replicates to account for DNA methylations that can be individual specific.

`Bismark` generated coverage files that consist of single CpGs in each row, and information on the location, methylation percentage, and number of methylated and unmethylated reads as the columns (for each CpG site). The `bsseq` package takes this data as input and applies a smoothing algorithm called `BSmooth`. `BSmooth` estimates methylation levels for a single sample and applies local averaging to improve precision.

Per CpG site, the smoothing function looks at the proportion of methylated reads M_j/N_j, where M_j is the number of methylated reads at the j^{th} CpG and N_j is the total number of reads at that location. It is assumed that M_j follows a binomial distribution with success probability π_j. M_j/N_j is an unbiased estimate of π_j.

The function makes use of the fact that methylation levels are strongly correlated across the genome: for example between neighboring CpG islands and shores, and thus assumes that π_j varies smoothly across the genome. Local likelihood smoothing is done to improve precision. The smoothness of the estimated profile depends on the genomic CpG density.

`BSmooth` also takes biological variation of replicates into account and looks for regions that show consistent differences. `BSmooth` thus detects DMRs by calculating a signal to noise statistic, similar to the statistic used in the t-test, to evaluate these consistent differences. There is a certain cut-off for this statistic [9]. Because biological variability is much greater in cancer samples we only used the variability between the normal samples as an estimate of the variance.

3.4 Methtuple

`Methtuple` [10] was used to look at consecutive CpG pairs. We thus get the following information for each CpG tuple:

1. MM reads: the number of reads where both Cs are methylated.
2. MU reads: the number of reads where the first C is methylated and the following one unmethylated.
3. UM reads: the number of reads where the first C is unmethylated and the following one methylated.
4. MM reads: the number of reads where both Cs are unmethylated.

These numbers were used in the scoring function that we made to assess the levels of methylation.

3.5 Our Scoring Function

3.5.1 Reason Behind the Function

`AmrFinder` gave us a multitide of regions. This is further discussed in section 4.1. We came up with a scoring function to not only limit the number of AMRs, but to judge how well `amrFinder` works and whether or not the regions it gave us do indeed show allele-specific methylation. Furthermore, a continuous score will give us the ability to look for changes in allele-specificity.

3.5.2 Filtering Conditions

For simplicity, the score was calculated for each genomic tuple that was produced by `methtuple` [10]. Some filtering was done on the tuples:

Coverage: A minimum coverage of 10 was set to each tuple. Any tuple covered by less than 10 total reads was discarded.

Tuple Distance: A maximum distance of 150 bp between the CpG sites in a tuple was set. This means that if in a tuple, the distance between the CpGs was bigger than 150 bp, the tuple was discarded.

Tuple Uniqueness: For each sample, the case of having the same first CpG site in multiple tuples can happen. For example, we can have the following tuples: {a,b}, {a,c}, {a,d}. They are all unique, but have the same start site. This can happen with the reads covering different portions of a region, since we have PE reads: the forward and reverse reads can have different gaps and extents of alignment. Figure 3.2 illustrates this. To solve this problem, we only consider the smallest tuple size – the tuples where the Cytosines are the closest together. We sorted the tuples by chromosome, position1, and position2, and only kept the tuples that were unique by chromosome and position1.

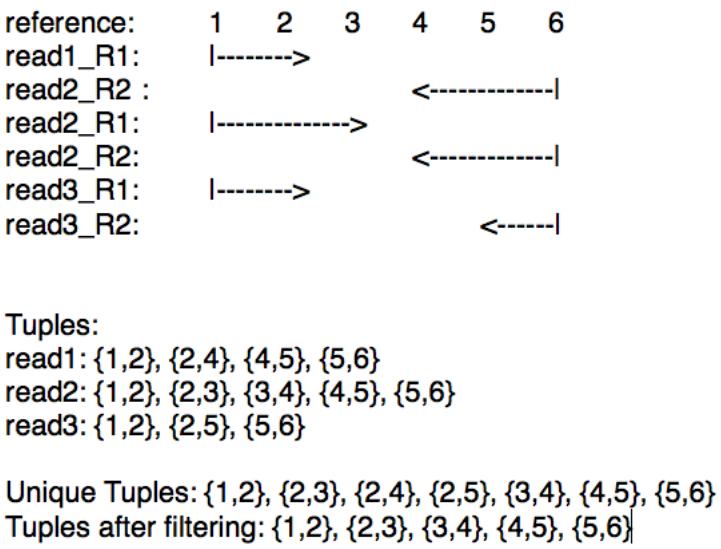


Figure 3.2: Filtering Tuples

After these filtering steps, a value of 1 was added to each of MM, MU, UM, UU, and total coverage. This step was necessary for the log odds ratio that was used in the scoring function.

3.5.3 The Score

The score was calculated for each genomic tuple that was produced by `methtuple` [10]. We used the log odds ratio as follows:

$$score = log\left\{\frac{MM * UU}{MU * UM}\right\} \tag{3.3}$$

To avoid situations where the denominator is zero, we added a value of 1 to every cell (to the MM count, the UU count, the MU count, and the UM count). Figure 3.3 depicts what we envisioned with this scoring method. In a situation of having ASM (the top situation in the figure), we would expect the majority of tuples to be a somewhat even mixture of the MMs and UUs, rather than a random mix of UM and MU, fully MM or fully UU.

3.5.4 Weighting the Score

The log odds ratio test in itself is not enough to look at methylation that is allele-specific. Having fully methylated or fully unmethylated regions will also give high scores. The ratio does not account for differences between MM and UU. It does indicate when we have a mixture of UMs and MUs which is insightful if this coincides

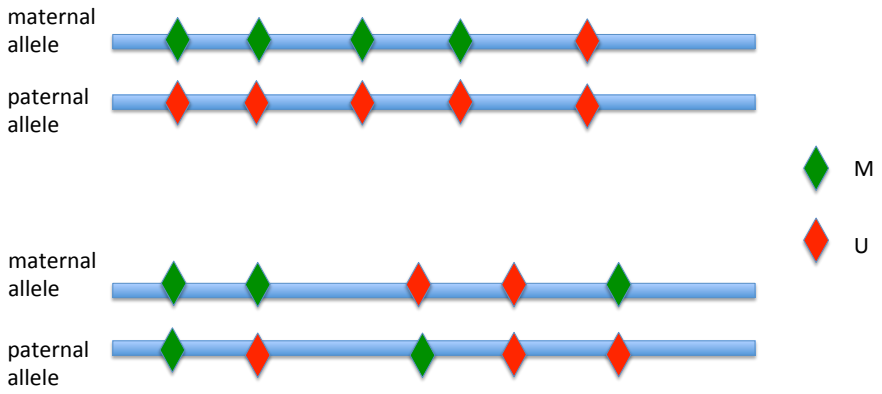


Figure 3.3: Score rationale: the top case is true ASM, whilst the bottom has localized regions that might be called ASM but are not consistent across the broader region.

with a region that `amrFinder` has given us, indicating that the region does not show true allele-specific methylation despite having an overall methylation level of around 50% in the region for example. To extend the functionality of the score to also give us an insight into true allele-specific methylation we added a weight to the odds ratios as follows.

We tested for the equal mixing of MM vs UU for each tuple. The null hypothesis is that the proportions are equal (when we have an ASM). The alternative hypothesis is that the proportions are not equal. We used *prop.test* in *R* for this (two tailed), and the corresponding *p-values* as weights. So the weight was a value between 0 and 1 that captures the statistical evidence of departures from 50% M and 50% U. The weight is exactly 1 when MM equals UU. The bigger the imbalance between the MM and UU counts, the bigger the penalty and the lower the score. The modified score is as follows:

$$score = log\left\{\frac{MM * UU}{MU * UM}\right\} * weight \tag{3.4}$$

Chapter 4

Results and Discussion

This Chapter discusses the results that were obtained with the tools presented in the Methods section and shows the comparisons and analyses that were done accordingly. We assessed the regions that had been predicted to be allele-specifically methylated by the tools and overlapped some results. We found regions that had lost the property of being ASMs in the adenoma stage in all samples, indicating that these regions, which were coding as well as non-coding regions, may have to be further investigated. We assessed the performance of our developed scoring method and compared it to that of `allelicmeth`, using the X chromosome as a measure or indicator of ASM. The effect that duplicate reads have on our analyses was also investigated.

4.1 AMRs

`AmrFinder` gave out a lot of AMRs per sample. The number of AMRs generated per sample ranged from 4,819 regions to 41,253 regions. Figure 4.1 shows the number of AMRs that were predicted for each sample.

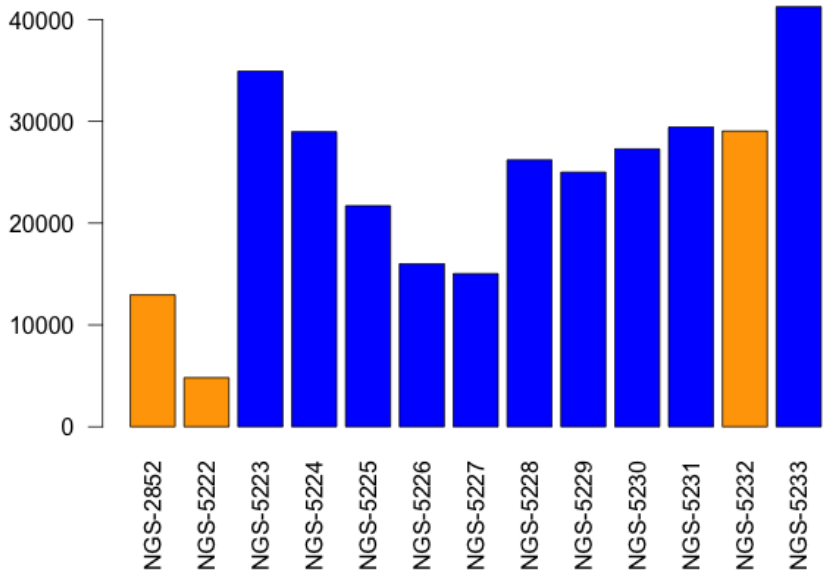


Figure 4.1: Number of AMRs generated by `amrFinder` by sample. The orange bars represent the normal crypts and the blue ones the adenoma samples.

Since we are interested in loss of imprinting from the normal to the adenoma condition, looking at the AMRs that were common to the three normal samples and then absent in the adenoma samples was the next step.

We found 61 AMRs that were present in the three normal samples and then completely lost in all 10 adenoma samples. Note that some of these regions overlap because we combined the AMRs from the three normal samples and kept the unique ones. Table 4.1 shows two of those regions that were found and the genes they overlap with. Most of them fell on the promoter regions of those genes or in non-coding regions, outside the genes. Box 4.1 shows the code that generated these regions.

CFAP58 is described as a protein binding protein in the extracellular matrix region. A question worth asking may be whether the proteins associating the cells to the extracellular region are less functional or change in function in a transition to the cancerous stage.

Figure 4.2 shows some example regions where we see the three normal samples' predicted AMRs and the methylation levels across all samples.

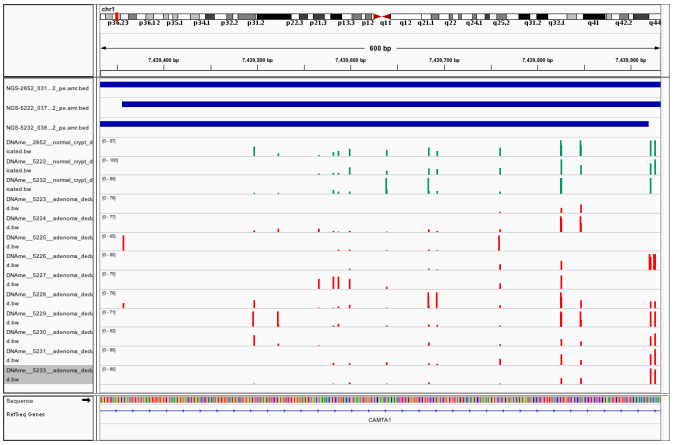
```
library(GenomicRanges)
library(data.table)

# Our file (BED files of the AMRs) paths are stored in the "files" vector
samples <- lapply(files, fread)

# keep unique AMRs in normals (some may overlap each other)
amr <- rbind(samples[[1]], samples[[2]], samples[[12]])
amr <- unique(amr)
colnames(amr) <- c("chr", "start", "end")
amr.gr <- makeGRangesFromDataFrame(amr)

# keep track on the overlap of the unique normal AMRs in all samples
for (i in 1:13) {
  s <- samples[[i]]
  colnames(s) <- c("chr", "start", "end")
  s.gr <- makeGRangesFromDataFrame(s)
  # overlap unique AMRs with those in s.gr
  count <- countOverlaps(amr.gr, s.gr)
  count <- replace(count, count>0, 1)
  mcols(amr.gr)[[nm[i]]] <- count
}

# sum of normals
mcols(amr.gr)[["normals"]] <- amr.gr$"2852" + amr.gr$"5222" + amr.gr$"5232"

# sum of adenomas
mcols(amr.gr)[["adenomas"]] <- amr.gr$"5223" + amr.gr$"5224" + amr.gr$"5225" +
    amr.gr$"5226" + amr.gr$"5227" +
  amr.gr$"5228" + amr.gr$"5229" + amr.gr$"5230" + amr.gr$"5231" + amr.gr$"5233"

# AMRs present in all normals and absent in all adenoma (normals=3, adenomas=0)

w <- which(amr.gr$normals==3 & amr.gr$adenomas==0)
amr.gr_3_0 <- amr.gr[w,]
```

Box 4.1: Generating the lost AMRs

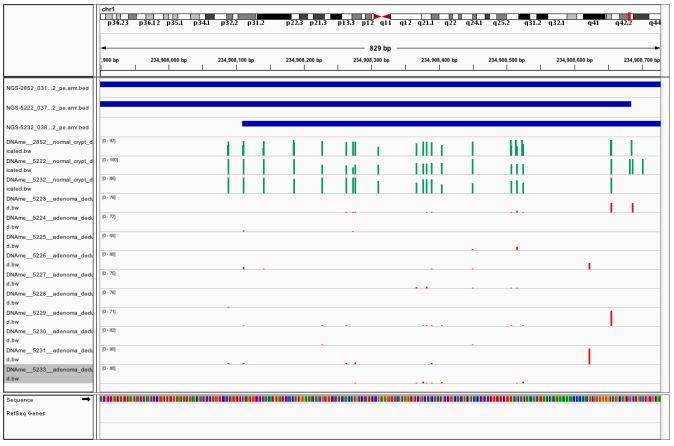
Table 4.1: Lost AMRs

chr	start	end	gene	full name	id in Ensembl
chr10	106,200,601	106,202,271	CFAP58	cilia and flagella associated protein 58	ENSG00000120051
chr1	7439355	7439955	CAMTA1	calmodulin binding transcription activator 1	ENSG00000171735

The adenoma stage is a pre-cancerous one. There was a lot of variation in the AMRs themselves between the adenoma samples. Some AMRs were detected in a considerable amount of adenomas and were lacking in others. It may be that the loss of imprinting was progressive, and only some of the samples had lost imprinting. It could also be the case that some AMRs are lost in some but not all of the cancers anyway, even at the final stages.



(a) Loss of ASM at the CAMTA1 gene



(b) Loss of ASM at a non-coding region

(c) Loss of ASM at the RUNX1 gene

Figure 4.2: Loss of ASM. The green methylation patterns are those of the normal samples, and they indicate a somewhat 50% methylation. The red patterns are those of the adenoma samples and we see a loss of methylation in most samples.

4.2 DMRs

Regions that showed differential methylation across the normal vs the adenoma samples were detected with bioconductor's `bsseq` package. The method accounts for the variance between the normal samples (biological variance).

Box 4.2 shows the code that was used to generate the DMRs. We set our conditions as follows for a CpG site to be considered in the process of generating DMRs: at least 3 adenoma samples (out of the available 10) and at least 2 normal samples (out of the available 3) had to have a coverage of at least 2 at this particular site.

```
library(bsseq)

# read in the meta file with all sample details
meta <- read.csv("meta_file.csv", header=TRUE)
rownames(meta) <- meta\$QUML_sampleid

# "sample_files" has the names of the coverage files generated by bismark
# "samples" has the names of the samples (they match the rownames of "meta")
# read in the coverage files
data <- read.bismark(sample_files, samples, verbose=TRUE)

# Smooth the data
data.fit <- BSmooth(data, mc.cores=13, verbose=T)

# get the coverage
data.cov <- getCoverage(data.fit)

# We take a look at the average coverage per sample
colMeans(data.cov)
[1] 4.303247 1.435212 3.936041 4.801395 2.740904 3.919605 2.462732 5.794801
[9] 5.553686 5.425147 3.451414 4.567880 6.352977

# set our conditions
keepLoci <- which(rowSums(data.cov[, meta\$condition=="adenoma"] >=2) >=3 &
    rowSums(data.cov[, meta\$condition=="normal_crypt", drop=F] >=2) >=2)
data.loci <- data.fit[keepLoci,]

# Estimate variance from normal crypts
data.tstat <- BSmooth.tstat(data.loci, group1=which(meta\$condition=="adenoma")
    , group2=which(meta\$condition=="normal_crypt"), estimate.var="group2",
    local.correct=T, verbose=T)

# DMRs
dmrs_quantil<-dmrFinder(data.tstat, qcutoff=c(0.005,0.995))
dmrs_subset<-subset(dmrs_quantil, n>=3 &abs(meanDiff) >= 0.1)
nrow(dmrs_subset)
[1] 2178

# plot top 200 DMRs
pdf(file="top200_DMRs.pdf", width=10, height=5)
plotManyRegions(data.loci, dmrs_subset[1:200,], extend=5000, addRegions=dmrs_
    quantil)
dev.off()
```

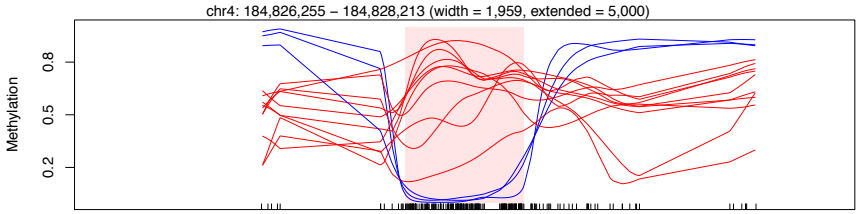
Box 4.2: Generating the DMRs

A total of 2,178 DMRs were produced. 2,694,588 CpG sites were covered completely by all 13 samples. The rest had coverage in only some of the samples. 3,680,703 loci were kept with our mentioned conditions as stated previously. Figure 4.3 shows the top three DMRs that were found.

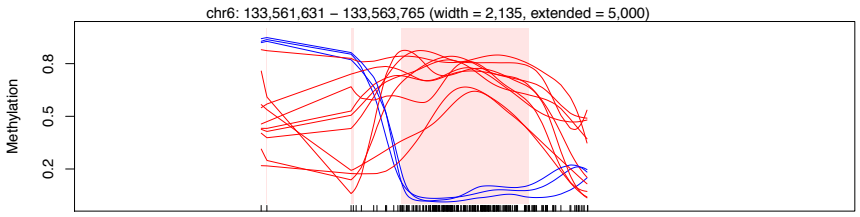
Table 4.2 summarizes the DMR results from trying a range of constraints and conditions. Even with very stringent conditions of a coverage of at least 30 reads per CpG site in at least 8 adenoma samples and 2 normal samples, `bsseq` generated 596 regions it considered to be differentially methylated.

Table 4.2: Various conditions for generating DMRs

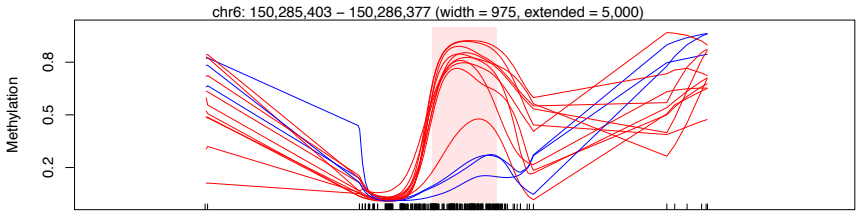
minimum normal coverage	minimum number of normal samples	minimum adenoma coverage	minimum number of adenoma samples	number of DMRs	number of CpG sites
2	2	2	3	2,178	3,680,703
4	2	4	5	1,865	2,896,081
10	2	10	5	1,498	1,894,073
20	2	20	5	1,142	1,124,307
20	2	20	8	975	941,856
10	3	10	5	906	888,411
10	3	10	8	889	875,294
30	2	30	8	596	541,806
20	3	20	8	596	341,168



(a) Highest ranked DMR



(b) Second ranked DMR



(c) Third ranked DMR

Figure 4.3: Top DMRs: the blue lines are the normals and the red lines the adenoma samples. The conditions for these DMRs were as follows: a coverage of at least 2 in at least 3 adenoma samples and in at least 2 normal samples

4.3 Overlapping Regions

We looked for regions that overlap between the AMRs and the DMRs. It is worth noting that DMRs are not necessarily specific to a situation where we have allele-specific methylation that is lost or gained (which is what we are interested in). DMRs also include regions that may have gone from a state of being completely unmethylated to being fully methylated for example. `AmrFinder` on the other hand predicted regions that it saw as allele-specifically methylated. It did so separately for each sample, whereas `bsseq` used all samples at once to call the DMRs.

Figure 4.4 shows the number of DMRs and AMRs for 3 of the 13 samples, and only for chromosome 14. No overlapping step between any of these regions was done. These are the total amounts of regions for this chromosome. We can see a clear difference in the number of regions that `amrFinder` and `bsseq` produce.

From a shear number perspective, `amrFinder` gave us a lot of regions it considered to show allele-specific methylation, so it was interesting to look at some of the overlaps with the DMRs from `bsseq`. It is expected for there to be less DMRs than AMRs since `bsseq` looks at the CpG positions with the conditions that were mentioned before across all 13 samples, applying smoothing techniques, weighting by coverage and looking at biological variation within a condition (for example the normal condition). `AmrFinder` works individually on every sample. This allows for more ASM regions to be detected if you consider individual specific methylation as one explanation, as well as differences in methylation that happened in one sample and not the other by chance.

Zhang et al [35] looked at individual specific DNA methylation in humans that is outside of imprinted loci. Looking at the leukocytes of healthy individuals they uncovered a number of DMRs that do not coincide with imprinted regions. Variability in DNA methylation was strongly influenced by the genetic differences of the individuals. Amplicons that showed intermediate methylation levels (as opposed to fully methylated or unmethylated ones) varied the most amongst the individuals. Their analysis further indicates that "allele-specific methylation is likely to affect about 10% of all human genes and to contribute to allele-specific expression and monoallelic gene silencing."

The DMRs that were taken to do the overlapping analysis with the AMRs were those generated by the conditions listed in the first row of table 4.2.

4.3.1 Lost AMRs vs DMRs

Previously, we found 61 regions that `amrFinder` deamed to be AMR in all normal samples and completely lost in all adenoma samples. Next, we overlapped these regions with the DMRs.

Table 4.3 lists the AMR regions that overlapped with the DMRs. SPATA18 encodes for a protein that mediates the repairing or degradation of unhealthy mitochondria

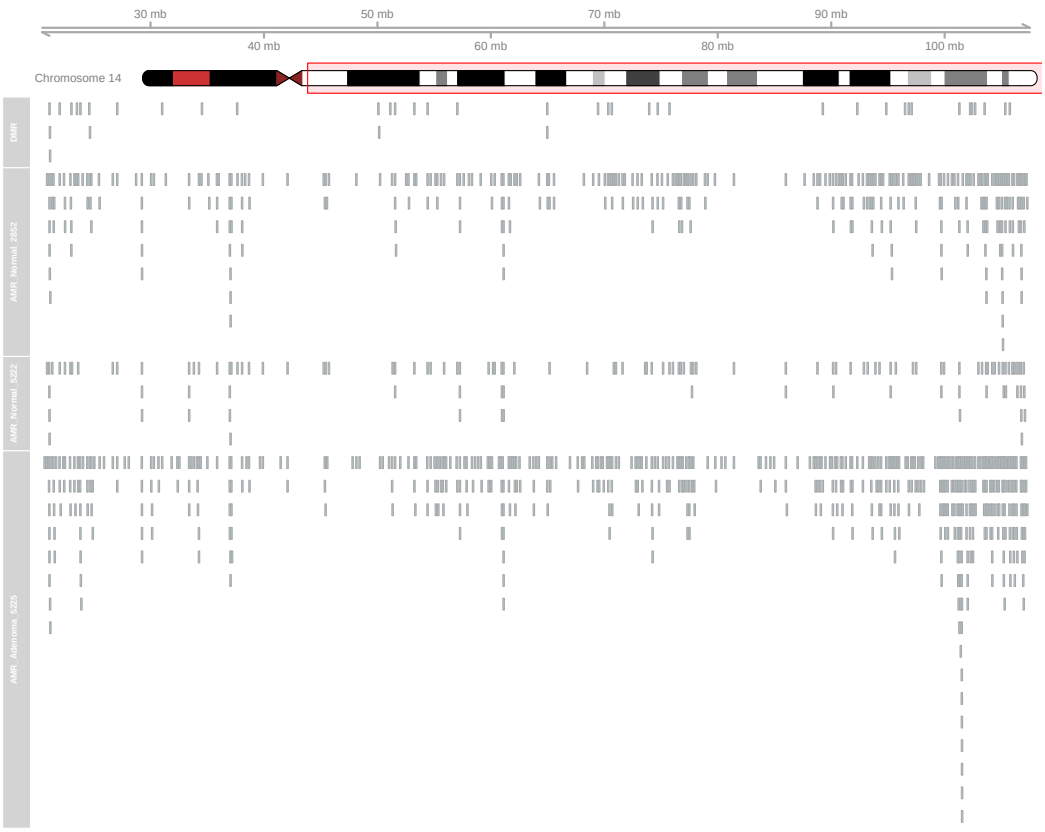


Figure 4.4: AMRs and DMRs for normal 2852, normal 5222 and adenoma 5225 on chromosome 14

in response to mitochondrial damage. It is also involved in mitochondrion degradation of damaged mitochondria by promoting the formation of vacuole-like structures (named MIV), which engulf and degrade unhealthy mitochondria by accumulating lysosomes. The physical interaction of SPATA18/MIEAP, BNIP3 and BNIP3L/NIX at the mitochondrial outer membrane regulates the opening of a pore in the mitochondrial double membrane in order to mediate the translocation of lysosomal proteins from the cytoplasm to the mitochondrial matrix [24].

RUNX1 encodes for proteins that bind to the core site of a number of enhancers and promoters. Chromosomal aberrations involving this gene have been linked to cases of acute leukemia. They are expressed in all tissues except the brain and heart. The highest levels are found in the thymus, bone marrow and peripheral blood [23].

To our knowledge, no association between either of these genes and colon cancer has been reported.

Table 4.3: Lost AMRs overlapping with the DMRs

chr	start	end	gene	full name	id in NCBI
chr1	234907852	234908938	non-coding region	—	—
chr1	234907891	234908684	non-coding region	—	—
chr1	234908109	234908938	non-coding region	—	—
chr4	52918066	52918743	SPATA18	spermatogenesis associated 18	ENSG00000163071
chr17	74260099	74261129	non-coding region	—	—
chr17	74260103	74261057	non-coding region	—	—
chr17	74260293	74261077	non-coding region	—	—
chr21	36419215	36420821	RUNX1	runt-related transcription factor 1	ENSG00000159216
chr21	36419215	36420754	RUNX1	runt-related transcription factor 1	ENSG00000159216

4.4 Our Scoring Function

We developed a scoring function of our own, as explained in section 3.5. Figure 4.5 shows the scoring before and after adding the weight to normal sample 2852. We see that sites that have a more balanced MM to UU ratio score higher (in the center). The colored dots represent known imprinted genes that scored higher, as expected: The orange dots are those of MEG3, the black of NDN, and the pink of RB1.

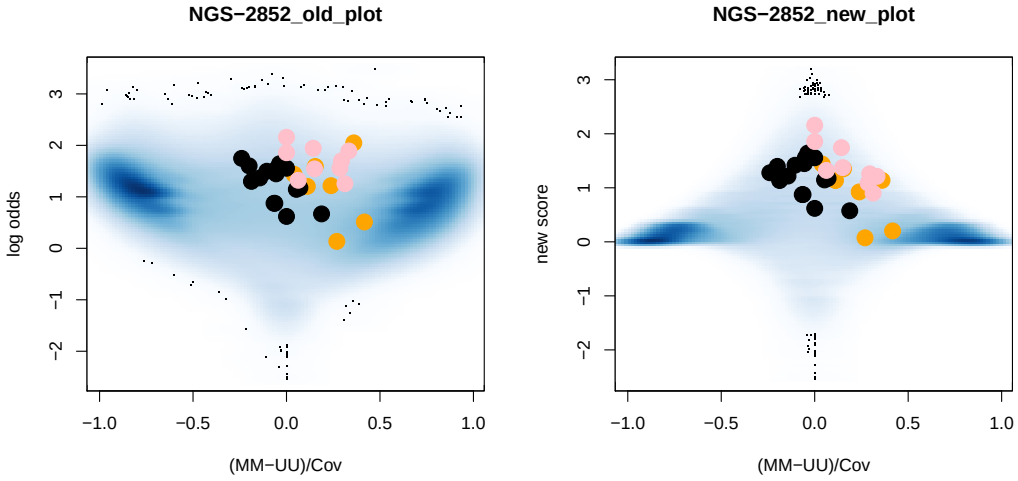
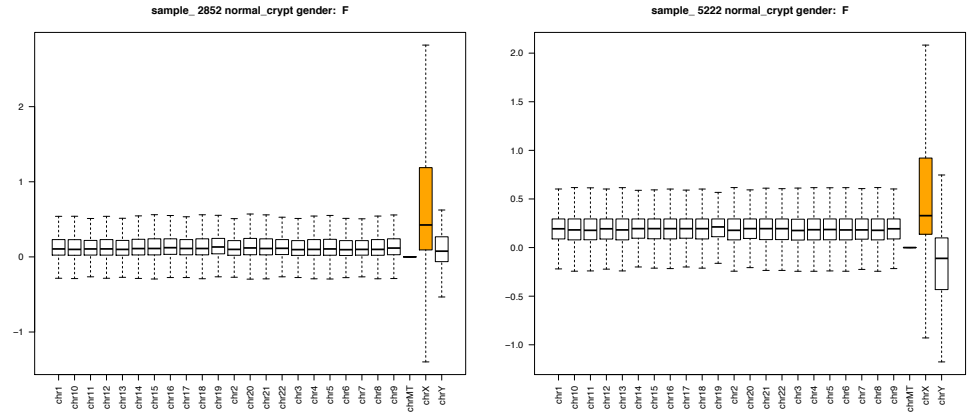
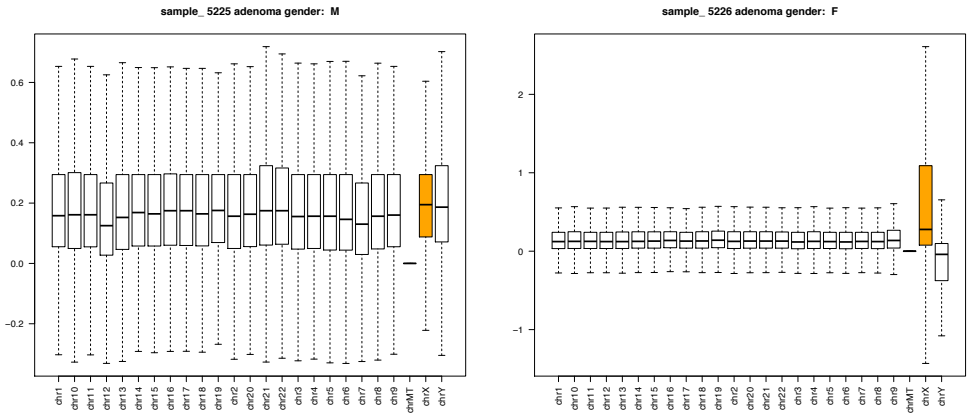


Figure 4.5: Scores with and without the weights for normal sample 2852.

To evaluate the success of the scoring, we looked at how the score performed on a chromosomal basis. In females, we expected the scoring to be higher for chromosome X than all other chromosomes due to the effect of the silenced X chromosome, where only one of the two X chromosomes is expressed (and the other one silenced by methylation). Looking at the boxplots of the score distributions in Figure 4.6, we see that the scores indeed were higher for chromosome X in females, whilst the rest of the chromosomes had a somewhat similar score distribution plots.



(a) Samples 2852 and 5222 (both normal)



(b) Sample 5225 and 5226 (both adenoma)

Figure 4.6: Scores (with weighting) across chromosomes for (a) 2 of the 3 normal samples, all of which were female, and (b) one male and one female from the adenoma samples.

Even without the weighting, we see a difference in the score distributions, with chromosome X scoring higher than the rest, except for the mitochondrial DNA. The score for the mitochondrial chromosome decreased after weighting as there were large regions of unmethylated DNA. Figure 4.7 shows the plots for two of the samples: sample 5229 (a female) and 5230 (a male). The results were similar in the rest of the samples. Looking at the beanplots of the methylation levels, we see what was to be expected. Chromosome X showed more methylation at around 50 % whereas the rest of the chromosomes showed more fully unmethylated or methylated percentages. The mitochondrial DNA showed very low methylation percentages overall. Figure 4.7b shows the scores across the chromosomes without having adjusted with the weights. This explains why the mitochondrial chromosome scores the highest. After the adjustment in

figure 4.7c, the scores worked as they ought to, with chromosome X exhibiting higher scores to allele-specific methylation, and the mitochondrial DNA being an example of a low scoring for fully unmethylated regions.

4.4.1 Allelicmeth

After running `allelicmeth` on each sample, we obtained the following information on each CpG site: the MM counts, UU counts, MU counts, UM counts, total coverage, and the p-value. This p-value was reflective of a score. We did the following order-preserving transformation on these p-values:

$$allelicmeth\ score = -log_{10}(p\_value)$$

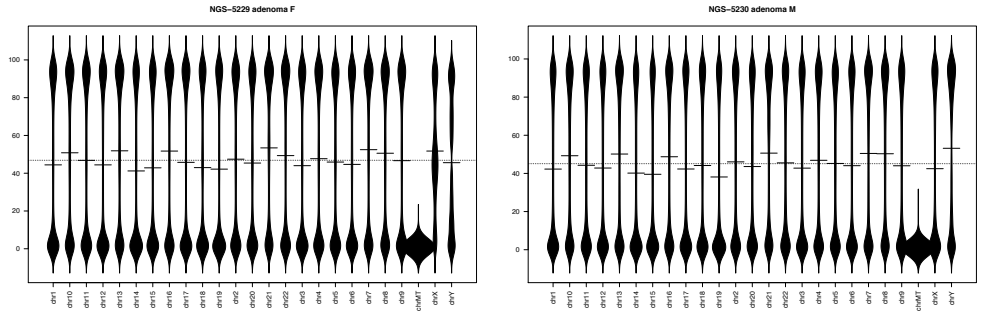
We kept the CpG sites that had a minimum coverage of 10 and looked at the score distributions of `allelicmeth` across the chromosomes. Figure 4.8 shows these distributions for the normal sample number 2852, a female, and adenoma sample number 5225, a male. Figure 4.8b shows the distributions for our scores that were computed using the information on the tuple counts that `allelicmeth` had generated for each CpG. We see that both scores had a higher distribution for chromosome X in sample 2852, as expected, since this sample was that of a female. For the male (sample 5225), the distribution of the scores also met our expectations.

Figure 4.10 shows the plots that depict our scores on the x-axis and those of `allelicmeth` on the y-axis. The red line is the y=x line. We see that at least for the low scores from `allelicmeth`, we also had low scores with our scoring function. Where our score was zero, `allelicmeth` had some quite high scores. This may further illustrate how conservative our scoring function is, and that we might be heavily penalizing regions that are ASM regions because the MM to UU imbalance isn't closer to a 50:50 balance but rather a 60:40 one for example, at high depth. The bigger the imbalance, the higher the penalty with our scoring method. This may be not reflective of a real situation since each sample reflects the data of a multitude of cells that were gathered, each of which is in a different epigenetic state. Moreover, there is a greater variety in the adenoma cells where we can envision that various cells were at different stages of developing into the colorectal cancer or simply had different methylation shifts despite being at similar stages. With this in mind, we can see that demanding a 50:50 read count of MM:UU is too strict, and allowing for a 66:34 imbalance may be just as important.

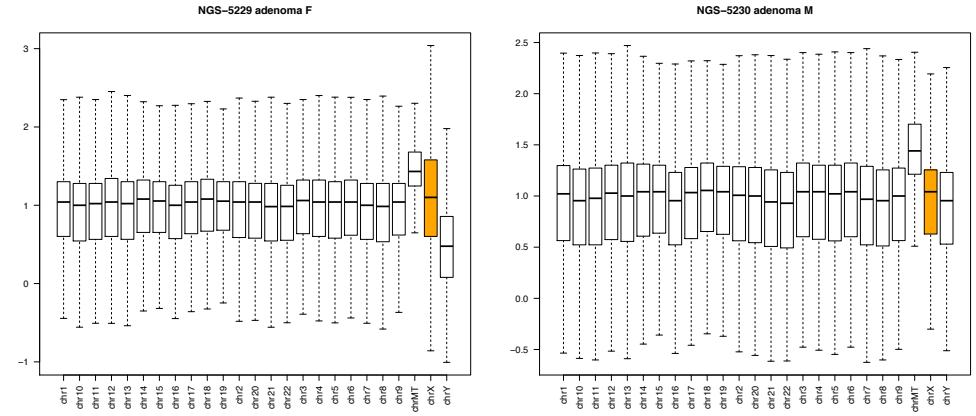
Figure 4.9 reflects the sensitivity vs (1 - specificity) plots for our scores and those produced by `allelicmeth`. We considered the score to be "true" if it was that of the X chromosome. We thus see all the female samples in the figure. 6 out of 7 samples showed better separations with the new score until 10% FDR.

4.5 The Duplicate Effect

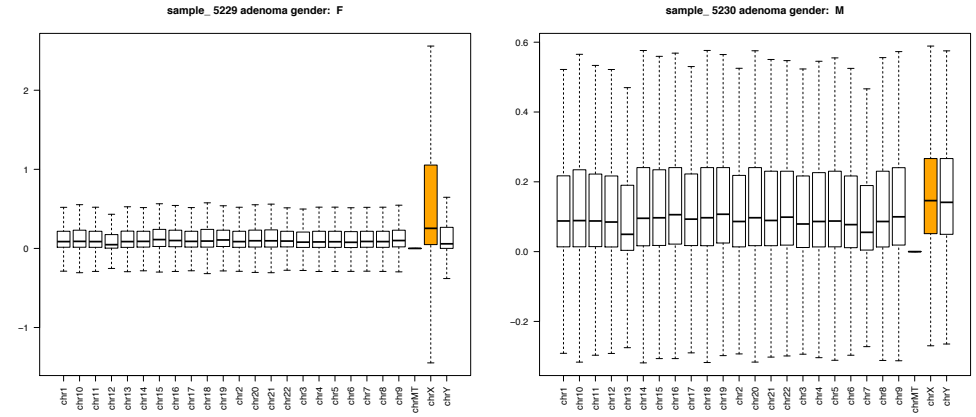
Here we see whether or not the removal of duplicate reads had any effect on our results.



(a) Beanplots for methylation percentages across chromosomes

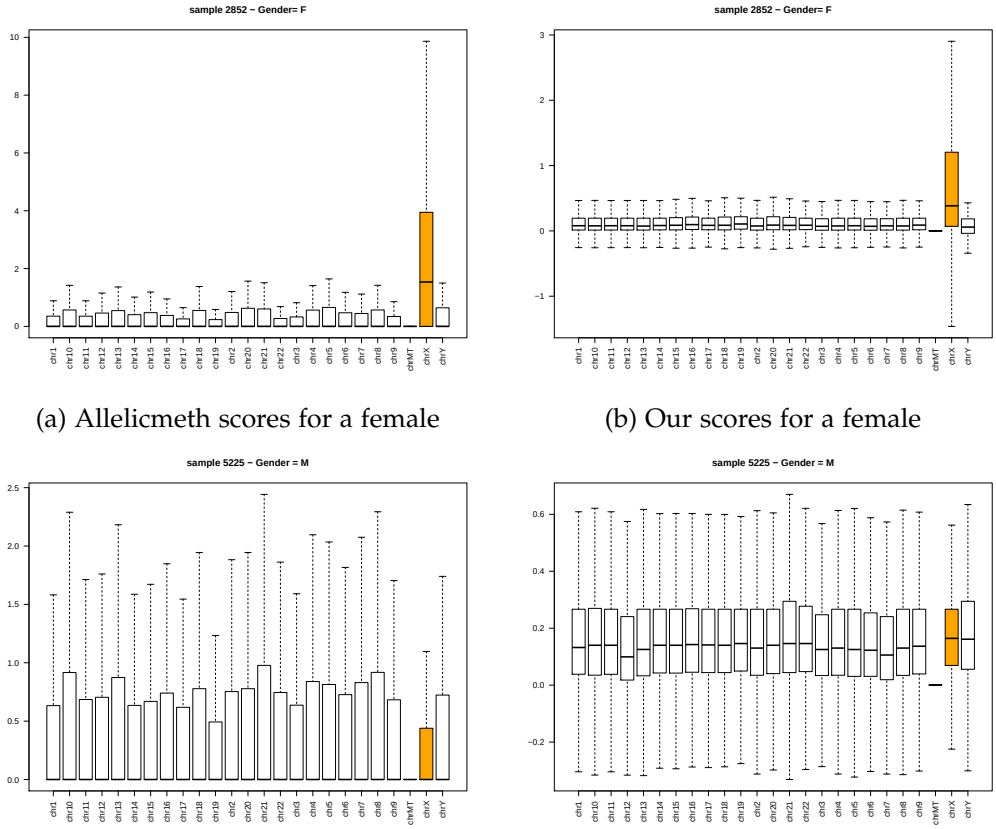


(b) Score distributions without weighting



(c) Score distributions after weighting

Figure 4.7: Adenoma samples 5229 (Female) and 5230 (male)



(a) Allelicmeth scores for a female

(b) Our scores for a female

(c) Allelicmeth scores for a male

(d) Our scores for a male

Figure 4.8: `Allelicmeth` and our scores for normal sample 2852 (female) and adenoma sample 5225 (male), with a minimum coverage of 10.

4.5.1 Effect of GC Content on PCR

Several studies have indicated a link between GC content of reads and the effect this has on their amplification during PCR. Benjamini and Speed [1] showed the effect the GC content of reads had on read coverage. The highest amounts of reads produced were at a GC content of around 50%.

This poses a problem for our BS-seq reads. Our aim lies in detecting regions that show allele-specific methylation. The unmethylated cytosines in the BS-seq reads have been converted to Uracils. We can thus predict a difference in coverage for the reads that retained their GC content because they happened to be methylated there and also had a GC content of around 50%. The coverage thus varies amongst the reads. Figure 4.11 shows the plot that Benjamini and Speed [1] presented. There is a clear amplification in coverage at a moderate GC content.

`Bsseq` gave us the predicted DMRs based on an analysis that took the coverage at every

CpG site into consideration. With the GC content having an effect on the coverage, we can see how there might be some variation in the DMRs that are predicted with and without duplicate removal.

4.5.2 DMRs With and Without Duplicates

We compared the DMRs that were produced with and without duplicate removal. In keeping the duplicate reads, we generated a total of 2,306 regions, as opposed to a total of 2,178 regions without the duplicates. There were 1,788 regions from the 2,178 that overlapped with the regions that had kept the duplicate reads. Overlapping the DMRs generated with duplicates with the other DMRs (generated without duplicates) 1,785 regions overlapped by any number of base pairs (even possibly 1 bp). That means that there were 521 unique regions that were considered to be DMRs when using duplicate reads, that were never considered to be DMRs after removing duplicates, not even partially. However, this does not rule out that they may have simply been below the cutoff when smoothing was done.

`Bsseq` generated the DMRs in order of significance. Next, we looked at some of the ranks of the mentioned DMRs that were unique with the duplicate reads, to see how significantly `bsseq` considered them to be DMRs. If they were considered significant, we might further question the power that duplicate reads had on the DMR predictions and our BS-seq reads.

For reference purposes, I refer to the DMRs produced from reads that kept the duplicates as DMRs from (a), and the DMRs produced from the reads without duplicates as the DMRs from (b).

Looking at the DMRs from (a), we found that 47 of the top 50 and 94 of the top 100 DMRs overlapped with the DMRs from (b). Table 4.4 shows a summary of the results. We see that a significant amount of the top DMRs from (a) overlapped with those from (b), indicating that removing duplicates did not have that much of an effect there. However, the overlapping was based on a minimum overlap of 1 bp.

4.5.3 Methylation Percentages

`Bismark` gave methylation percentages at every CpG site. We compared the methylation percentages before and after duplicated removal.

We filtered out sites that had a coverage of less than 10, since a difference in methylation might have appeared more drastic there. For example, having only one read mapping at a CpG that is methylated gives a methylation percent of a 100. After this filtering, the site from the one methylation profile were mapped to those of the other file, and we plotted the common regions and what methylation levels `bismark` had reported.

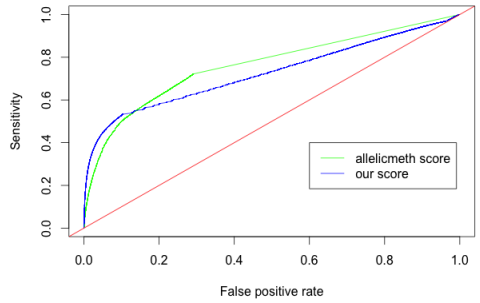
Figure 4.12 depicts some of the comparisons made. We see that figure 4.12a shows the results for sample number 2852 which had the lowest duplicate percentage of around

Table 4.4: The top DMRs when considering duplicates that overlapped with the DMRs that ignored duplicates

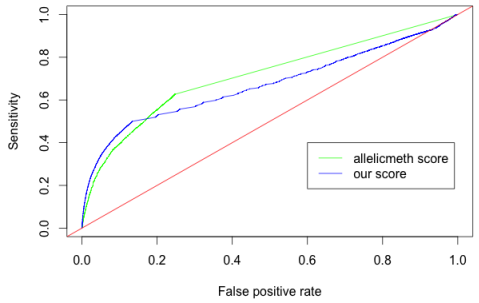
Number of Top DMRs (with duplicate reads)	Number of DMRs that overlapped
50	47
100	94
150	140
200	186
250	231
300	272
400	358
500	443
700	622
1000	887
1500	1296
2306	1785

7%. Figure 4.12b on the other hand is that of sample number 5225 which had the highest duplicate percentage of about 80%.

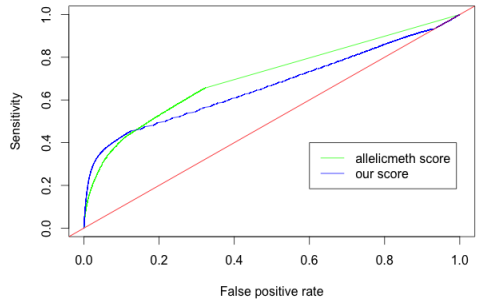
We see that keeping the duplicate reads has consequences on the methylation percentages that `bismark` predicts per CpG site. With the overlapping DMRs, we saw less of a drastic effect. Most of the highest ranking DMRs that had been predicted with the duplicate reads were maintained after duplicate removal. 47 of the first 50 DMRs and 186 of the top 200 were kept.



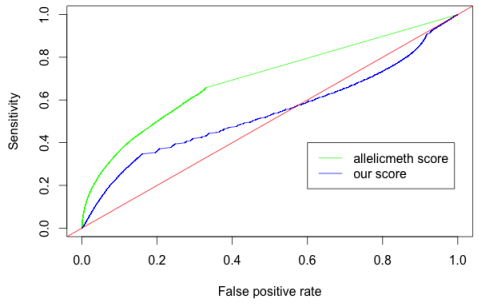
(a) Normal sample 2852



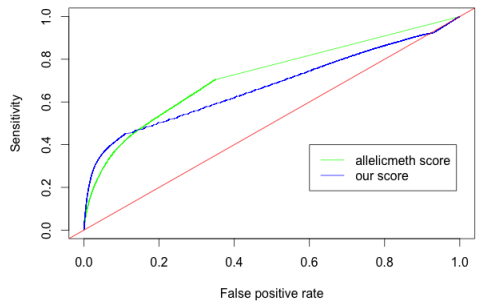
(b) Normal sample 5222



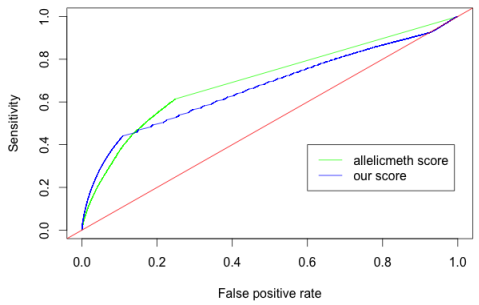
(c) Adenoma sample 5226



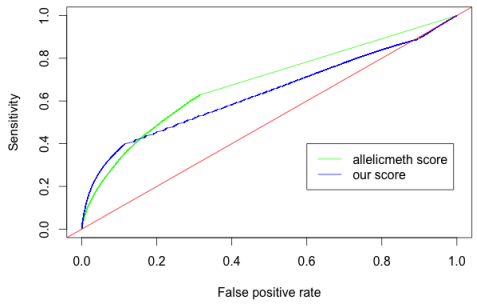
(d) Adenoma sample 5227



(e) Adenoma sample 5229

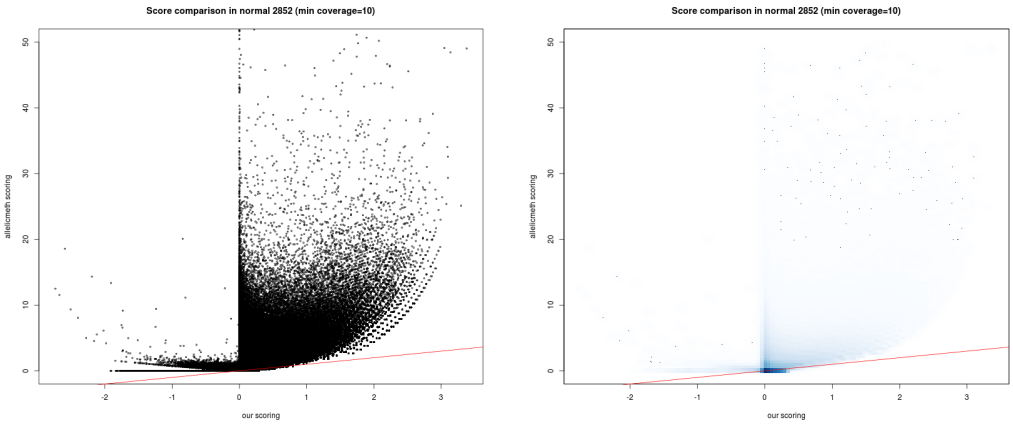


(f) Normal sample 5232



(g) Adenoma sample 5233

Figure 4.9: Sensitivity vs (1-specificity) plots for female samples



(a) Comparison plot (b) Smooth scatter

Figure 4.10: `Allelicmeth` vs our scores for normal sample 2852

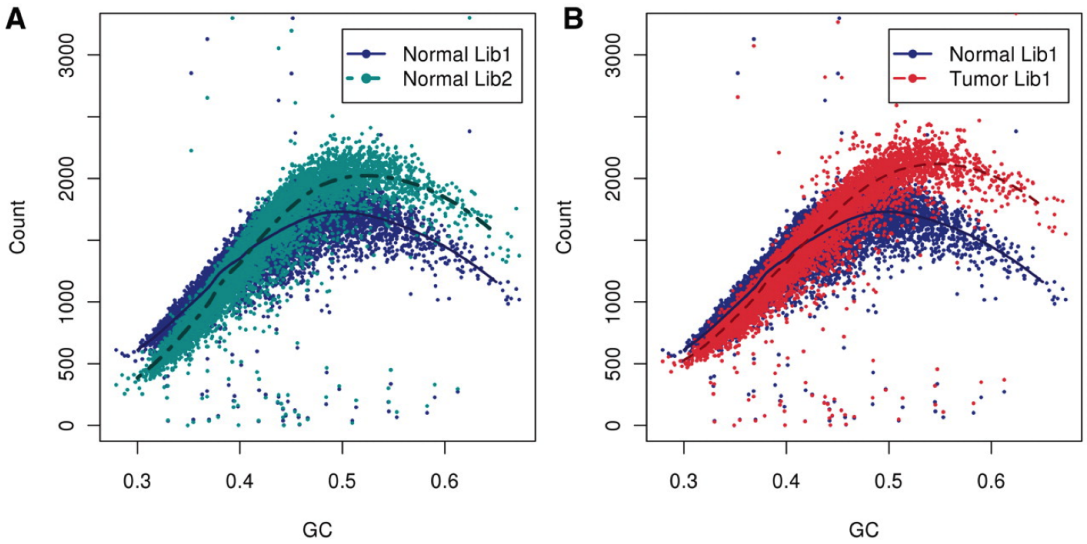
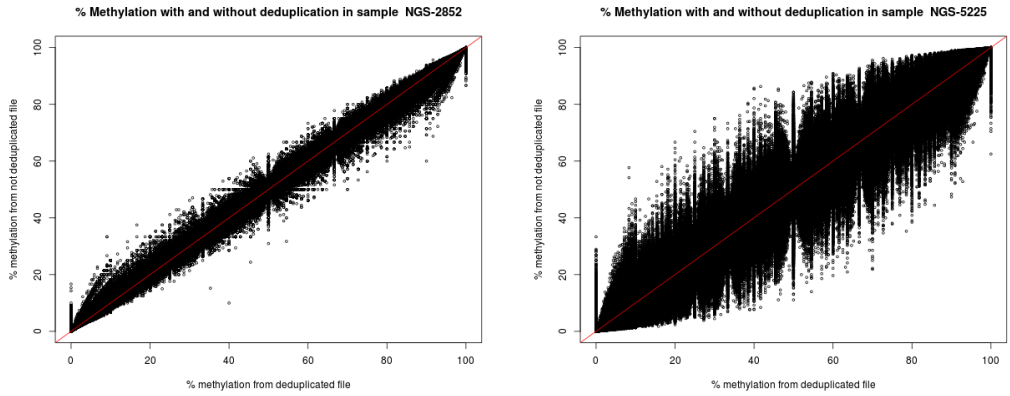


Figure 4.11: Effect of GC content on coverage: The left figure shows the effect in two normal libraries from the same sample, and the right one the tumor library in red and its matched normal library in blue [1]



(a) Normal sample 2852 (b) Adenoma sample 5225

Figure 4.12: Methylation percentages with (x-axis) and without (y-axis) the removal of duplicate reads

Chapter 5

Conclusion

We can conclude that we did find regions that showed loss of allele-specific methylation from the normal to the adenoma state. This is in regards to the ASMs, as predicted by `amrFinder`, that we found to be lost in all adenoma samples. What also ought to be kept in mind though, is the variety of cells in the adenoma stage and that we may want to be more lenient and look at ASMs that were lost in a smaller percentage of the samples, instead of demanding 100%. We did get a total of 61 regions however, with this strict demand which means we would get a vast amount of AMRs the less restraints we allow. This leads us back to the initial problem we posed: How reliable is `amrFinder`?

Our developed scoring method tried to evaluate the allele-specificity of the regions that `amrFinder` predicted to be AMRs. This score did perform well, with chromosome X as a benchmark of its performance. Nevertheless, we also found that our current penalty for the imbalance was too extreme, considering there is more variation amongst cells in adenoma tissue.

What also ought to be considered is the age of the individuals. Two of the three normal samples were from individuals in their 40s (42 and 44), whilst most of the adenoma samples had been taken from individuals aged from 61 to 85, with 73.67 being the average age, keeping in mind that one of the samples had an unknown age (NA). Methylation is also age related as certain CpGs are more likely to be methylated with age [29]. On the other hand, maybe cell type variation rather than age is what plays a bigger role in methylation differences. Jaffe and Irizarry [17] demonstrate that the change in cell composition in blood is what explains the observed variability in DNA methylation. Their analysis was done on blood but we may think of a similar situation for our samples, where cellular composition plays the defining role for methylation variability. With this in mind, the younger patients represented a reasonable control for our purposes.

With regard to the effect of removing duplicate reads, the consequences were not so drastic in terms of `bsseq`'s ability to predict most of the same DMRs. The methylation

levels that `bismark` predicted per CpG site however did seem to vary immensely. It is not so clear how much keeping duplicate reads can influence our analyses. Removing them was a reasonable option since our calculations for our developed scoring method were sensitive to the number of reads per CpG site, especially when considering a MM:UU imbalance. Moreover, as discussed before, there is an amplification bias for GC content that centers at around 50%. Since we are interested in allele-specific methylation removing duplicates seemed most appropriate.

5.0.4 Future Work

To further evaluate our developed scoring method and its performance, it ought to be simulated on data with known ASM regions. We may also want to be less strict in the penalty or weighting we apply for the MM:UU imbalance. There is great variety amongst cells, even more so amongst adenoma tissue. Demanding a 50:50 balance for allele-specificity may be neglecting regions of importance. One solution may be looking into the posterior distribution of MMs and considering the values that lie between 0.25 and 0.75.

We used a tuple size of two CpG sites. In the future, we may increase the tuple size, which may also increase our confidence in the evaluation of whether or not a site shows ASM. However, this may decrease the number of regions that we can interrogate.

To further demonstrate whether or not a site shows methylation that is allele-specific, it is useful to look at the SNPs. Having a heterozygous SNP that happens to be on an AMR can truly reflect if the region in question is an AMR, as predicted by `amrFinder`. For this, `BisSNP` was employed to look the SNPs at the level of AMRs and the reads that align. This will be further assessed in the future.

Appendix A

More Results

A.1 Coverage Histograms from TEQC

We present the rest of the coverage histograms as generated by `TEQC`.

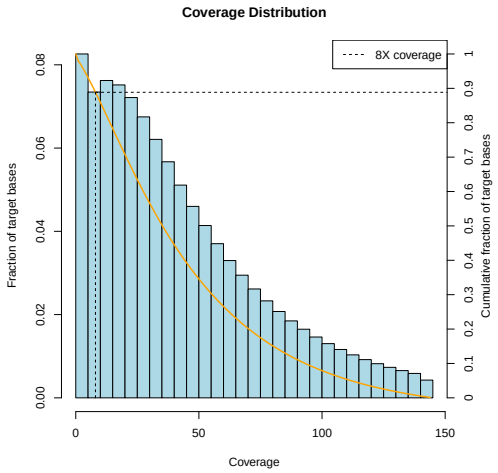


Figure A.1: Normal sample 2852

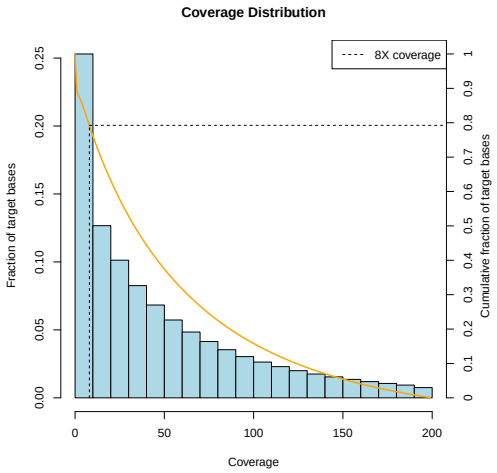


Figure A.2: Normal sample 5222

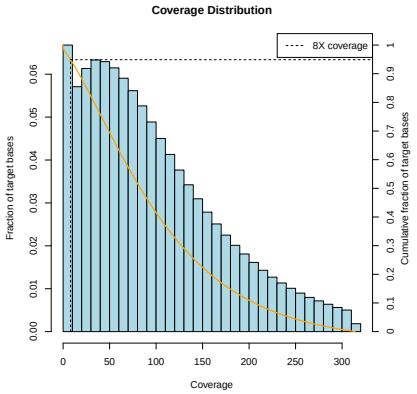


Figure A.3: Adenoma sample 5223

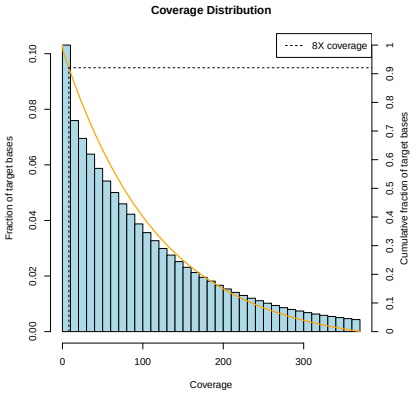


Figure A.4: Adenoma sample 5224

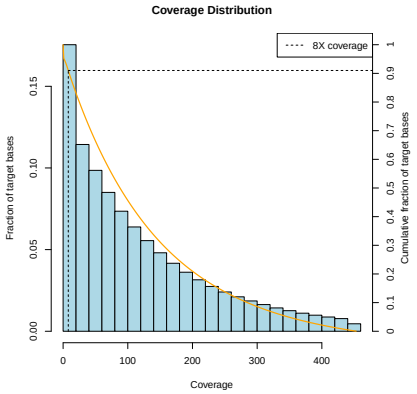


Figure A.5: Adenoma sample 5225

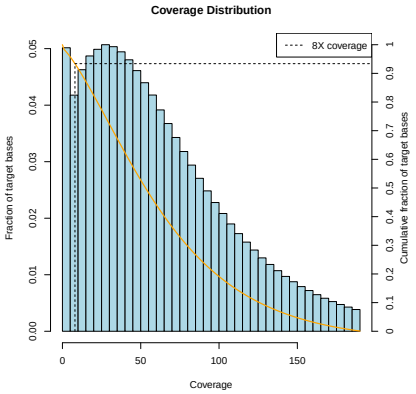


Figure A.6: Adenoma sample 5226

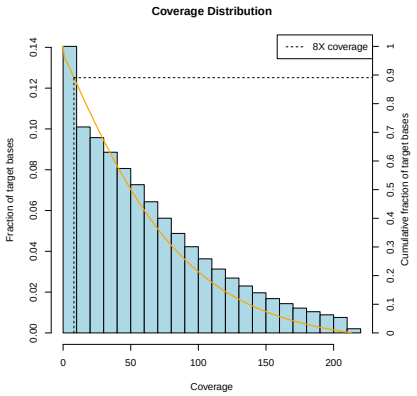


Figure A.7: Adenoma sample 5227

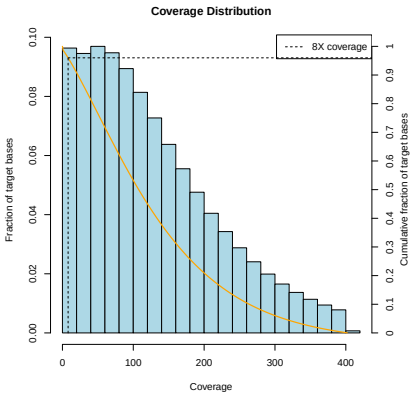


Figure A.8: Adenoma sample 5228

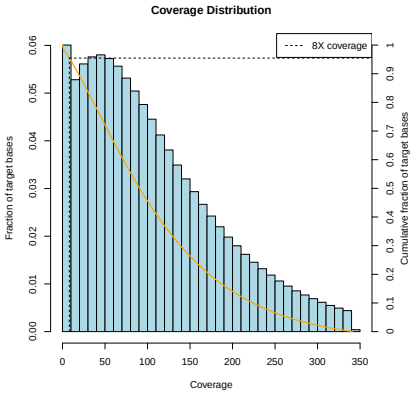


Figure A.9: Adenoma sample 5229

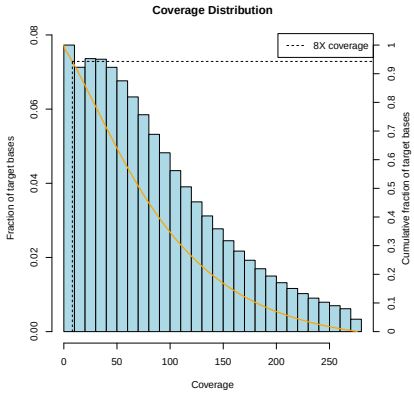


Figure A.10: Adenoma sample 5230

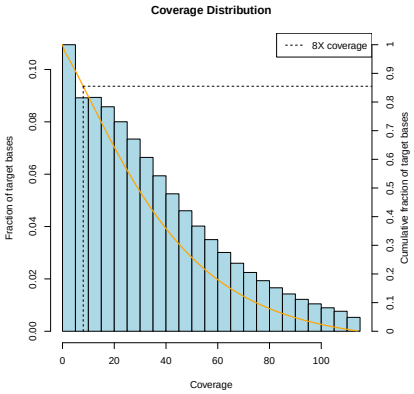


Figure A.11: Adenoma sample 5231

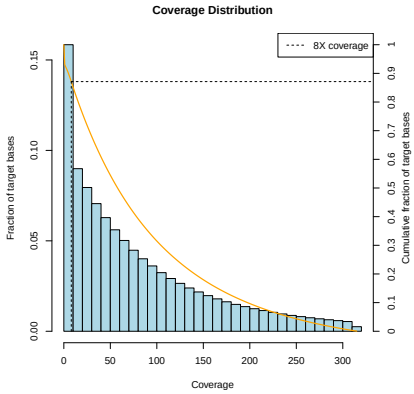


Figure A.12: Normal sample 5232

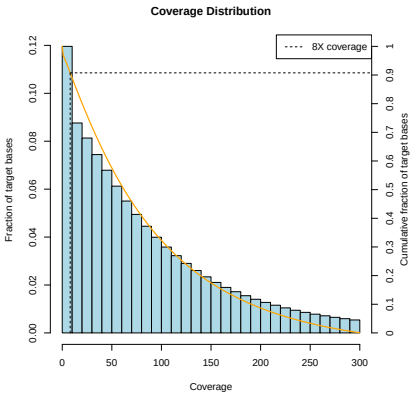


Figure A.13: Adenoma sample 5233

A.2 Lost AMRs

We present a full list of the AMRs that were present in all normal crypts and then lost in all adenoma samples as explained in section 4.1.

Table A.1: List of Lost AMRs

chr	start	end
chr1	7438997	7439955
chr1	113304600	113305990
chr1	166458061	166459858
chr1	234907852	234908938
chr10	106200601	106202271
chr11	34426004	34429220
chr12	70082181	70084382
chr13	49323359	49324361
chr14	20881768	20882846
chr14	21341359	21342129
chr17	74260293	74261077
chr2	136743316	136743920
chr20	5589667	5590558
chr21	36419215	36420754
chr22	32056705	32057473
chr5	106723364	106725895
chr5	113687838	113692663
chr6	45644732	45646111
chr6	79314327	79316836
chr6	108141210	108142925
chr7	154719421	154720721
chr8	436063	437268
chr8	61570774	61573554
chr1	7439355	7439955
chr1	113304573	113306033
chr1	166458184	166459433
chr1	234907891	234908684
chr11	34426042	34428466
chr12	70082181	70084087
chr13	49323359	49324111
chr14	20881768	20882839

chr	start	end
chr14	21341428	21342331
chr17	74260103	74261057
chr2	136743318	136743917
chr20	5589453	5590622
chr22	32056705	32057471
chr4	52918066	52918743
chr5	106723795	106725741
chr6	45644766	45645801
chr6	79312878	79318086
chr6	108141298	108142888
chr7	154719421	154720687
chr8	436063	437181
chr8	61570774	61573521
chr1	7438997	7439919
chr1	113304600	113306033
chr1	166458061	166459984
chr1	234908109	234908938
chr11	34426004	34428804
chr13	49322856	49324171
chr14	21341359	21342449
chr17	74260099	74261129
chr2	136743308	136743920
chr20	5587959	5590907
chr21	36419215	36420821
chr22	32056705	32057551
chr5	106723333	106725895
chr6	45644688	45646111
chr6	108141210	108142888
chr7	154719421	154720670
chr8	61570712	61573554

A.3 Boxplots of Our Weighted Score

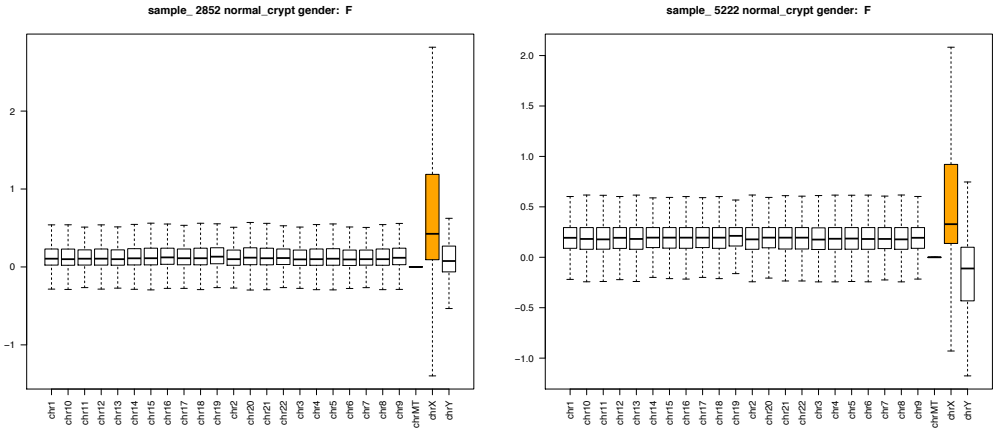


Figure A.14: Normal samples 2852 (F) and 5222 (F)

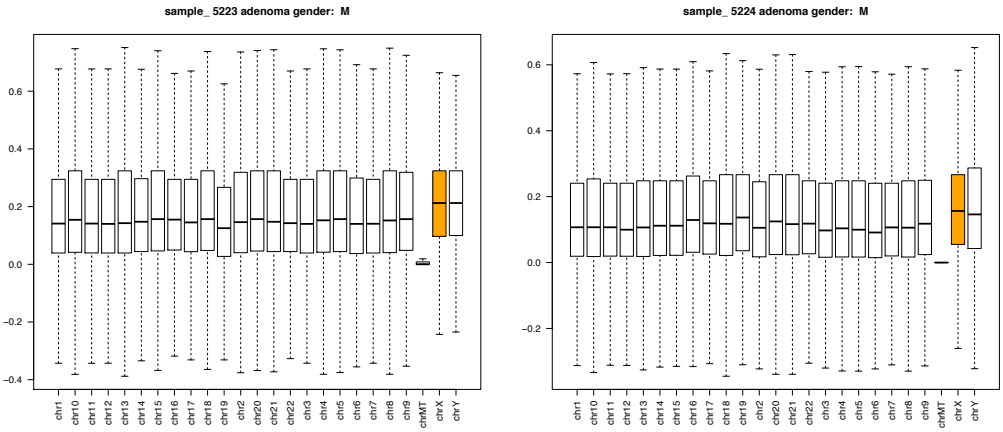


Figure A.15: Adenoma samples 5223 (M) and 5224 (M)

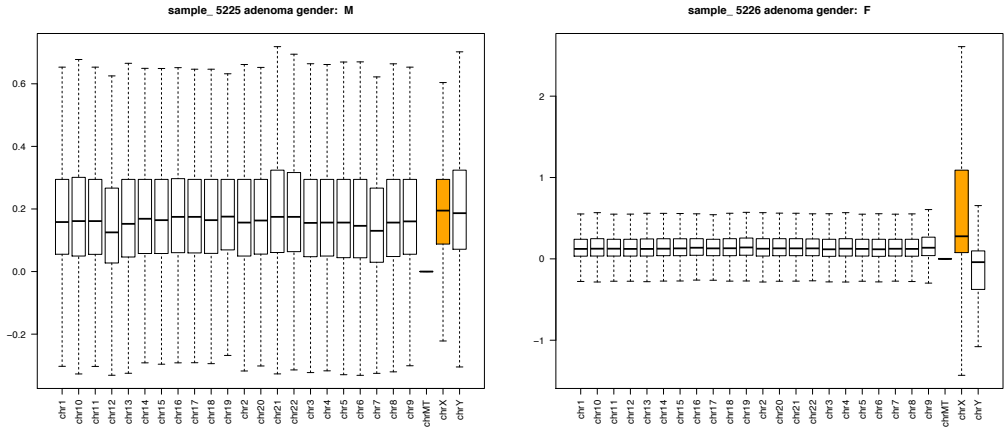


Figure A.16: Adenoma samples 5225 (M) and 5226 (F)

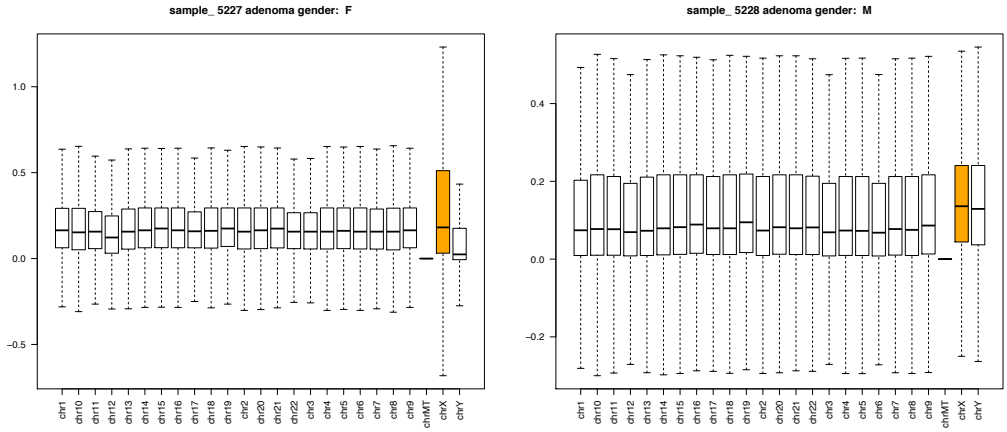


Figure A.17: Adenoma samples 5227 (F) and 5228 (M)

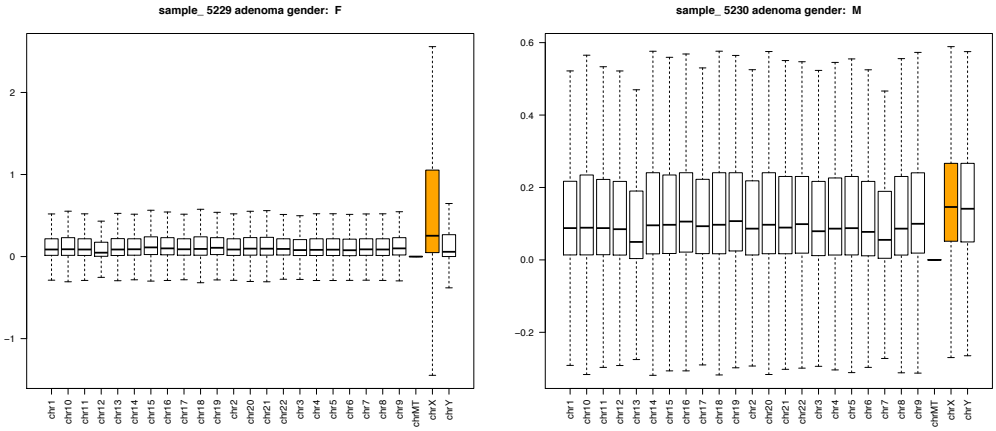


Figure A.18: Adenoma samples 5229 (F) and 5230 (M)

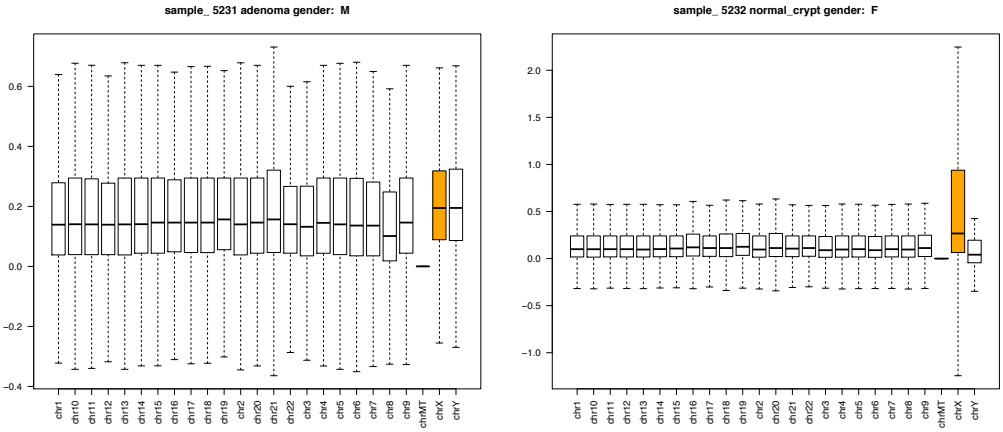


Figure A.19: Adenoma sample 5231 (M) and normal sample 5232 (F)

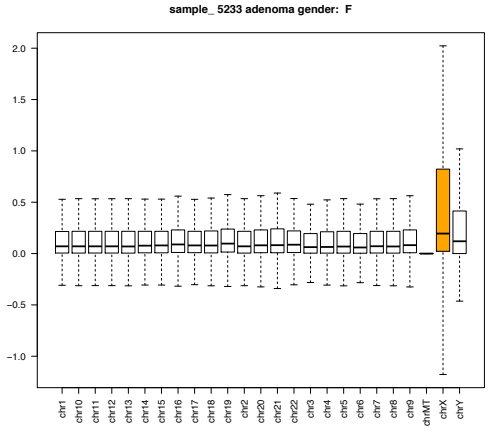


Figure A.20: Adenoma sample 5233 (F)

A.4 Top 400 DMRs

Table A.2: List of DMRs

chr	start	end
chr4	184826255	184828213
chr6	133561631	133563765
chr6	150285403	150286377
chr4	96468589	96471218
chr8	97506021	97507754
chr11	7272759	7274235
chr13	93879177	93881111
chr3	32858656	32860287
chr1	114695678	114697403
chr13	36919675	36921175
chr4	155663530	155666013
chr8	145105235	145106922
chr12	5018339	5020322
chr8	69242727	69244828
chr4	166794379	166796272
chr3	142837819	142839328
chr15	84748514	84749285
chr3	28616405	28618167
chr5	127872642	127873910
chr13	110959124	110960743
chr4	156680217	156681476
chr11	91957608	91959895
chr9	113341139	113342256
chr10	25464419	25465810
chr3	134514596	134515871
chr18	74961748	74963420
chr3	192126633	192128027
chr7	43152022	43153598
chr17	32906321	32907345
chr4	110223188	110224477
chr2	182321784	182322811
chr20	24450706	24451989
chr5	38257886	38259136
chr14	51560321	51561502
chr14	74706955	74708081
chr3	128719903	128721216
chr7	121512773	121513997
chr6	127440105	127441238
chr15	47476261	47477412
chr12	41582662	41583929
chr19	12266870	12267898

chr	start	end
chr11	88241521	88242360
chr4	82135578	82136780
chr4	142053329	142054396
chr3	2140759	2141938
chr17	62774951	62775860
chr15	83348841	83349827
chr1	65990767	65991910
chr7	134143072	134144235
chr11	104034007	104035152
chr4	30722520	30723580
chr6	32063940	32064765
chr3	185911421	185912478
chr4	144621244	144622182
chr8	53477318	53478662
chr21	22369475	22370515
chr3	139258010	139258823
chr13	36044769	36045655
chr5	134825480	134826485
chr2	31360516	31361105
chr8	68864376	68864982
chr13	103052382	103053653
chr6	84562837	84563752
chr7	132260852	132261563
chr11	112832507	112833250
chr10	7452347	7453557
chr6	71666111	71667178
chr8	85094524	85095813
chr6	391305	392247
chr5	136834000	136834584
chr8	79427938	79429099
chr4	6200918	6201722
chr3	24870809	24872350
chr11	120434877	120435721
chr2	115919870	115920849
chr15	79724122	79724747
chr15	48936812	48937449
chr12	104850682	104851661
chr3	35680792	35681706
chr13	96743228	96744162
chr7	49814368	49815693
chr10	7450500	7451315

Table A.3: List of DMRs (continued)

chr	start	end
chr2	149633184	149633851
chr12	66122805	66123506
chr22	48884899	48885562
chr7	149917322	149917887
chr12	41086275	41087222
chr21	34443342	34444107
chr8	31496527	31497241
chr3	134369502	134370095
chr6	152957549	152958518
chr5	54179431	54180171
chr4	101110975	101112318
chr1	20879028	20879645
chr11	125035539	125036426
chr7	49812962	49813518
chr10	22765371	22766094
chr6	39759898	39760608
chr2	115419532	115420183
chr6	117086407	117086997
chr13	37247896	37248676
chr2	74742153	74742914
chr4	122301540	122302205
chr9	3181008	3181490
chr7	84814982	84815941
chr1	107682962	107683854
chr18	11149276	11149842
chr8	139508891	139509729
chr3	156008947	156009944
chr6	6546302	6547167
chr16	6533021	6533701
chr19	57862403	57863217
chr6	100441461	100442106
chr11	30606512	30606996
chr7	45614710	45615558
chr5	83679698	83680285
chr21	32930102	32930685
chr8	67344302	67345107
chr9	13278513	13279241
chr14	70655152	70655806
chr3	157155269	157156397
chr8	49647519	49648218
chr18	59000801	59001611

chr	start	end
chr13	92051202	92051825
chr5	38556939	38557635
chr5	132947002	132947512
chr13	96296182	96296800
chr5	16179855	16180644
chr4	156129305	156130199
chr10	15761309	15762025
chr22	28196227	28196862
chr6	80656746	80657314
chr3	140770421	140771080
chr16	87636535	87636972
chr8	35092501	35093286
chr17	42635311	42635666
chr19	37406806	37407615
chr5	127874395	127875069
chr6	150358873	150359170
chr11	92702736	92703434
chr6	110678713	110679251
chr19	37288451	37288752
chr16	22824774	22825413
chr8	10590455	10591059
chr8	494182	494583
chr1	14925578	14925982
chr2	70994803	70995404
chr2	198650881	198651437
chr8	10588372	10589153
chr2	237145666	237146166
chr10	3822014	3824618
chr10	118031585	118032205
chr3	192232255	192232864
chr19	56904816	56905190
chr11	69632065	69632526
chr2	73518744	73519431
chr7	50344115	50344519
chr17	35165363	35165763
chr2	139537460	139537916
chr10	83633877	83634393
chr8	109095265	109095678
chr3	132757123	132757678
chr2	238535563	238536033
chr4	4388932	4389510

Table A.4: List of DMRs (continued)

chr	start	end
chr12	3602287	3602866
chr1	76080411	76080945
chr13	38443519	38444165
chr17	56234422	56234907
chr2	945524	946237
chr7	86273444	86274444
chr3	79067884	79068318
chr17	46673534	46674850
chr2	119067517	119067972
chr18	67067538	67067990
chr17	10101427	10101779
chr2	175546749	175547226
chr6	11043783	11044155
chr18	22930019	22930422
chr20	39319189	39319653
chr12	111471377	111471860
chr5	10564837	10565557
chr17	8533006	8533717
chr2	100938860	100939228
chr3	16554240	16554892
chr6	118228234	118228889
chr21	34442485	34443011
chr4	128544254	128544903
chr9	100615569	100615985
chr7	108095251	108095806
chr1	14924669	14925228
chr5	113391552	113392095
chr13	36704531	36705129
chr1	231298687	231299052
chr18	49866532	49867277
chr4	168155068	168155626
chr11	111411801	111412291
chr10	128077172	128077501
chr2	131720858	131721178
chr8	91997078	91997490
chr4	176986846	176987304
chr3	2139961	2140330
chr17	63532965	63534570
chr8	72755814	72756150
chr16	82660483	82660943
chr8	495818	496192

chr	start	end
chr1	76082024	76082569
chr12	41581794	41582350
chr8	11205362	11205785
chr8	67873541	67874152
chr3	5024126	5025663
chr11	112833366	112833975
chr2	233284239	233284775
chr5	37839685	37840414
chr2	164593066	164593412
chr7	54609769	54610075
chr5	115151823	115152327
chr11	123301362	123301711
chr19	46387439	46388419
chr8	11204744	11205013
chr3	49939356	49940698
chr22	50720245	50721530
chr9	98111366	98111688
chr20	62959174	62959388
chr19	56879646	56879995
chr4	156588197	156588602
chr3	139653531	139653938
chr2	115918468	115918972
chr11	15094959	15095437
chr22	46474731	46477145
chr12	50354841	50355308
chr11	69633798	69634136
chr5	88185306	88185774
chr17	46682916	46684765
chr7	18126786	18127238
chr10	108924642	108924964
chr7	132262270	132262562
chr11	105480925	105481779
chr8	16884369	16884632
chr17	66596132	66596638
chr18	49868166	49868651
chr2	29338728	29339077
chr7	142494493	142495271
chr3	38690212	38690714
chr7	101006102	101006313
chr3	186857129	186857407
chr11	22214932	22215342

Table A.5: List of DMRs (continued)

chr	start	end
chr7	79081696	79082305
chr8	10589590	10590099
chr22	39954187	39954503
chr3	195599992	195601156
chr5	63461450	63461638
chr2	121104180	121104437
chr7	45613411	45613813
chr2	161127253	161128447
chr6	152129011	152129420
chr3	142681928	142682292
chr11	69061109	69063401
chr1	2427602	2428603
chr2	133426528	133426891
chr2	100937597	100937981
chr10	7454295	7454623
chr7	870022	872218
chr11	8284248	8284529
chr4	21950037	21950366
chr19	53635778	53636188
chr6	168167669	168168651
chr1	115880398	115880731
chr1	86621626	86622114
chr7	151106846	151107102
chr2	142887726	142888073
chr5	82768921	82769268
chr11	111383662	111383892
chr8	142240470	142241381
chr5	138729770	138729967
chr2	101033610	101033858
chr17	78793303	78794367
chr6	161188131	161188397
chr11	115530548	115531177
chr3	142839904	142840174
chr6	170491742	170492379
chr7	18125703	18126071
chr5	38556204	38556421
chr11	116451330	116451698
chr1	2980095	2980419
chr7	330556	330850
chr2	175547780	175548165
chr4	55098322	55098693

chr	start	end
chr6	94126274	94126559
chr10	102590152	102590444
chr20	58180269	58180561
chr10	26223707	26224048
chr13	88327724	88329318
chr17	46685117	46685873
chr13	36871523	36871963
chr11	69633137	69633366
chr18	22929113	22929480
chr3	26664544	26664941
chr17	79816428	79817160
chr7	154996875	154997546
chr2	66808617	66808897
chr10	125851534	125851788
chr13	114770157	114771155
chr5	11384582	11384890
chr11	30605858	30606110
chr7	90895084	90895328
chr3	101396659	101397389
chr7	18126170	18126297
chr11	94502136	94502476
chr7	149744630	149744945
chr1	165414244	165414545
chr16	20359877	20360260
chr2	230578124	230578455
chr6	110679567	110679730
chr17	37011404	37011650
chr13	114875869	114876748
chr2	154334553	154334867
chr16	87441257	87441795
chr7	49813871	49814209
chr16	85981191	85981752
chr19	30019540	30019753
chr17	79315784	79316098
chr7	145813480	145813848
chr16	10652485	10653304
chr8	12989287	12989635
chr14	23821230	23821446
chr10	100993837	100994050
chr13	114497597	114498384
chr22	46460253	46461424

Table A.6: List of DMRs (continued)

chr	start	end
chr7	1098970	1099695
chr11	47358943	47359273
chr11	8040323	8040595
chr20	61050501	61050696
chr2	240196172	240197267
chr2	105459713	105459939
chr3	159944288	159944522
chr2	71693210	71693456
chr10	102821566	102822050
chr11	8102312	8102569
chr6	118229611	118229837
chr5	133449322	133449584
chr1	47915469	47916075
chr3	193587524	193588047
chr2	71017421	71017629
chr7	3341473	3341871
chr1	98519238	98519505
chr19	57019217	57019388
chr1	86622331	86622629
chr5	15500691	15500838
chr5	179780601	179780782
chr7	852607	853314
chr19	57049890	57050099
chr4	5889936	5890082
chr16	85516952	85517811
chr5	15500078	15500230
chr11	68610242	68611025
chr10	26223240	26223372
chr16	1598764	1599254
chr9	77113233	77113381
chr4	156589081	156589292
chr2	1747687	1747825
chr5	891005	891800
chr15	78912607	78912786
chr10	108923927	108924133
chr19	57018989	57019190
chr3	150237739	150238558
chr3	142683111	142683330
chr3	150803002	150803136
chr11	24518540	24518679
chr1	1267194	1267871

chr	start	end
chr7	2149698	2150483
chr8	89340032	89340206
chr18	6929477	6930453
chr20	24450041	24450168
chr14	27067621	27067917
chr4	183369632	183369853
chr20	36150939	36151625
chr1	243431211	243432313
chr12	8171311	8171415
chr7	1025677	1026441
chr2	68546744	68546867
chr15	93632677	93632830
chr4	156297497	156297702
chr7	5567393	5568285
chr13	113648952	113649527
chr10	43600344	43600618
chr11	17497465	17497598
chr18	77542922	77543423
chr19	1439912	1440306
chr3	12046409	12046517
chr10	93392705	93392905
chr7	80549873	80551770
chr8	58907730	58907845
chr7	157406033	157406144
chr5	177030964	177031571
chr10	102586127	102586340
chr2	29338070	29338184
chr13	36705514	36705682
chr1	234908089	234908524
chr10	83634994	83635136
chr4	3386070	3387208
chr20	34894501	34894691
chr13	91827455	91827601
chr6	168812138	168812847
chr19	37464402	37464675
chr6	11044842	11044960
chr4	21950682	21950844
chr20	62679561	62679693
chr8	76318489	76319159
chr1	182809948	182811225
chr22	19868605	19869425

Appendix B

Abbreviations

Abbreviation	Meaning
AID	Activation-Induced Cytosine Deaminase
AMR	Allelically Methylated region
APOBEC1	Apolipoprotein B RNA-Editing Catalytic Component 1
ASM	Allele Specific Methylation
bp	base pairs
BsSeq	Bisulfilte Sequencing
CGI	CpG Islands
CpG	Cytosine phosphate Guanine
DMR	Differentially Methylated Region
DNMT	DNA Methyltransferase
fastQC	Fast Quality Control
MeDip	Methylated DNA Immunoprecipitation
MM	Methylated Methylated
MU	Methylated Unmethylated
PCR	Polymerase Chain Reaction
PE	Paired End
RRBS	Reduced Representation Bisulfite Sequencing
TEQC	Target Enrichment Quality Control
UM	Unmethlated Methylated
UU	Unmethylated Unmethylated
WGBS	Whole Genome Bisulfite Sequencing

Bibliography

[1] Yuval Benjamini and Terence P. Speed. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research (Oxford Journals)*, May 2012.

[2] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford Journals)*, April 2014.

[3] Johns Hopkins Colon Cancer Center. *From Polyp to Cancer*. found under http://www.hopkinscoloncancercenter.org/CMS/CMS_Page.aspx?CurrentUDV=59CMS_Page_ID=5DE6-4CB4-B387-4158CC924084.

[4] David Cunningham, Wendy Atkin, Heinz-Josef Lenz, Henry T Lynch, Bruce Minsky, Bernard Nordlinger, and Naureen Starling. Colorectal cancer. *The Lancet*, 375:1030–1047, March 2010.

[5] diagenode. Applications — Bisulfite conversion. found under http://www.diagenode.com/en/applications/bisulfite-conversion.php.

[6] Fang Fang, Emily Hodges, Antoine Molaro, Matthew Dean, Gregory J. Hannon, and Andrew D. Smith. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *PNAS*, 109, March.

[7] Fang Fang and Andrew Smith. *The amrfinder Manual*, April 2012.

[8] Marianne Frommer, Louise E McDonald, Douglas S Millar, Christina M Collis, Fujiko Watt, Geoffrey W Grigg, Peter L Molloy, and Cheryl L Paul. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Genetics*, 89:1827–1831, March 1992.

[9] Kasper D Hansen, Benjamin Langmead, and Rafael A Irizarry. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biology*, 13:R83, October 2012.

[10] Peter Hickey. *Methtuple*, 2012. found under https://github.com/PeteHaitch/methtuple.

[11] Manuela Hummel, Sarah Bonnin, Ernesto Lowy, and Guglielmo Roma. TEQC: an R package for quality control in target capture experiments. *Bioinformatics (Oxford Journals)*, February 2011.

[12] Inc. Illumina. *Infinium Methylation Assay*. found under http://www.illumina.com/technology/beadarray-technology/infinium-methylation-assay.html.

[13] Inc. Illumina. Paired-end sequencing. found under http://www.illumina.com/technology/next-generation-sequencing/paired-end-sequencing_assay.html.

[14] Inc. Illumina. *Field Guide to Methylation Methods*, September 2013. found under http://www.illumina.com/content/dam/illumina-marketing/documents/products/other/field_guide_methylation.pdf.

[15] NIH National Cancer Institute. *General Information About Colon Cancer*. found under http://www.cancer.gov/types/colorectal/patient/colon-treatment-pdq.

[16] Rafael A. Irizarry, Christine Ladd-Acosta, Bo Wen, Zhijin Wu, Carolina Montano, Patrick Onyango, Hengmi Cui, Kevin Gabo, Michael Rongione, Maree Webster, Hong Ji, James Potash, Sarven Sabunciyan, and Andrew P. Feinberg. Genome-wide methylation analysis of human colon cancer reveals similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nature Genetics*, 41:178–186, January 2009.

[17] Andrew E Jaffe and Rafael A Irizarry. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biology*, 15, February 2014.

[18] Randy L. Jirtle. *Geneimprint*. found under http://www.geneimprint.com/site/what-is-imprinting.

[19] Igor Kovalchuk and Olga Kovalchuk. *Epigenetics in Health and Disease*. FT Press, 2012.

[20] Felix Krueger and Simon R. Andrews. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics (Oxford Journals)*, 27, June.

[21] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Oxfor Journals (Bioinformatics)*, 25:2078–2079, May 2009.

[22] Yuanyuan Li and Trygve O. Tollefsbol. DNA methylation detection: Bisulfite genomic sequencing analysis. *Methods in molecular biology (Clifton, NJ)*, 791:11–21, December 2011.

[23] Magrane M. and the UniProt consortium. Q01196 - runx1_human. found under http://www.uniprot.org/uniprot/Q01196.

[24] Magrane M. and the UniProt consortium. Q8tc71 - mieap_human. found under http://www.uniprot.org/uniprot/Q8TC71.

[25] Alexander Meissner, Andreas Gnirke, George W. Bell, Bernard Ramsahoye, Eric S. Lander, and Rudolf Jaenisch. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Oxford Journals - Nucleic Acid Research*, 33:5868–5877, September 2005.

[26] Tom Moore and David Haig. Genomic imprinting in mammalian development: a parental tug-of-war. *Trends in Genetics*, 7:45–49, February 1991.

[27] UCSF School of Medicine. *Genes and Genomes – Epigenetics*, July 2007. found under http://missinglink.ucsf.edu/lm/genes_and_genomes/methylation.html.

[28] Lionel A Sanz, Satya K Kota, and Robert Feil. Genome-wide DNA demethylation in mammals. *Genome Biology*, 11:11–110, March 2010.

[29] Nobuyoshi Shimoda, Toshiaki Izawa, Akio Yoshizawa, Hayoto Yokoi, Yutaka Kikuchi, and Naohiro Hashimoto. Decrease in cytosine methylation at CpG island shores and increase in DNA fragmentation during zebrafish aging. *American Aging Association*, June 2013.

[30] Agilent Technologies. *SureSelect - How it Works*. found under http://www.genomics.agilent.com/article.jsp?pageId=3083.

[31] Agilent Technologies. *Agilent SureSelect Human Methyl-Seq for the Quantitative Analysis of DNA Methylation with Single-Base Resolution*, May 2012. found under http://www.agilent.com/genomics/sureselect.

[32] Helga Thorvaldsdóttir, James T. Robinson, and Jill P. Mesirov. *Briefings in Bioinformatics*.

[33] Hong Tran, Jacob Porter, Ming an Sun, Hehuang Xie, , and Liqing Zhang. Objective and Comprehensive Evaluation of Bisulfite Short Read Mapping Tools. *Advances in Bioinformatics*, April 2014.

[34] Wikipedia. *Illumina Methylation Assay*. found under https://en.wikipedia.org/wiki/Illumina_Methylation_Assay.

[35] Yingying Zhang, Christian Rohde, Richard Reinhardt, Claudia Voelcker-Rehage, and Albert Jeltsch. Non-imprinted allele-specific DNA methylation on human autosomes. *Genome Biology*, December 2009.

[36] Michael J. Ziller, Hongcang Gu, Fabian Müller, Julie Donaghey, Linus T.-Y. Tsai, Oliver Kohlbacher, Philip L. De Jager, Evan D. Rosen, David A. Bennett, Bradley E. Bernstein, Andreas Gnirke, and Alexander Meissner. Charting a dynamic DNA methylation landscape of the human genome. *Nature*, 500:477–481, August 2013.

ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

ANALYZING THE LOSS OF ALLELE-SPECIFIC METHYLATION IN HUMAN COLORECTAL ADENOMA WITH BISULFITE SEQUENCING

Authored by (in block letters):

For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**
MACHLAB	DANIA

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date	**Signature(s)**
Zürich 07.07.2015	MachlabD

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.