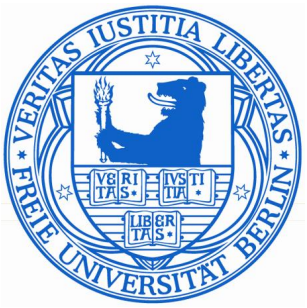
Master Thesis

Developing a BS-Seq Analysis Workflow for Genomic Variation and Methylation Level Calling

Sabrina Krakau

October 29, 2013



Freie Universität Berlin
Bioinformatik

Supervisors:

Prof. Dr. Knut Reinert
Freie Universität Berlin
Algorithmic Bioinformatics

Prof. Dr. Martin Vingron
Max-Planck-Institute for Molecular Genetics
Berlin

Declaration of Originality

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institute of tertiary education. Information derived from the work of others has been acknowledged.

Sabrina Krakau

Next-Generation Sequencing technologies combined with a bisulfite treatment of DNA (BS-Seq) present an efficient method in the field of epigenomics, allowing for a precise analysis of methylation patterns at single-nucleotide resolution. The expanding field leads to the production of enormous amounts of data, that need to be analysed using precise and efficient software tools. The analysis needs to address the specific challenges arising from the bisulfite treatment, which converts unmethylated Cs to Ts. We present a powerful analysis workflow, implemented using the C++ software library SeqAn, for accurate bisulfite read mapping, the detection of single-nucleotide polymorphisms and methylation level calling at single-nucleotide resolution. Furthermore, we provide a multiple sequence realignment method to improve the alignment accuracy for indel reads. We show that in comparison to existing methods, we achieve comparable and even better results.

Acknowledgements

I like to thank my two supervisors, Prof. Dr. Knut Reinert and Prof. Dr. Martin Vingron, for giving me the opportunity to work on this exciting and challenging project.

I want to thank Dr. David Weese for the continuous support, ideas and advises and acknowledge that you were always available when I had questions or difficulties. I enjoyed working on the project, even though I had some challenging times.

Thank you Felix and Jochen for your continuous and last minute support. In addition I want to thank the whole SeqAn team for their your useful advises.

Last but not least I want to thank my family and friends for their continuous and unconditional support.

Contents

1 Introduction

1.1 Introduction

Until the last decades of the 20th century, researchers thought that the entire hereditary information of an organism is encoded in its DNA sequence [29]. DNA is built from two long biopolymers, consisting of the nucleotides adenine (A), cytosine (C), guanine (G) and thymine (T), which form together a double stranded helix. The sequence of the nucleotides encodes the genetic information. In 1975 an epigenetic mechanism was discovered, in which methylation groups attached to the DNA influence its gene expression. In 1987 [28] it was then shown that such DNA methylations do not only have a regulatory role, but even are inherited to the next generation. This was an important turn towards the understanding that the DNA sequence is just one part of a much more complex genetic system. The information stored in the DNA sequence itself is more or less stable, and is identical in each cell of one organism. Nevertheless, cells differ significantly between different tissues, different points in time during development or cell proliferation, and between different individuals.

Today it is known that DNA methylation is one of the most important epigenetic mechanisms in higher eukaryotes. It enables the modification of gene expression without changing the underlying DNA sequence. Furthermore, the patterns of DNA methylation are flexible and can be influenced by exogenous factors. In other words, the regulatory elements of one individual can be adjusted to external conditions and passed on to its descendants. This explains the wide range of diseases that are linked not only to genomic differences, but to epigenetic variations.

In order to understand how such diseases arise, the genetic and epigenetic information must to be considered together. In 2001 one important milestone was set in history when the Human Genome Project, a collaboration of researchers from all over the world, released the entire sequence of 3 billion base pairs of the human genome. This laid the foundation for further basic research, answering questions about genomic variations, regulatory functions and the cause of diseases, always with the aim to develop possible treatments. The first draft of the human genome was obtained by Sanger sequencing, a method developed in 1977 [56]. However, using this method, it took researchers 10 years and $3 billion. Nowadays, next-generation sequencing (NGS) technologies, such as Illumina sequencing, allow for a much faster analysis while producing only a small fraction of the original costs. As a consequence, today NGS sequencing methods are commonly used to explore genetic variations in the field of diagnostics.

Based on this, the research could be extended for epigenetic information. Roughly at the time when the Human Genome Project came to an end, the Wellcome Trust Sanger

Institute (UK), the Epigenomics AG (Germany) and the Centre National de Génotypage (France) started the Human Epigenome Project [54]. The objective of this EU funded project was to identify, catalogue and interpret genome-wide DNA methylation patterns throughout the human genome [4]. Therefore special sequencing technologies were used to determine methylations encoded within the sequence. In 2006 the first DNA methylation map was published for the human chromosomes 6, 20 and 22.

However, even though today whole genome methylation patterns are released for the human, it is still an open task to analyse the methylomes of individuals, specific cell types, development states or other organisms. The recent high-throughput sequencing technologies help to answer the question how genes are switched on and off in the genome and open new possibilities in the field of diagnostics and personalised medicine. Clearly, these technologies come along with the need for fast and accurate tools in order to analyse the data in a high resolution.

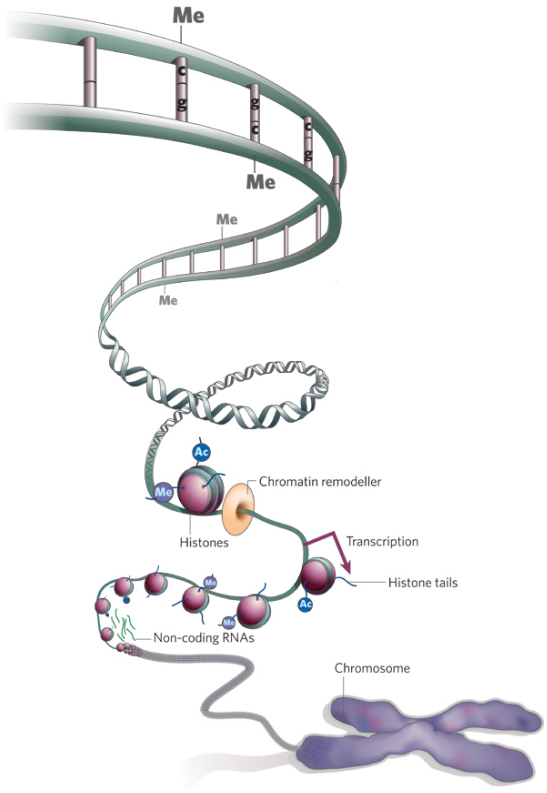


Figure 1.1.1: The double stranded DNA is wrapped around histone proteins, which organize a higher-order chromatin structure that finally builds the chromosome. Methylations (Me) on cytosine bases (C) play an important role for the formation of the chromatin structure, which in turn regulates the DNA transcription and thus activates or inactivates genes. (Source: *Nature* 2008; 454: 711–755)

1.1.1 Motivation

In order to contribute to the expanding field of epigenomics and improve the analysis of the in large quantities produced data, the aim of this thesis is the development of an accurate method to analyse DNA methylations.

The analysis can be done based on read sequences obtained with NGS bisulfite sequencing technologies. These contain, due to a specific bisulfite treatment prior the the actual sequencing, the information about methylation sites encoded in their sequence. The first goal of this thesis is to implement a read mapping method in order to compute the genomic origin of the bisulfite reads. In a second step, we implement a method using all mapped reads to decode the information and to analyse the DNA methylations at single-nucleotide resolution. For both methods the aim is to achieve highly accurate and precise results. In order to address this tasks, existing methods will be examined and if required modified or extended.

Further, this methods should be implemented in efficient data structures to enable a reliable and fast analysis of real biological data. Therefore SeqAn, an efficient and generic C++ template library for sequence analysis [17] will be used, which offers already a range of data structures that can be used for this work.

1.2 Biological Background

DNA methylation is beside histone modification the most important epigenetic modification mechanism in eukaryotes based on methyl groups that are added to the C5 carbon residue of cytosine nucleotides (C). Enzymes called DNA methyltransferases ensure the methylation maintenance of cytosines that occur in the context of the dinucleotide CG (CpG) [53] through DNA replication and thus through the process of cell division. This enables the inheritance of those non-sequence based information to the next generation. Methylations of non-CpG cytosines, on the other hand, evolve most often anew after each DNA-replication [53]. Despite their stability with respect to DNA replication, methylations vary between different cell lines, points in time and even between cells of the same type.

1.2.1 DNA methylation patterns

For a better understanding of how to analyse DNA methylations, we will first have a look at how the methylation groups are distributed over the DNA.

In the double-stranded DNA molecule complementary nucleotides, namely A and T as well as C and G, are paired such that the genetic information of the forward strand is indirectly encoded in the reverse strand. Methyl groups on the contrary are only bound to C nucleotides resulting in distinct methylation patterns for both strands. Cytosine methylation rates are highly dependent on the sequential context in the DNA, more precisely on the subsequent bases in 5' to 3' direction of the polymer. One distinguishes mainly between three different contexts: CG, CHG and CHH, where H stands for any nucleotide other than G. The by far highest methylation rate occurs in the CG context. Due to the complementary properties, the CG context is the only context occurring symmetrical on both DNA strands. Such CpG dinucleotides are mostly methylated symmetrically regarding the two C nucleotides. However, several studies revealed also asymmetric CpG methylations ([62], [60]), which might play a role in cancer [19] among other cases.

In a long term view, methylated Cs have a higher probability to be converted to Ts than unmethylated Cs caused by a deamination of the 5-methylcytosine (5mC) [40]. Given the relatively high methylation rate of the CG context, this conversion causes a genome wide under-representation of CpGs compared to the expected rate considering the single nucleotide frequencies of C and G. Unlike most other cytosines in CG context, cytosines in CpG islands – genomic regions that contain a fairly high percentage (>50%) of CpG dinucleotides – show a very low methylation rate. As a result, CpG islands are associated with genomic conservation [10].

CpG islands are typically located near or within promoter regions and act as regulatory functional elements. However, the DNA methylation itself has a big impact on the transcriptional regulation. Methylation in promoter regions or nearby CpG islands as well as in coding regions is shown to have an inhibitatory effect on gene transcription [33]. Thus promoter regions and CpG islands are mostly unmethylated while intragenetic regions show in general a higher methylation rate [9]. The underlying regulatory

mechanism of DNA methylation is complex and not fully known yet. One way of regulation is based on the methyl-CpG-binding domain proteins (MBDs), which are involved in different regulatory processes, e.g. the modification of the chromatin structure.

Since methylation plays a crucial role in many regulatory processes, it comes as no surprise that those epigenetic variations are associated with different kind of cancers. "Cancer is an epigenetic disease at the same level that it can be considered a genetic disease." [21]. Hypermethylation of tumour-suppressor gene promoters, for example, is a common cause for the inactivation of tumour suppression, while global hypomethylations lead to a decreased stability of the chromatin structure of the genome [12]. Embryonic stem cells on the other hand feature a significantly higher non-CG context methylation rate as well as methylation pattern modifications during the differentiation. This allows the cellular differentiation of different cell types with various gene expression patterns [45].

If we now shift the view from the genomic methylation pattern of one single cell to a group of cells, eg. of the same type, we can assign each genomic C a methylation level between 0 and 1. This methylation level represents the frequency of this specific C being methylated across the sample. Previous research displayed that methylation levels are beta distributed [8] [18]. In a CG context this beta distribution is bimodal, where the most Cs are either high frequently methylated or unmethylated throughout the cell mixture, while just a small fraction of Cs is fuzzy methylated. In the CHG and CHH context the methylation levels show an unimodal beta distribution with the most Cs being unmethylated in all or most of the cells.

The genome wide context dependent methylation rates differ not only significantly between different organisms as plants and animals, but also between insects and mammalians [48].

1.2.2 Classical DNA methylation detection

A range of different strategies exists to detect methylation patterns with a more or less high resolution. One example is the widely used methylated DNA immunoprecipitation (MeDIP), which takes advantage of antibodies binding to methylated Cs [39]. This allows for the purification of methylated DNA fragments. Afterwards the selected DNA fragments can be hybridized to a microarray feature and detected (MeDIP-chip). The drawback of this method is the limitation to the microarray design. Another method makes use of Next-Generation Sequencing (NGS) techniques to sequence the purified fragments (MeDIP-seq). In this way, the methylation rates of genomic regions are indirectly estimated by the coverage. A similar strategy is based on proteins containing a methyl-CpG-binding domain (MBD), again purifying methylated DNA fragments [39]. These approaches provide the information about methylated regions in a rather low resolution, as no position specific detection is possible.

1.2.3 Bisulfite sequencing strategies

However, Next-Generation Sequencing (NGS) techniques provide a single-nucleotide sequence resolution and thus open new possibilities in this field. Today bisulfite sequencing (BS-Seq) is the gold-standard technology to analyse methylation patterns at a single-base resolution. It makes use of the different characteristics of methylated and unmethylated Cs under bisulfite treatment. When single-stranded DNA gets treated with bisulfite, unmethylated Cs deaminate to the nucleotide uracil (U), whereas methylated Cs remain unaffected [43]. In the subsequent sequencing process, these Us – previous unmethylated Cs – are sequenced as Ts. That implies the encoding of the individual methylation states in the read sequence and enables a later decoding.

Two common, but substantially different, protocols exists: The directional protocol (Lister et al. [44]) and the non-directional protocol (Cokus et al. [11]). The directional protocol is less complex. In short, genomic DNA of a cell mixture is randomly fragmented, adapters are ligated and gel electrophoresis is applied to select fragments of the desired size. Afterwards the DNA is treated with bisulfite. Next, the DNA is amplified with PCR (polymerase chain reaction) and another set of adaptors is ligated for the final sequencing process. The detailed steps of the directional sequencing process are presented in Figure 1.2.2. Since methylation patterns are not symmetrical on the two DNA strands, it is crucial to distinguish between the top and the bottom strand. Furthermore, during the amplification reverse complements of the original top and original bottom strands are synthesized. The directional protocol ensures with the help of special adapters, that in the case of single-end reads only fragments from the original top or bottom strand are sequenced [32]. Hence unmethylated genomic Cs will be represented as Ts in all reads. For paired-end reads additionally, the reverse complements of the original strands obtained by PCR amplification are sequenced. Again the adapters ensure that the left read is always originating from one of the original strands, while the right read in this case is always originating from the reverse complement (see Figure 1.2.1). In this way the directionality is maintained, or in other words, for each read, given its orientation regarding the reference sequence, the origin is known. Orientation can be obtained using a range of different bisulfite read mapping strategies. For the subsequent analysis, this means knowing whether each read contains the information about the top or about the bottom strand DNA methylations. Since the right reads of read pairs are resulting from the reverse complements, previous genomic bisulfite C>T conversions are displayed in this case as G>A conversions.

The non-directional protocol differs from the directional protocol in that the first adapter set is removed, followed by the ligation of an additional adapter set and an additional PCR amplification. This process yields to four different kind of reads, both for single- and paired-end reads. The first two hold the information of either the original top strand or the original bottom strand and contain Ts resulting from unmethylated genomic Cs. Additionally reads are sequenced from the reverse complements of the original strands, consequently containing possibly G>A conversions. The main drawback of this protocol is the loss of the directionality, i.e. each read can be originating from any strand. Therefore in the subsequent analysis it is not known whether a read contains

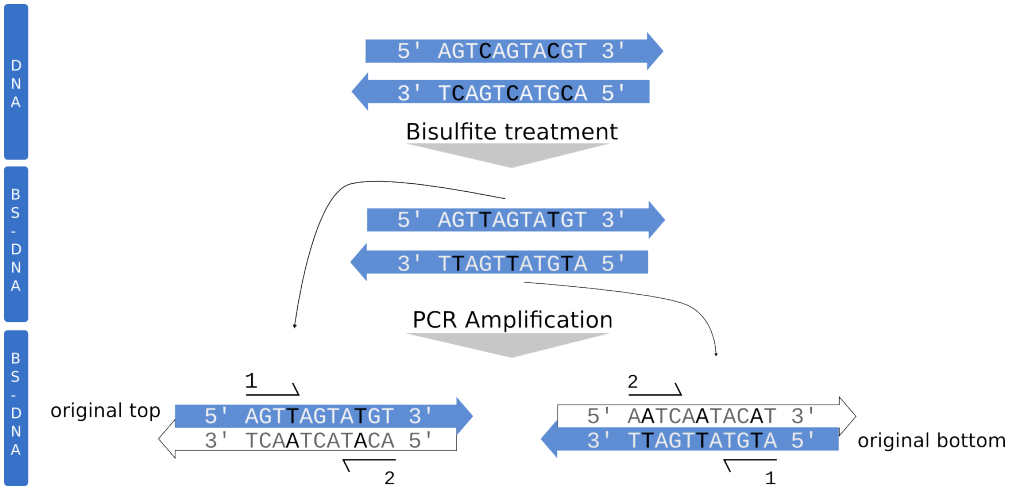


Figure 1.2.1: Directional BS-Seq for paired-end reads: Left reads (1) are either originating from the original top or from the original bottom strand. Right reads (2) are always originating from the reverse complements of the original top or bottom strand.

---

the information about the top or the bottom strand. However, if the read has a mapping giving the orientation and a certain number of observable conversions, this can be estimated by the type of conversion, i.e. C>T or G>A. The most widely used protocol is the directional one [46], most likely as a consequence of its more unambiguous data. The Illumina BS-Seq protocol is directional.

BS-Seq protocols differ fundamentally from other sequencing protocols in multiple points. One difference is, that whole genome amplification (WGA) cannot be applied upstream as in conventional sequencing strategies, because this would destroy the methylation patterns of interest. However, BS-Seq protocols deal with DNA fragments from cell ensembles. Since the methylation status of one specific genomic C position can differ between different cells, the resulting reads hold the information about the methylation levels across the cell ensemble rather than about the absolute methylation patterns in one cell. As a result, methylation differences between different cell types or between healthy and cancer cells can be analysed fairly well, but not between single cells. One way to circumvent this inaccuracy is single-cell sequencing. This on the other hand causes problems due to the difficult isolation process itself [35]. Moreover, the bisulfite conversion rate of unmethylated Cs is not always 100% and since each genomic C is only represented once, imperfect bisulfite conversions would cause wrong results.

1.2.4 Reduced representation bisulfite sequencing (RRBS)

Another approach is the reduced representation bisulfite sequencing (RRBS), which limits the genomic fraction that has to be sequenced. Therefore the genomic DNA of interest gets digested by the methylation-insensitive MspI restriction enzyme cleaving

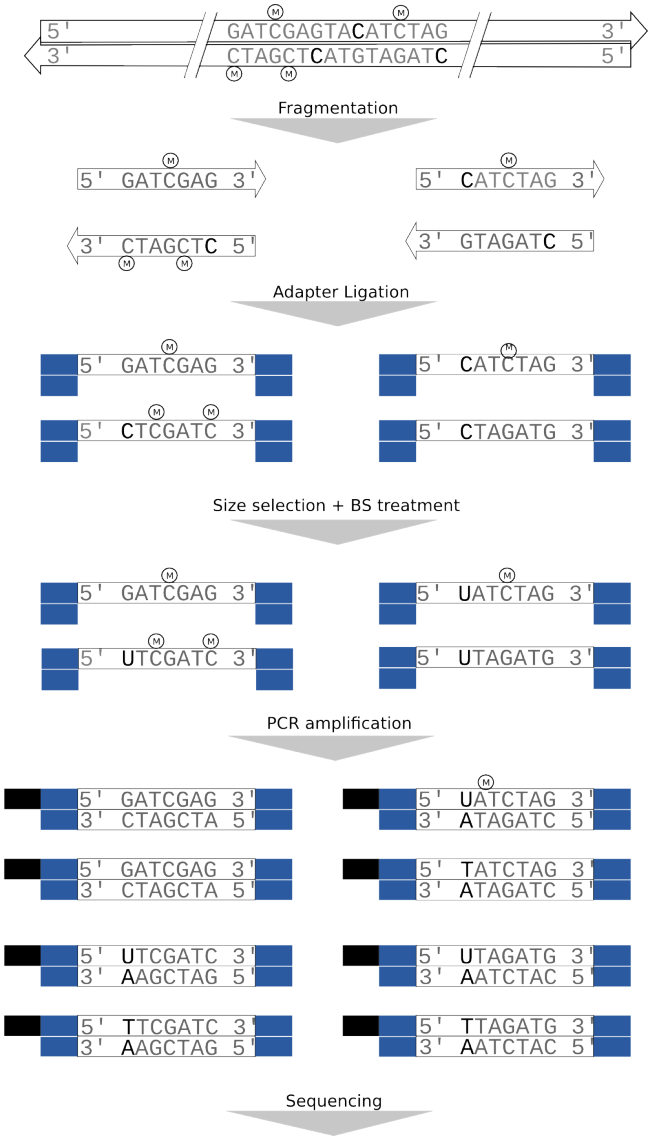


Figure 1.2.2: Directional BS-Seq protocol (Lister protocol [44]): Double-stranded asymmetric adapters are ligated to the single-stranded fragmented genomic DNA. Gel electrophoresis is applied for size selection. Bisulfite treatment now causes the conversion of unmethylated Cs to Us. Subsequent PCR with primers, which are complementary to the ligated adapter sequences, amplifies the fragmented DNA and yields to double-stranded DNA. Due to the asymmetric adapters it is now possible to ligate another set of adapters to the strands which are coming from the original top or bottom strand. Single-end reads can be sequenced now from fragments originating from the original strands.

phosphodiester bonds of the DNA upstream of CpG dinucleotides. This combined with a size selection leads to an enrichment of fragments with CpG content, able to reduce the genome fraction to be sequenced, for example, up to 1% [49]. If the main interest lies in CG context methylations and completeness is not required, this might be a cost-effective alternative to whole genome bisulfite sequencing (WGBS). One key aspect of this approach is that the coverage of the sequenced genomic fraction can be significantly higher than in whole genome sequencing protocols. Thus coverages greater than 100 are not uncommon [25].

1.2.5 Genomic variations

SNPs Contrary to the bisulfite treatment scenario, in natura methylated Cs have a higher probability to convert into Ts than unmethylated Cs (see Section 1.2.1). Due to this reaction, C>T single-nucleotide polymorphisms (SNPs) are with 65% [46] the most frequent of all possible SNPs in mammalian genomes. Especially in the analysis of bisulfite data, where C>T converted bases are measured, this can have a big influence. SNPs can be mistakenly interpreted as bisulfite conversions.

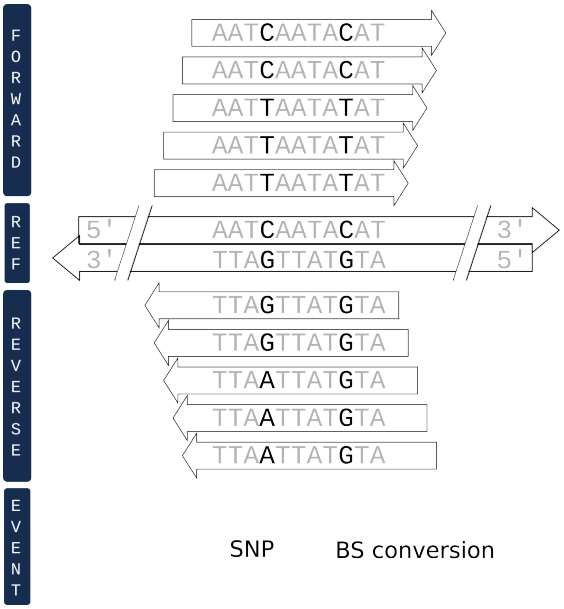


Figure 1.2.3: Distinction between heterozygous C>T SNP and bisulfite C>T conversions given reads sequenced with the directional BS-Seq protokol (displayed only for single-end data): SNPs can be observed in reads originating from both the top and the bottom strand. Top strand bisulfite C>T conversions only occur in reads originating from the top strand, while reads originating from the bottom strand contain Gs.

In diploid genomes with two copies of the genome, heterozygous SNPs (in one allele) and homozygous SNPs (in both alleles) can occur. Heterozygous C>T SNPs might be

interpreted as fuzzy methylation levels and could lead to wrong conclusions. On the other hand, such heterozygous SNPs might change the methylation patterns allele-specific and are of high interest. Such allele-specific DNA methylations (ASMs) play an important role since they are linked to a range of diseases [57]. ASMs can be further caused by genomic indels (insertions or deletions) or near by located SNPs changing the context and thus preventing DNA methylation [57].

Even though SNPs cause conversions in both DNA strands, whereas bisulfite conversions only occur in one strand, the distinction of such C>T SNPs from bisulfite C>T conversions is not trivial. Especially given data produced with the non-directional protocol, because this might produce reads of unknown origin (see Section 1.2.3). Directional reads have the advantage of providing unambiguous information for both strands individually. Given single-end reads mapping against the forward and the reverse reference strand, the former contain information about methylations on the top and the latter about methylations on the bottom strand. Thus additionally to the observed C>T conversion on the top strand, the information about the bottom strand can be used to confirm or not confirm a SNP (see Figure 1.2.3). Paired-end data hold the same information, but with a different mapping.

Not only C>T SNPs have an individual frequency, also other genomic substitutions are uniquely distributed due to different chemical properties. As one example, substitutions between purine nucleotides (A, G) and pyrimidine nucleotides (C, T) – also called transitions – are more likely than transversions, where two purine nucleotides or two pyrimidine nucleotides are substituted.

Indels Genomic deletions or insertions (indels) cause gaps in the read alignments and complicate the read mapping process. Moreover, indels give rise to ambiguously mapping bases flanking this gaps. Figure 1.2.4 illustrates an example with multiple sequences containing indels aligned against a reference sequence. Computing accurate alignments in such cases is particularly challenging in combination with mismatching bases caused by SNPs or sequencing errors. For BS-Seq data this fuzziness is increased further due to bisulfite C>T conversions.

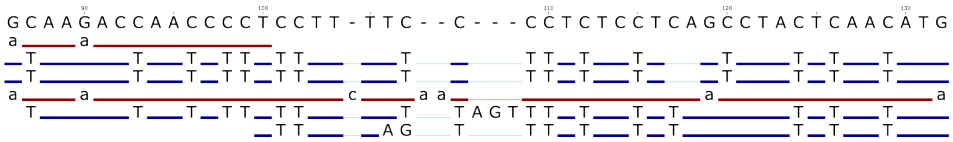


Figure 1.2.4: Multiple sequence alignment (MSA) for bisulfite reads containing indels. The blue and red lines represent reads mapped against the top and the bottom strand respectively. Bases, deviating from the reference base are separately displayed (including bisulfite conversions). Gaps in the alignment are denoted with transparent lines.

1.2.6 Sequencing errors

One challenge for Next-Generation Sequencing tools are errors introduced into the read during the sequencing process. This can be either wrong base-calls or additional introduced or missing bases. Subsequent analyses, like read mapping or variant calling, have to take this possibility into consideration. Recent studies revealed that these sequencing errors in high-throughput sequencing techniques such as Illumina are not equally distributed over the whole set of possible error types [16] [50]. Different analyses showed significant biases in sequencing errors dependent on the genomic and erroneous base type [16] [50]. Altogether, twelve different substitution errors are possible, but analyses of an eukaryotic data set [16] revealed an error frequency of only 2% for C>G substitutions, whereas the substitution error T>C occurs in 15% of the cases. C>T substitution errors again occur low-frequently with only 4%. A similar strong bias holds for indel errors, where insertion as well as deletion errors of A and T nucleotides occur up to 6 times more often [50] than indel errors of the nucleotides C and G.

Due to C>T conversions bisulfite sequencing data contains a higher fraction of Ts (and As in the case of paired-end or non-directional sequencing data) than non-bisulfite sequencing data. We expect that these non-uniform background frequencies combined with non-uniform sequencing errors can cause significant biases in the final methylation analysis results, if not taken into account. However, NGS sequencing technologies often provide base-call qualities assigned to each read base, representing the probability of the call being wrong. Such base qualities are phred-scaled and given by:

$$Q = -10 \log_{10} P(\text{call is wrong}). \tag{1.1}$$

This quality information is stored together with the read sequence in the FASTQ format and can be used in the downstream analysis to prevent erroneous bases from biasing the result.

1.3 Related Work

In the following chapter we describe the required steps for the analysis of BS-Seq data, its computational challenges and the already existing methods. There exists a range of different tools in this field, some of them using similar methods providing similar performances and running times. Thus we will focus on the various approaches and present the most widely used tools for BS-Seq data with single nucleotide resolution, from bisulfite read mapping to methylation level calling. Further, we will describe the main challenges involved in MSA computation and present a short overview of existing methods addressing this task. Finally, we review bisulfite read simulation tools.

Bisulfite read mapping

Clearly, due to the nature of bisulfite reads, mapping them is more challenging than conventional read mapping tasks. The loss of information through bisulfite conversions results in a highly increased search space, where Ts in the read sequence are allowed to map against Cs and Ts of the reference sequence. Furthermore, we don't know how many of those genomic Cs are converted into Ts. There might be reads coming from a highly methylated region, i.e. containing only few bisulfite conversions. The information contained in these Cs allows for a relatively high mapping efficiency. But there might be reads coming from unmethylated regions, i.e. most of the Cs are converted into Ts. This high percentage of Ts can cause ambiguous mappings. In the case of non-directional or paired-end reads, G-A conversions have to additionally be taken into account. Read mapping tools need to address these problems when trying to achieve a good mapping efficiency for highly methylated, highly unmethylated and alternating methylated regions. Two fundamentally different approaches exist.

Wild-card mapping The first one makes use of all given information and simply counts the mapping of a genomic C against a read T as a match, while counting the mapping of a genomic T against a read C as a mismatch. This can be done by using an additional wild-card letter for reference Cs matching against Cs and Ts (see Figure 1.3.1). Wild-card mappings come along with a relatively high sensitivity. The drawback of this method is that the result can be biased, as reads containing a higher percentage of Cs can be mapped with a greater efficiency. For this reason, the overall methylation rates are slightly overestimated.

The probably most common read mapper in this category is BSMAP [63]. It uses a bitwise mapping strategy combined with a hash table seeding. Read Ts are masked as Cs if the corresponding reference position is a C. The masked reads are then divided into seeds – allowing also for real mismatches – and mapped bitwise against the reference sequence using its hash table. The hash table contains all possible bisulfite conversion combinations for each seed. Finally, a simple bitwise operation is used to count real mismatches. BSMAP (v1.2) provides a high sensitivity, but it has a very long runtime compared to other existing tools [6]. The tool RMAPBS [58] is based on the RMAP read mapper and also uses a wild-card mapping implemented in a seed and hashing

procedure, but a bit different from BSMAP. Segemehl is based on enhanced suffix arrays and uses the Myers bit-vector algorithm to extend seeds [51]. In contrast to BSMAP and RMAPBS, Segemehl explicitly takes indels into account.

Another noteworthy tool is Last [23], a read mapper developed originally for non-bisulfite reads. It uses adaptive seeds, makes use of base quality values and allows for indels. The authors of Last implemented an extension for bisulfite reads, using new scores considering C-T conversions and bisulfite specific base abundances. Last uses subset seeds [36], i.e. a reduced alphabet can be used at certain positions in the seed for the mapping. For bisulfite reads a three-letter alphabet is used, with Cs and Ts being treated equally. Different combinations of the three- and four-letter mapping strategy for the seeds result in different sensitivities. The Last authors could show that Last is on various bisulfite datasets significantly faster and at the same time more sensitive and accurate than RMAP and BSMAP [23].

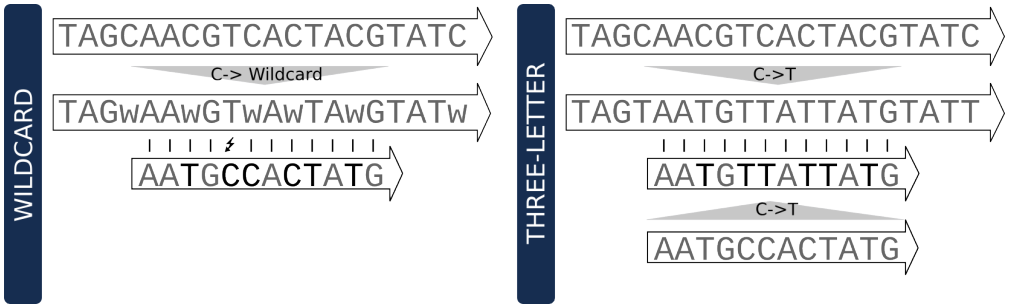


Figure 1.3.1: Wildcard read mapping versus three-letter read mapping: The wildcard mapping allows read Cs and Ts to map against reference Cs, while read Cs mapping against a reference T are counted as mismatches. In the three-letter mapping this information gets lost and real (not bisulfite induced) C-T mismatches are not visible any longer.

Three-letter mapping In the second approach, reads and reference sequences are reduced into a three-letter alphabet, treating Cs and Ts as one letter. As a consequence, the mapping of a genomic T against a read C is counted as a match as well (see Figure 1.3.1). This comes with a slightly reduced mapping efficiency due to the increased search space and more false positive hits. The advantage is that the methylation rate can be assumed to be unbiased. Moreover, tools in this category can benefit from existing and well approved read mappers for the three-letter mapping by adding a pre- and post-processing.

Bismark [38] is probably the most widely used tool in this category, using internally the Bowtie2 read mapper. For the pre-processing it converts reads and reference sequences into a three-letter alphabet. In case of directional single-end reads, all reads are C>T converted. Those are then mapped against a C>T and a G>A converted reference, representing bisulfite conversions on the top and on the bottom strand respectively.

Therefore two instances of Bowtie2 run parallel. For non-directional or paired-end reads Bismark additionally performs G>A conversions in order to cover reads sequenced from the reverse complements of the original reference strands that contain G>A conversions (see Section 1.2.3). Note that no further post-processing steps are applied to filter out reads with a large number of real C-T mismatches.

BS Seeker [7] uses a very similar approach, but discards in an additional post-processing step reads mapping against both the top and the bottom reference strand or those that have too many real C-T mismatches. BS Seeker shows a slightly better running time than Bismark [38]. Both Bismark and BS Seeker support the directional and the non-directional BS-Seq protocol and use Bowtie2 as an internal read mapper. In addition, Bowtie2 takes base qualities for the mapping score of mismatches into account.

The tool MethylCoder [52] is more flexible and able to use either GSNAP or Bowtie2 for the three-letter mapping. Further it uses a similar strategy as Bismark and BS Seeker. In all three tools, the external mapping instances are set to output only unique reads. MethylCoder additionally realigns unmapped reads with their original sequence against the original reference sequence. In this way, reads originating from highly methylated regions that suffer from non-unique three-letter alignments are able to be mapped with a unique four-letter alignment.

In general, the three-letter mapping tools offer a significantly better runtime than the mentioned wild-card mapping tools [6]. An overview of various existing bisulfite read mapping tools is presented in Table 1.1.

Methylation level calling

Some of the previous described mapping tools, such as Bismark and MethylCoder, additionally estimate methylation levels for genomic C positions based on the C-T ratio of the mapped reads. However, this is a rather inaccurate method. The main challenge is that the data can be significantly influenced by sequencing errors, genomic variations, wrong mappings and incomplete bisulfite conversions. In the case of CT genotypes for example, the C-T ratio deviates radically from the true methylation level of the C base.

An advanced approach is implemented in Bis-SNP [46], which takes base qualities into account, identifies SNPs and estimates methylation levels. It is based on a Bayesian approach and makes use of the information about the top and about the bottom DNA strand to distinguish between SNPs and bisulfite conversions. In this way, C>T SNPs are no longer interpreted as unmethylated Cs. Since it is not always known from which strand non-directional reads originate (see Section 1.2.5), Bis-SNP supports only the directional BS-Seq protocol. The weakness of Bis-SNP is that, despite its enhanced model for genotype calling, the methylation levels are simply estimated using the C-T ratio.

Multiple sequence alignment

In general, NGS analyses at single-nucleotide resolution, such as variant calling, critically depend on correctly mapped bases. However, genomic indels can significantly bias the

		non-directional	mapping tool[1]	paired-end	use of base qualities[2]	non-unique three-letter hits	three-letter post-processing[3]	output C-T ratios	runtime
three-letter	Bismark	✓	Bowtie2	✓	✓	✗	✗	✓	-
	BS Seeker	✓	Bowtie2	✗	✓	✗	✓	✗	-
	MethylCoder	✗	Bowtie2/ GSNAP	✓	✓	✗	✓	✓	-/+
wildcard	BSMAP	✓	SOAP	✓	✗			✓	+
	RMAPBS	✓	RMAP	✓	✓			✗	+
	Segemehl	✓		✓	✗			✓	
	BRAT	✓		✓	✗			✓	+
	Last	✗		✓	✓			✗	

Table 1.1: Overview of existing bisulfite read mapping tools and supported features. [1] Segemehl, BRAT and Last are stand-alone read mapping tools. [2] Denotes whether base-qualities are taken into account for the alignment computation or not. [3] Basic post-processing, for example discarding reads with too many non-bisulfite mismatches in four-letter space. [4] Indicator whether runtime is strength or weakness of the tool [6] [23][51], where - denotes a relatively short runtime, whereas + denotes a long runtime. For Segemehl and Last no reliable third-party sources are available.

column-wise alignment accuracy (see Section 1.2.5). For this reason it is common practice in accuracy critical fields to perform a realignment, taking all reads into account that span those indels. This enables the distinction between genomic variations, visible in a high fraction of reads, and sequencing error caused variations, only occurring in one read, both for indels as well as for base substitutions. The information of all reads can then be used to compute a consistent multiple sequence alignment (MSA).

A set of pairwise alignments can simply be converted into an MSA by adjusting the gaps throughout all sequences. However, such simple MSAs are rather inaccurate regarding the column-wise mappings. A more accurate strategy would be to compute the MSA directly by aligning all sequences together.

Bis-SNP/GATK realigner The Bis-SNP method for genotype and methylation level calling mentioned in Section 1.3 is based on the Genome Analysis Tool Kit (GATK)

[15]. Optionally, it performs a realigning step for reads spanning known indels, provided as VCF input [3], prior to the final analysis. This is implemented by using the existing non-bisulfite GATK realigner, adjusted to allow read Ts for mapping against references Cs. In order to limit the time consuming MSA computation, GATK first carefully selects candidate regions. Therefore it synthesizes possible haplotypes using the reference sequence and observed indels, either in the given read set or in the known indel set. The reads are then realigned against these haplotypes, without allowing gaps. The highest scoring haplotype over all reads is chosen and the reads are either assigned to this alternative haplotype or to the reference sequence. Only if the model including the alternative haplotype has a significantly higher likelihood than the original model the actual MSA is computed.

The general problem of computing optimal MSAs is its enormous computational cost. A dynamic programming approach for MSAs is shown to be NP-complete and thus impractical for larger data. This challenge involves heuristic algorithms. One common method is the progressive alignment computation, which makes use of the distance information from all pair-wise alignments, represented as a hierarchical tree, in order to guide the final MSA computation. Similar sequences are aligned first followed by the next similar sequence (or sequence group) corresponding to the guide tree. This approach simulates biological relations between the sequences in order to improve the individual less cost-intensive alignments yielding in the MSA.

Anson-Meyers' ReAligner In our context, we deal with reads from one individual organism, making the progressive strategy needless. Anson and Meyers [1] developed a round-robin realignment strategy that is able to improve MSAs constructed by single pairwise alignments. Iteratively, in each step each sequence is removed one after another from the MSA and realigned until there is no further improvement. This is done by performing a pairwise dynamic programming algorithm that allows one of the sequences to be an MSA. Therefore a specific scoring function is required. In order to align the base a against a given MSA column C, two different scores are combined weighted equally. The first one indicates whether a equals the consensus base, i.e. the most frequent base in C. The second one is the frequency of base a in column C. For the purpose of gaining speed a banded alignment version can be performed, since the rough mapping location is already known.

Bisulfite read simulation

In order to enable a precise benchmarking of the BS-Seq analysis tools respective to sensitivity and specificity, it is crucial to simulate reads with characteristics as realistic as possible. Such characteristics are for example genomic SNPs and indels, sequencing errors and in this context also bisulfite conversions. Many published BS-Seq analysis tools were tested along with rather imprecise read simulations, developed probably only for the purpose of benchmarking mappings and not methylation level calls. In one exam-

ple, first non-bisulfite reads were simulated and after that Cs were randomly converted into Ts dependent on the context [27]. Read mappings can be benchmarked in this way, since reads are mapped independently of each other. However, for the methylation level calling it would be more precise to model SNPs, methylation levels, bisulfite conversions and sequencing errors separately from each other. First, this would prevent simulated wrong base-calls from being additionally converted. And second, this would allow for a more accurate model distinguishing between the methylation probability and the bisulfite conversion rate.

One existing read simulation tool is Sherman [37], able to model reads for the directional and non-directional protocol using context dependent C>T conversion rates. The main drawback is that base substitutions are only applied read-wise, i.e. no genomic SNPs can be simulated.

The software package DNemulator [22] can simulate realistic SNPs and sequencing errors for bisulfite reads. This is done by using real genomic variants for a given reference and empirical error probabilities given as input. The methylation levels on the other hand are randomly assigned, distinguishing only between two contexts (CG on non-CG). Furthermore, the global methylation rates cannot be modified by the user.

BSSim is the most advanced bisulfite read simulator the authors are aware of [47]. It can simulate SNPs for different haplotypes and uses sequencing error and quality distributions. The methylation and bisulfite conversion rates can be specified dependent on context. Both the directional and the non-directional protocols are supported. An apparent disadvantage is that no indels can be simulated.

1.4 Contribution

In this thesis we present a BS-Seq analysis workflow implemented to map and analyse bisulfite reads. The aim is the estimation of accurate methylation levels at single-nucleotide resolution.

We implemented two core modules, a bisulfite read mapper and a subsequent methylation caller, the latter processing the alignments from the former module. In order to interpret the bisulfite data unambiguously, it is crucial to know, whether a specific read is holding the encoded methylation information about the top or the bottom DNA strand. For that reason we decided to focus this workflow on the directional BS-Seq protocol (see Section 1.2.3). Furthermore we decided to support both single-end and paired-end read data to support the current standard of sequencing technologies. Some of the existing tools only analyse CpG methylations or make assumptions on the symmetry of CpG methylations regarding the top and bottom DNA strand. Clearly, this is an open field and non-CpG methylations and asymmetric methylations are worth analysing [62] [60]. Thus, in this workflow, the three contexts CG, CHG and CHH are examined and no assumption is made about symmetric methylations.

Previous research has proved already that the use of base quality values can significantly improve the accuracy of alignments [59] [24]. This discovery was to be expected, since the information content of such reads is increased and biases caused by sequencing errors can be reduced. Accordingly we decided to make use of quality information for both the mapping and the calling module. Beside this, we know that such sequencing errors do not occur with the same frequency for all possible types, but are rather base dependent (see Section 1.2.6). In the Chapters 2 and 3 we will present methods in order to incorporate non-uniform error probabilities for alignment and methylation level computation.

The workflow presented here is implemented using SeqAn [17], which offers already efficient data structures specifically designed for read mapping and comes with a comprehensive alignment module, laying the foundation of this work.

In the following we provide a brief overview of the different parts of this BS-Seq analysis workflow presented in this thesis.

Bisulfite read mapping

In order to analyse the methylation levels at a single-nucleotide resolution, we first need to compute the column-wise accurate alignments between the bisulfite reads and the reference. For the mapping we use the three-letter mapping approach (see Section 1.3), since it proved to be quite efficient while allowing for decent runtimes. This three-letter mapping can be performed using external tools. In Chapter 2 we present the overall mapping method, including the appropriate pre-processing of the given sequences. For the post-processing step we describe an advanced statistical method for realignment and verification in four-letter space.

Methylation level estimation

In Chapter 3 we present a probabilistic model to estimate accurate position-wise methylation levels based on the given alignments. We decided to additionally perform genotype calling, in order to allow for the computation of genotype specific methylation levels and to distinguish between C>T SNPs and C>T bisulfite conversions (see Section 1.2.5). This is particularly important because these C>T SNPs occur highly frequently and thus have an important impact on the overall methylation level accuracy. Our method is based on the Bayesian approach implemented in Bis-SNP [46] (see section 1.3) for combined methylation and SNP calling. We present an extended method to optimize these methylation levels and to prevent non-uniform sequencing error frequencies from biasing the result.

Local multiple sequence realigning

We decided to additionally include a module for multiple sequence realignment into our BS-Seq analysis workflow, in order to reduce the bias caused by indels. Since our approach is designed for de-novo analysis, able to discover genomic variants independently of known sites, a rather high fraction of indel candidate regions needs to be examined. For this reason, the GATK multiple sequence realignment method implemented in Bis-SNP (see Section 1.3) is rather impractical due to its expensive MSA computation.

For this reason we base our realignment method on the basic principle of the Anson-Meyers' heuristic (see Section 1.3). In each step one read is first removed and then realigned against the remaining MSA, until the alignment score does not further improve. The aim is a column-wise accurate and consistent MSA in order to improve the subsequent analysis of genomic SNPs and position specific methylation levels. In Chapter 4 we present the detailed method.

Annotation mapping

Since methylation rates highly depend on the genomic region and its function, the idea is to optionally map the position-wise estimated methylation levels to genomic annotations. Region specific methylation rates are of high interest, for example for the analysis of cancer methylomes. Tumour-suppressor inactivations are often caused by the hypermethylation of its promoter region. A region-wise mapping could allow a first insight and the detection of regions that may be worth investigating further. The detailed method is described in Chapter 5.

Bisulfite read simulation

The first step prior to the actual analysis is to simulate bisulfite reads to enable an exact implementation and benchmarking process. Due to the incompleteness of the existing simulation tools (see Section 1.3), we decided to implement our own bisulfite read simulation tool. The presented workflow is designed to deal with small indels, so that even the BSSim simulation tool would be rather impractical. We based our method

on the SeqAn tool Mason [30] which uses an advanced method for simulating non-bisulfite reads for different sequencing platforms. Mason is able to simulate genomic SNPs and indels and further applies model specific sequencing errors and base quality values [30]. Chapter 6 describes the detailed method to simulate bisulfite reads representing realistic bisulfite conversions and sequencing errors.

In the following we will present the core modules for the SeqAn bisulfite workflow in detail. An overview is presented in Figure 1.4.1.

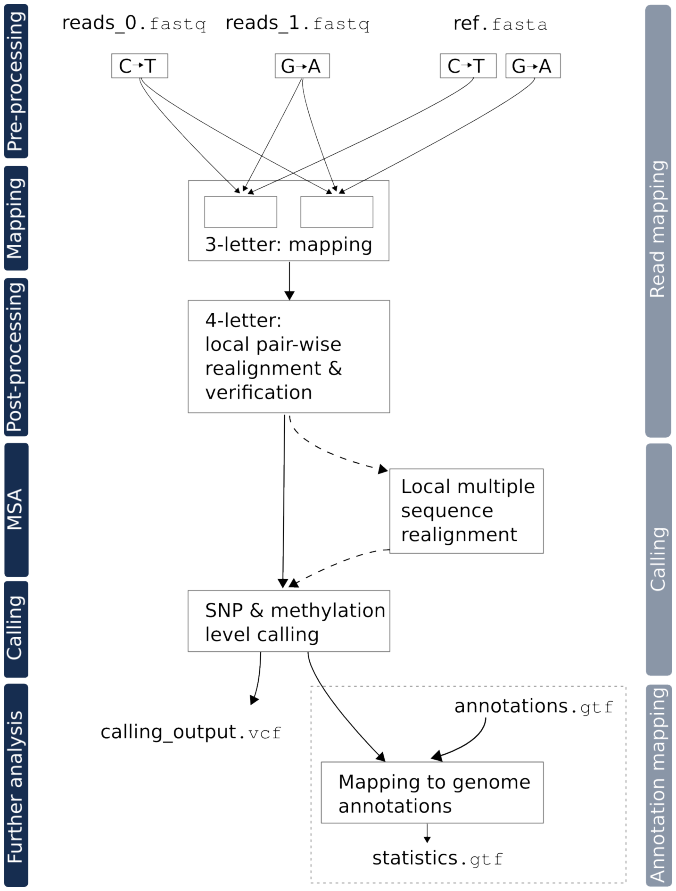


Figure 1.4.1: BS-Seq analysis workflow: From read mapping to methylation level calling.

2 Bisulfite Read Mapping

In this section we will describe our method to address the bisulfite read mapping challenge, which lays the foundation for the whole analysis workflow. Wrongly mapped reads would cause wrong methylation callings in the subsequent step. Moreover it is not only important to achieve a precise mapping, i.e. to find the true genomic location of the read, it is also essential for the calling to compute an column-wise accurate alignment. In other words, we need to find the precise mapping of each single read base, which is particularly important if the alignment contains genomic or sequencing caused indels.

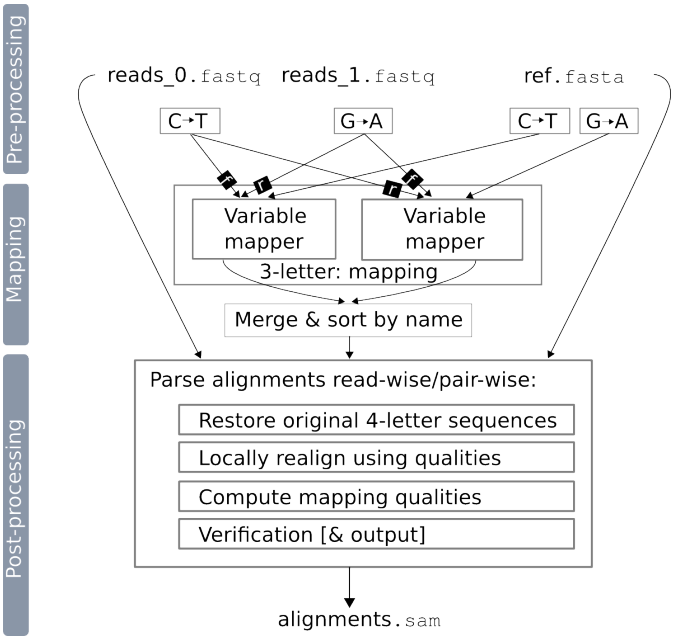


Figure 2.0.1: Bisulfite read mapping: Reads and reference sequences are converted into three-letter sequences. Left reads (reads_0.fastq) are mapped against the forward strand of the C-T converted references and against the reverse strand of the G-A converted reference. Right reads are treated the other way round. Afterwards the SAM files resulting from the two mapping instances are merged and sorted by read name. Finally this SAM file together with the original sequence files is used in the post-processing step for the local realignment and verification.

In order to meet these requirements, we first map the reads with a simple three-letter mapping approach to find candidate regions, followed by a local realignment and

verification process. For the latter we convert the sequences back into the four-letter alphabet and apply an advanced statistical model to compute a column-wise accurate alignment. This is of great importance, especially since in the three-letter alignment the base mappings C-T, T-T, C-C and T-C are all treated equally. Afterwards the scores can be used to filter out low quality and – in the case of reads mapping to multiple locations – non-optimal alignments. Additionally, mapping qualities are computed based on the scores of all occurrences of one read, which again has a filter function and can be handed over to the subsequent methylation calling module. An overview of the different steps is presented in Figure 2.0.1.

2.1 Preprocessing

Prior to the actual three-letter mapping, the reads and reference sequences have to be appropriately converted. Remember we are using reads from the directional bisulfite sequencing protocol. In this way we know for single-end reads that they have their origin in one of the original top or bottom strands and thus contain possibly C>T conversions (see Section 1.2.3). For paired-end reads we know that the left reads behave like single-end reads, while the right reads originate from a reverse complement, meaning they contain possibly G>A conversions.

On account of this, first a three-letter version of the reference sequence is built with all Cs converted into Ts, covering possible bisulfite conversions caused by unmethylated Cs in the original top strand. Second, a G>A three-letter version is built, covering possible conversions in the bottom strand.

Next, in the case of single-end reads, all reads are C>T converted, allowing for a mapping against the forward strand of the C-T reference version and against the reverse strand of the G-A reference version. In the case of paired-end reads, left reads are treated the same way as single-end reads. Right reads get G>A converted, such that they can be mapped against the reverse strand of the C-T reference version and against the forward strand of the G-A reference version (see Figure 2.1.1).

2.2 Three-letter mapping

For the three-letter mapping process itself, external read mapping tools can be used in the workflow as long as they provide a valid SAM output, e.g. the SeqAn tool RazerS3 [61]. Due to the reduced alphabet, such mappings result in an increased number of reads mapping to multiple genomic locations compared to conventional read mapping scenarios. Since these multiple hits are subsequently verified to find the best mapping in the 4-letter space, it is strongly recommended to use a mapping tool able to output all mappings above a given error-threshold. Moreover it has to be taken into account that true read-C reference-T mismatches can be hidden behind C>T conversions, adding up to the errors allowed by the read mapper. This again increases the search space and thus the runtime of the three-letter as well as of the subsequent four-letter post-processing module. To avoid this, the allowed error rate for the mapping process should be carefully

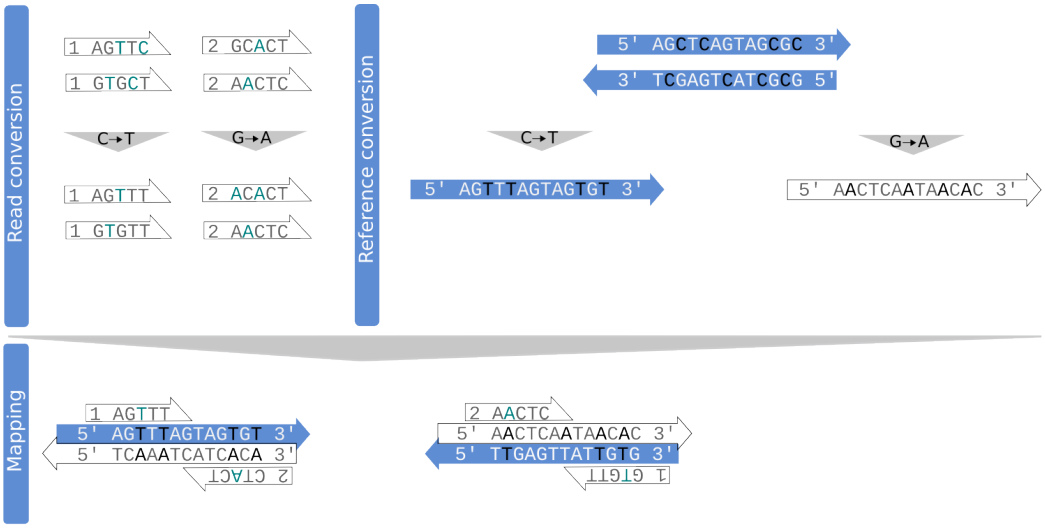


Figure 2.1.1: Three-letter mapping process for paired-end reads: Left reads (1) and right reads (2) are C>T and G>A converted. (Single-end reads would be treated equally to left reads.) The reference sequence is converted to a C-T and a G-A three-letter version. Finally reads are mapped correspondingly.

---

defined as low as possible, while still maintaining the mappability of the reads, taking realistic rates of sequencing errors and genomic variations into account.

2.3 Four-letter local realignment

The raw results obtained in the previous step still contain a high number of wrongly aligned reads and bases. In the following section we present the realignment strategy and its underlying advanced statistical model, aiming to get the best out of the given data. Prior to the actual alignment computation, for each read or read-pair given in the raw alignment results, the four-letter sequence is restored using the original FASTQ read files. Then the three-letter mapping locations are used to perform a semi-global alignment on the read and an infix of the reference sequence at the respective position. In order to improve the accuracy of the alignment, our score function makes use of target frequencies incorporating different SNP, methylation and bisulfite conversion probabilities. Moreover, base-call error probabilities as well as mapping scores of non-observed bases are taken into account corresponding to their probability, based on a model developed by Frith et al. for non-bisulfite reads [24]. Finally we make use of base type dependent sequencing error probabilities for our score computation.

2.3.1 Traditional scoring matrices

Before we describe in detail our bisulfite specific scoring method, we will have a look at how traditional scoring matrices are used to obtain the score for mapping two bases in an alignment. These scoring matrices usually contain positive values for matches and negative values for gaps and mismatches. In a very simple case this could be zero for a match and minus one for a gap or mismatch. For a more accurate statistical computation, log likelihood ratios are used [34]. Assume we want to compute our scoring matrix S with S_{xy} defining the score for mapping the bases x and y. The likelihood ratio R_{xy} of two bases is the ratio of the probability of observing base y aligned to base x in our model $\mathcal{A}$ to the joint probability of these bases, assuming they are independently of each other:

$$R_{xy} = \frac{P(xy|\mathcal{A})}{P(x)P(y)}. \tag{2.1}$$

As an example, $P(xy|\mathcal{A})$ can be the probability of a genomic substitution of base x to base y. For $P(x)$ and $P(y)$ simply background frequencies of the bases can be used. In this way S_{xy} can be computed as follows, given an arbitrary scaling factor T:

$$S_{xy} = T \cdot log(R_{xy}). \tag{2.2}$$

The target frequencies $P(xy|\mathcal{A})$ can be modelled given substitution probabilities, e.g. based on the different probabilities of genomic transitions and transversions [24](see section 1.2.5). Target frequencies always sum up to one. Such log likelihood ratios serve as an accurate method to build scoring matrices, if the background and substitution frequencies are known. Assume we know the overall substitution probability in our model, which is denoted by φ, then the match probability is $1 - \varphi$. Let parameter ω be the probability of a substitution to be a transition. Given the total number of possible matches (4), transitions (4) and transversions (8), we can now compute the target frequencies as follows:

$$P(xy|\mathcal{A}) = \begin{cases} (1-\varphi) \cdot 1/4 & \text{if } x = y \\ \varphi\omega \cdot 1/4 & \text{if } \{x, y\} \in \{\{A, G\}, \{C, T\}\} \\ \varphi(1-\omega) \cdot 1/8 & \text{else.} \end{cases} \tag{2.3}$$

2.3.2 Bisulfite conversion aware scoring matrices

The described traditional scoring matrices using log likelihood ratios build the foundation of our method. Due to simplicity reasons, in the following we will describe the model only for reads containing C>T bisulfite conversions. The strategy for paired-end reads is equivalent and mentioned at the end of this chapter (see Section 2.5).

Bisulfite reads show different characteristics compared to conventional reads caused by bisulfite conversions and we can model this in our scoring function with corresponding

target and background frequencies. These lead to asymmetric scoring matrices, so that in the following we denote with $P(xy|\mathcal{A}_{BS})$ the probability to align the read base y to the reference base x under our bisulfite model $\mathcal{A}_{BS}$. For the mapping of base types not involved in bisulfite conversions, the model stays conform to the traditional one. For read Cs or Ts mapped against a reference C, we extended the model by methylation and bisulfite conversion probabilities. Let the parameter β denote the global methylation probability of a genomic C. The parameter α denotes the conversion probability of an unmethylated genomic C under bisulfite treatment. Given this, we can compute the target frequency of mapping a read base T against a reference C by adding the probability that this T is caused by a bisulfite conversion – and in reality represents a match – to the regular transition probability:

$$P(CT|\mathcal{A}_{BS}) = \varphi\omega \cdot 1/4 + (1-\varphi) \cdot 1/4 \cdot (1-\beta)\alpha, \tag{2.4}$$

with

$$\begin{aligned} \varphi\omega \cdot 1/4 &: \text{regular transition probability (C>T)} \\ (1-\varphi) \cdot 1/4 &: \text{match probability (C-C)} \\ (1-\beta)\alpha &: \text{probability that genomic C is unmethylated and converted to T.} \end{aligned} \tag{2.5}$$

The mapping of a read base C against a reference C can be explained by a genomic C, which is methylated and thus not converted, or by an unmethylated C, which is not converted. This can be represented by the target frequency:

$$P(CC|\mathcal{A}_{BS}) = (1-\varphi) \cdot 1/4 \cdot (\beta + (1-\beta)(1-\alpha)), \tag{2.6}$$

with

$$\begin{aligned} (1-\varphi) \cdot 1/4 &: \text{match probability (C-C)} \\ \beta &: \text{methylation probability (C is not converted)} \\ (1-\beta)(1-\alpha) &: \text{probability that genomic C is unmethylated and not converted.} \end{aligned} \tag{2.7}$$

Besides the target frequencies, also the read background frequencies have to be adjusted due to the bisulfite conversions. Therefore the background frequencies of the read bases $P(x_{read})$ involved in bisulfite conversions are estimated based on the original genomic background frequencies $P(x_{ref})$, the global methylation probability β and bisulfite conversion probability α:

$$P(C_{read}) = P(C_{ref}) - P(C_{ref})(1-\beta)\alpha \tag{2.8}$$

$$P(T_{read}) = P(T_{ref}) + P(C_{ref})(1-\beta)\alpha. \tag{2.9}$$

2.3.3 Incorporating base qualities by Frith et al.

So far we have seen how to compute accurate scores for aligning two bases under the assumption that both observed bases are correct. We will now describe a method, implemented in the read mapping tool LAST [24] for conventional read mapping, which additionally takes sequencing error probabilities into account. In this context the authors could show that the incorporating of base qualities into the score function does significantly improve the read mapping. The likelihood ratio R_{xy} is generalized to allow for base-calling probabilities instead of assuming a correctly observed base. Given observed data d obtained from the sequencing process itself (e.g. image intensities), this can be interpreted as the probability $P(y|d)$ that a certain base y is the true base. Thus a new likelihood ratio R' can be computed by:

$$R'_{xd} = \frac{P(xd|\mathcal{A})}{P(x)P(d)}. \tag{2.10}$$

Applying the Bayes formula and some rearrangements, this can be proven [24] to be equal to:

$$R'_{xd} = \sum_{y \in \{A,C,G,T\}} \frac{P(xy|\mathcal{A})}{P(x)P(y)} \cdot P(y|d). \tag{2.11}$$

In other words, the new likelihood ratio of aligning a base with quality data d to the reference base x is the sum of the original likelihood ratios for all possible bases weighted with the corresponding base probability. Since we deal with FASTQ quality values and not with image intensities, $P(y|d)$ has to be computed based on the assigned Phred quality value q [24] of the current base. The error probability ε of base y given q is:

$$\varepsilon = 10^{\frac{-q}{10}}. \tag{2.12}$$

The probability that the true base is y given a quality value q and an observed base a can now be computed using ε. Phred quality values contain no information about the three possible non-observed bases and in the Last method they are assumed to have uniform distributions:

$$P(y|q,a) = \begin{cases} 1-\varepsilon & \text{if } y == a \\ \varepsilon/3 & \text{if } y \neq a. \end{cases} \tag{2.13}$$

We decided to work with this method for our bisulfite alignment problem, since it efficiently makes use of the per base given quality data and prevents poorly called bases from causing biases. Moreover it takes the original likelihood ratios of the non-observed bases into account correspondingly to their probability. Especially in the case of bisulfite reads with an increased ambiguity due to conversions, this could be proved to be a powerful method to improve the column-wise accuracy of alignments.

2.3.4 Extension for base dependent sequencing substitution errors

The previous presented generalized log likelihood ratio computation assumes uniform sequencing error distributions over all possible bases. However, several studies proved that sequencing error frequencies are biased for all types of errors (see Section 1.2.6). Considering for example the mapping of a read C against a reference T, knowing T>C substitution errors are the most frequent ones, this information can be used to increase the weight of the corresponding – relatively high scoring – likelihood ratio R_{TT} to the new score.

We extended the previous generalized likelihood ratio computation (see 2.11) to take base dependent sequencing error probabilities into account:

$$R''_{xa} = \frac{P(xa|\mathcal{A}_{\mathcal{BS}})}{P(x_{\text{ref}})P(a_{\text{read}})} \cdot (1-\varepsilon) + \sum_{b \neq a} \frac{P(xb|\mathcal{A}_{\mathcal{BS}})}{P(x_{\text{ref}})P(b_{\text{read}})} \cdot \varepsilon \cdot P(b|a). \tag{2.14}$$

where $P(b|a)$ represents the probability that the true base is b under the assumption that the observed base a is caused by a substitution error.

2.3.5 Extension for indels

DNA sequencing techniques introduce not only substitution errors, but base deletion and insertion errors with a frequency dependent on the genomic and inserted base respectively (see ection 1.2.6). However, the methods mentioned so far are only defined for substitution errors. In order to provide an accurate and complete model, we additionally take such indel probabilities for the computation of alignment gap scores into account. In this context it is crucial to differentiate between genomic indels and sequencing indel errors. Moreover, due to missing quality values at read gap positions, scores for reference gaps and read gaps have to be modelled separately.

First we will describe our method regarding the score for mapping a read base a with a given quality value against a gap '−' in the reference. This case can be caused by mainly three different scenarios:

1. A genomic insertion of the base a occurred (and no sequencing error).
2. A sequencing insertion error of base a occurred.
3. A genomic insertion of a base different to a occurred, which was erroneously sequenced as an a.

The likelihood ratio for this mapping is denoted with R''_{-a} and can be computed based on these three scenarios as follows:

$$R''_{-a} = \frac{P(-a|\mathcal{A}_{\mathcal{BS}})}{P(-_{\text{ref}})P(a_{\text{read}})} \cdot (1-\varepsilon) + \varepsilon \cdot P(-|a) + \sum_{b \neq a} \frac{P(-b|\mathcal{A}_{\mathcal{BS}})}{P(-_{\text{ref}})P(b_{\text{read}})} \cdot \varepsilon \cdot P(b|a), \tag{2.15}$$

where $P(-|a)$ is the base dependent sequencing insertion error probability. The assigned likelihood ratio for insertion errors is one. The computation of the likelihood ratio for aligning the read base a against a reference gap depends on the target and the background frequency of genomic base deletions. Since it is often difficult to know these values in advance, we replace this likelihood ratio with a simple gap penalty score. Therefore we make use of the following formula derived from equation 2.2:

$$R_{xy} = e^{S_{xy}/T}. \tag{2.16}$$

Given a simple gap score S_{-y} we can now approximate R''_{-a} with:

$$R''_{-a} \approx e^{S_{-y}/T} \cdot (1-\varepsilon) + \varepsilon \cdot P(a|-) + \sum_{b \neq a} e^{S_{-y}/T} \cdot \varepsilon \cdot P(b|a). \tag{2.17}$$

Thus genomic insertions are modelled base independent, while insertion errors are modelled base dependent. Note that S_{-y} needs to be proportional to the underlying simple match and mismatch scores S_{xy}, indirectly given in equation 2.14 by the likelihood ratios. Unsuitable scores would cause globally wrong weighted gaps.

In the second case we model the score for mapping a read gap against a reference base x. The drawback here is that no quality value is given due to the missing read base. One possible strategy might be to use the neighbour base qualities to estimate a quality value indicating whether this case is caused by a genomic deletion or by a sequencing error. However, this is not trivial to estimate, since an accurate model describing the influence of an insertion error on the neighbour quality values would be needed. Therefore we decided to approximate the error probability by using a global deletion error rate ϕ. There exist three scenarios able to cause the mapping of a read gap against a reference base x:

1. A genomic deletion occurred (and no sequencing error).
2. A deletion error of base x occurred.
3. A genomic substitution to a base different than x occurred and a sequencing deletion error removed this base.

All three cases are taken into account in the following formula:

$$R''_{x-} \approx e^{S_{-y}/\lambda}(1-\phi) + \sum_{b \in \{A,C,G,T\}} \frac{P(xb|\mathcal{A}_{\mathcal{BS}})}{P(x_{\text{ref}})P(b_{\text{read}})}(\phi \cdot P(b|-) \cdot F), \tag{2.18}$$

with $P(b|-)$ being the base dependent deletion error probability and F being a scaling factor to maintain the global deletion error rate. Especially in the bisulfite case with a significantly higher background frequency of base T combined with a relatively high deletion error probability of base T the scaling is crucial. F is consequently dependent on the various background and the deletion error probabilities.

2.3.6 Final alignment score computation

Given the mapping and gap scoring functions based on the likelihood ratio R''_{xy} with $x, y \in \{A, C, G, T, -\}$, the alignment score S_A for a read aligned against a reference sequence can be computed. Let A be the global alignment of the reference subsequence and the read sequence given the alignment rows a_1 and a_2 of length n respectively. For the alignment score the following holds:

$$S_A = T \cdot \log\left(\frac{P(a_1 a_2|\mathcal{A}_{\mathcal{BS}})}{P(a_1)P(a_2)}\right) \tag{2.19}$$

$$= T \cdot \log\left(\prod_i^n \frac{P(a_{1i} a_{2i}|\mathcal{A}_{\mathcal{BS}})}{P(a_{1i})P(a_{2i})}\right) \tag{2.20}$$

$$= T \cdot \sum_i^n \log\left(\frac{P(a_{1i} a_{2i}|\mathcal{A}_{\mathcal{BS}})}{P(a_{1i})P(a_{2i})}\right) \tag{2.21}$$

$$= T \cdot \sum_i^n \log\left(R_{a_{1i}a_{2i}}\right) \tag{2.22}$$

Given quality data for the read sequence and using our extended method we obtain:

$$S_A = T \cdot \sum_i^n \log\left(R''_{a_{1i}a_{2i}}\right) \tag{2.23}$$

For the actual alignment computation the sum of the log likelihood ratios is maximized.

2.4 Verification

As the three-letter mapping finds all hits below a certain error threshold, each read can now have multiple recomputed alignments with given scores and error rates corresponding to the four-letter space. For the downstream analysis though, this ambiguity needs to be resolved keeping the best – and most likely true – mapping. Additionally, low quality mappings might need to be discarded given user defined thresholds. The scores themselves are already a good measurement for the quality of a mapping and could be used for verification, beside the error rates. Nevertheless, if a read has multiple high scoring alignments at different genomic locations, this measurement is rather poor, since it does not take this uniqueness into account.

2.4.1 MAQ Mapping qualities

A more efficient measurement, taking all information of one read into account, is the mapping quality used in the read mapping tool MAQ introduced by Li et al. [41]. The quality Q_M of a mapping M is – equivalent to base qualities – the phred-scaled probability that the read is mapped wrong:

$$Q_M = -10 \cdot \log P(\text{read mapping is wrong}) \tag{2.24}$$
$$= -10 \cdot \log\left(1 - P(\text{read mapping is right})\right) \tag{2.25}$$

The MAQ mapping quality computation makes use of the fact that the probability of a read coming from a certain genomic location can be expressed by the product of error probabilities of bases at mismatch positions in the alignment [41]. The more mismatches, the lower the probability that the read is originating from this location. But the higher the error probabilities at the mismatch positions, the higher the probability that the mapping is true and the mismatch is caused by a sequencing error. Due to this definition, this method is restricted to ungapped alignments.

The naive way of computing the posterior probability that a given genomic location is the origin of a read is to take all other possible mapping locations into account. Since the computation of all possible mapping locations would be highly expensive, the MAQ authors developed an approximation for Q_M [41]. It takes different mapping error types into account. If a read mapper is used, which guarantees to find all possible mapping locations given a certain error threshold, then the MAQ mapping quality computation can be reduced to this formula [41]:

$$Q_M = q_2 - q_1 - c \cdot \log n_2, \tag{2.26}$$

where q_1 is the sum of quality values at mismatch positions of the best mapping location and q_2 correspondingly to the second best mapping location. The parameter n_2 is the number of mappings having the same number of mismatches as the second best hit and c is a scaling constant. The parameters q_1 and q_2 are based on the probability of the mapping being right:

$$\begin{aligned} q_1 &= -c \cdot \log P(\text{mapping is right}) \\ &\approx -c \cdot \log\left(\prod_{i \in MM} \varepsilon_i\right) \\ &= \sum_{i \in MM} -c \cdot \log \varepsilon_i \\ &= \sum_{i \in MM} q_i, \end{aligned} \tag{2.27}$$

where MM gives the set of mismatch positions. Thus the greater q_2, i.e. more high-quality mismatches, and smaller q_1, i.e. less mismatches and lower qualities, the higher the mapping quality. Additionally it holds that the higher the number of hits with the second best score, the lower the mapping quality.

2.4.2 Generalized mapping qualities

Real world sequencing reads not always map to the reference genome without gaps, since reads might contain genomic as well as sequencing error indels. For this reason the MAQ mapping quality definition can be rather inaccurate for some data. Moreover, only information from mismatch positions is taken into account, even though high quality match positions indicate a greater probability that the mapping is right than low quality match positions. Since it is essential for our method to allow for indels and compute an accurate mapping quality, we adjusted the MAQ method in order to use our previously described alignment score:

$$q_1' = -c' \cdot \sum_i^n \log\left(R''_{a_{1i}a_{2i}}\right), \tag{2.28}$$

where c' is again a scaling constant dependent on c and T (see equation 2.23).

Let S_{A_1} and S_{A_2} be the alignment scores of the best and second best alignment, then the new mapping quality is given by:

$$Q'_M = (-c' \cdot S_{A_2}) - (-c' \cdot S_{A_1}) - c \cdot \log n_2. \tag{2.29}$$

Using alignment scores comes with the advantage that the information about the whole alignment – and thus about matches, mismatches and gaps – is taken into account.

Note that in this module only the hits reported by the three-letter read mapping tool are analysed. Therefore, possible mappings containing errors above a certain threshold are not considered. Three different mapping scenarios can occur for single-end reads:

1. The read has a unique hit, i.e. exactly one best hit, and at least one second best hit.
2. The read has a non-unique hit, i.e. more than one best mapping.
3. The read has only one hit (reported from the read mapper).

In the first case of a unique hit, equation 2.29 is applied to compute the mapping quality. In the second case the mapping quality is set to zero, since no information is available about the true mapping location. The mapping in the third case clearly should get a relatively high quality assigned, since all other possible hits were discarded in the upstream read mapping process. In order to compute the mapping quality, the error threshold used for the three-letter mapping is now used to compute the score of the best theoretically possible second best mapping. Therefore mismatches are assumed to be at high quality positions (avoiding the construction of too high mapping qualities, possibly even higher than of the reported best hit), while all other positions are assumed to contain matches. No further second best hits are assumed for this case. This method

provides a mapping quality that is – despite its limitation – more accurate than the original MAQ qualities.

Paired-end read mapping qualities are based on the qualities of the individual read mappings. Two scenarios have to be distinguished:

1. The read pair has a unique pair hit, while the individual read mappings are not necessarily unique.
2. The read pair has a non-unique pair hit.

In the first case both reads get the sum of the individual mapping qualities assigned:

$$Q_{M_L} = Q_{M_R} = Q_{M_L} + Q_{M_R}, \tag{2.30}$$

with L and R denoting the left and right read of the pair.

In the second case each individual read gets its own individual mapping quality assigned:

$$\begin{aligned} Q_{M_L} &= Q_{M_L} \\ Q_{M_R} &= Q_{M_R}. \end{aligned} \tag{2.31}$$

This assigns a lower quality to non-unique pair hits – compared to unique pair hits – and maintains the information about the uniqueness of the individual reads.

2.4.3 Verified alignment output

Out of all given alignments, only those that are unique regarding the four-letter space are written to the SAM file. Optionally, the results are filtered based on user defined thresholds regarding the mapping quality, the alignment score or the error rate.

2.5 Implementation

The previously described realignment and verification step is done for each read or read pair independently. This allows a read- or pair-wise parsing and post processing of the three-letter alignments given in the (by queryname sorted) SAM file, keeping the memory consumption of this module fairly low. The semi-global alignment computation is based on the existing SeqAn alignment module, which provides an efficient and flexible alignment interface. We implemented new scoring classes and functions in order to materialize the presented realignment method using quality values. Reads with a mapping against a reverse strand of the reference – either of the G-A or of the C-T reference version (see Section 2.1) – are internally projected to the forward strand for the alignment computation. Accordingly, reads containing the information about the bottom DNA

strand get correspondingly adjusted scoring schemes assigned, i.e. allowing for bisulfite G>A conversions instead of C>T conversions. Moreover, specific scoring schemes are used dependent on the original mapping direction, ensuring that the used sequencing error probabilities always refer to the original sequenced bases. Beside that, the score for mapping two bases depends on the reference base, the read base and its assigned quality value. Since such quality values are located within a limited range of integer values [1], scoring tables covering all scenarios are precomputed in order to decrease the runtime. For the alignment computation itself the mapping location is already known, so we used a banded alignment function to lower the complexity.

Our method for this module is dependent on various parameters, such as the methylation probability, the bisulfite conversion probability, indel rates, simple gap scores and sequencing error probabilities among others, that can be defined by the user. If required, a uniform model can be switched on for the target frequencies assuming uniform distributed genomic substitutions, while maintaining the bisulfite specific properties. The equivalent applies for the sequencing errors.

[1] The range of quality values is dependent on the sequencing technology. Recent Illumina quality values range from 0 to 41.

3 SNP and Methylation Level Calling

In the previous chapter we presented a method for the computation of column-wise accurate alignments of bisulfite reads. In the following chapter we describe how to use this information for precise methylation level calling.

The aim of this method is to take into account the information about all mapped bases for each genomic position, while minimizing the bias caused by sequencing errors, wrongly mapped reads and genomic variations. To address this challenge, we apply not only call methylation levels, but also call genotypes, based on a Bayesian model implemented in Bis-SNP. This model takes into account base qualities and possible bisulfite conversions. We extended the method to find the methylation level explaining the data best. Non-uniform sequencing error probabilities and mapping qualities are additionally used to prevent erroneous base-calls or mappings from biasing the result.

Prior to the position-wise analysis, the given pairwise alignments are parsed and converted into a global multiple alignment to enable a column-wise access.

3.1 Bayesian model

In this section we describe in detail the underlying Bayesian approach used in our method. It models for each genomic position the posterior probabilities of all diploid genotypes given the observed mappings and a variable methylation level, that has to be determined.

How to efficiently determine the optimal methylation level will be presented in Section 3.2.

3.1.1 Bis-SNP by Liu et al.

Our main idea is based on the Bis-SNP model [46], which in turn uses the Genome Analysis Tool Kit (GATK), a framework for variation discovery and genotyping [15]. In short, Bis-SNP reads in the alignments given by a bisulfite read mapper, optionally realigns them locally and performs SNP calling using a Bayesian model. All ten possible diploid genotypes are considered: AA, AC, AG, AT, CC, CG, CT, GG, GT, TT. The posterior probability for each possible genotype for each genomic position is computed given the read bases mapped at this position

For the final SNP calling the genotype with the maximum probability is chosen.

The posterior probability of genotype G under the observed data D is computed as follows:

$$P(G|D) = \frac{\pi(G)P(D|G)}{P(D)}, \tag{3.1}$$

where $\pi(G)$ is the prior probability of this genotype under the reference base, obtained from public SNP data. $P(D|G)$ is the likelihood to observe exactly this data under genotype G. $P(D)$ is the general likelihood to observe this data, given by the following sum over all possible genotypes:

$$P(D) = \sum_{G} \pi(G) P(D|G). \tag{3.2}$$

In order to calculate the posterior probability of genotype G, first the likelihood $P(D|G)$ needs to be determined. This is done by summing up the individual likelihoods over all reads to observe the respective base under the given genotype:

$$P(D|G) = \prod_{j=1}^{n} P(D_j|G), \tag{3.3}$$

with n being the number of reads mapped at the current position. The individual base likelihoods in turn can be calculated by taking the different haplotypes H_1 and H_2 equally weighted into account:

$$P(D_j|G = XY) = \frac{1}{2} P(D_j|H_1 = X) + \frac{1}{2} P(D_j|H_2 = Y). \tag{3.4}$$

Now, given an observed base and its quality value, the likelihood to observe this base under haplotype H can be computed. For the haplotype $H = A$ or $H = T$ simply the error probability ε_j obtained from the base quality value q_j is used:

$$P(D_j|H = T) = \begin{cases} 1 - \varepsilon_j & \text{if } D_j = T \\ \varepsilon_j/3 & \text{else.} \end{cases} \tag{3.5}$$

The likelihood to observe a base under a haplotype $H = C$ or $H = G$, which is involved in bisulfite conversions on the top or bottom strand respectively, is dependent on the genomic methylation level and the bisulfite conversion rate of unmethylated Cs. Moreover, the Bis-SNP model assumes that also methylated Cs might convert into Ts. Let the following parameters be given:

β_j : Methylation level at the current position.

γ : Global bisulfite conversion rate for methylated Cs.

α : Global bisulfite conversion rate for unmethylated Cs.

For a better understanding, we will first have a look at how such base likelihoods can be modelled. A C mapping against the top strand under the haplotype C can be caused by two different scenarios:

1. The base is called correct $(1 - \varepsilon_j)$ and the genomic C is methylated (β_j) and not converted $(1 - \gamma)$.

2. The base is called correct and the genomic C is unmethylated $(1 - \beta_j)$ and not converted $(1 - \alpha)$.

Equivalently the likelihood for a T mapped against the top strand can be modelled. In general, the likelihood to observe a mapped base D_j under the haplotype $H = C$ can be computed in the following way:

$$P(D_j|H = C) = \begin{cases} (1 - \varepsilon_j)\,(\beta_j(1 - \gamma) + (1 - \beta_j)\alpha) & \text{if } D_j = C^+ \\ \varepsilon_j/3 + (1 - \varepsilon_j)(\beta_j\gamma + (1 - \beta_j)(1 - \alpha)) & \text{if } D_j = T^+ \\ 1 - \varepsilon_j & \text{if } D_j = C^- \\ \varepsilon_j/3 & \text{else,} \end{cases} \tag{3.6}$$

where C^+ denotes a C mapped against the top strand and C^- denotes a G mapped against the bottom strand. This distinction between bases mapped against the top strand (C^+, T^+) and bases against the bottom strand (C^-) is crucial to distinguish between C>T SNPs and bisulfite C>T conversions. The corresponding likelihoods under haplotype $H = G$ are modelled equivalently by using error probabilities for bases mapped against the top strand and additionally using methylation and bisulfite conversion rates for bases mapped against the bottom strand. Since methylated Cs usually do not get converted, the default setting in Bis-SNP for the bisulfite conversion rate of methylated C is 0. The position specific methylation level β_j, which is crucial for the base likelihood computation, is by default estimated by the percent of observed Cs regarding the total number of Cs and Ts. Optionally, the user can chose a context-specific estimation strategy, in which in a first run the context-specific methylation rates are obtained and then used in a second run for a more accurate likelihood computation. However, the authors showed that the default method achieves a higher accuracy.

Finally, the genotype with the highest posterior probability is chosen and a score is assigned, representing the odds ratio between the best genotype G_1 and the second best genotype G_2:

$$score = \log\left(\frac{P(G_1|D)}{P(G_2|D)}\right) \quad . \tag{3.7}$$

Since the Bis-SNP approach is proved to achieve accurate SNP calls and improved methylation level estimates compared to other methods [46], we build our method on this.

3.1.2 Extension for precise methylation level calling

The described method implemented in Bis-SNP uses the C-T ratio of the mapped reads at the current genomic position to estimate the methylation level. This in turn is used to compute the genotype posterior probabilities. However, this is a rather naive method, since this ratio might be already influenced by possible sequencing errors, (heterozygous)

SNPs or the bisulfite conversion rate. In the worst case, if the estimation is not precise, this can cause false genotype calls. Since we aim for a high accurate calling method for this bisulfite workflow, we extended this method to compute for each position the methylation level that explains the data best. More precisely, for each genotype containing a C or a G the methylation level is determined that maximizes the likelihood to observe the data under this genotype.

Before looking at the concrete methylation level computation, we describe the extension of the Bayesian model. We split up the individual base likelihood computation given in equation 3.6 to differentiate between methylated $H = C^M$ and unmethylated haplotypes $H = C^U$. Thus, the methylation level is indirectly given by either one or zero. For the sake of completeness, we extended the model further by the probability that a methylated or unmethylated C might be bisulfite converted and erroneously sequenced as a C^+:

$$P(D_j|H = C^M) = \begin{cases} (1-\varepsilon_j)(1-\gamma) + (\varepsilon_j/3)\gamma & \text{if } D_j = C^+ \\ \varepsilon_j/3 + (1-\varepsilon_j)\gamma & \text{if } D_j = T^+ \\ 1-\epsilon_j & \text{if } D_j = C^- \\ \varepsilon_j/3 & \text{else} \end{cases} \tag{3.8}$$

$$P(D_j|H = C^U) = \begin{cases} (1-\varepsilon_j)\alpha + (\varepsilon_j/3)(1-\alpha) & \text{if } D_j = C^+ \\ \varepsilon_j/3 + (1-\varepsilon_j)(1-\alpha) & \text{if } D_j = T^+ \\ 1-\varepsilon_j & \text{if } D_j = C^- \\ \varepsilon_j/3 & \text{else.} \end{cases} \tag{3.9}$$

The likelihoods $P(D_j|H = C^M)$ and $P(D_j|H = C^U)$ can then be used to calculate the likelihood to observe base D_j under the haplotype $H = C$ and a given methylation level β as follows:

$$P(D_j|H = C, \beta) = (1-\beta) \cdot P(D_j|C^U) + \beta \cdot P(D_j|C^M). \tag{3.10}$$

The methylation level β serves as a weight for the individual methylation state dependent likelihoods and can take any value between 0 and 1. The likelihood to observe the data D under the genotype $G = CT$ is now given by:

$$P(D|G = CT, \beta) = \prod_{j=1}^{n} \left(\frac{1}{2} \cdot P(D_j|H_1 = C, \beta) + \frac{1}{2} \cdot P(D_j|H_2 = T) \right). \tag{3.11}$$

Finally, each genotypes specific methylation level β maximizing $P(D|G, \beta)$ is determined. Given the posterior probabilities, the most likely genotype is called.

In the case of heterozygous SNPs the C-T ratio does not necessarily correspond to the methylation level, e.g. in the case of the CT genotype. This extended method comes with the advantage that for heterozygous Cs the methylation levels are automatically appropriately determined. Besides that, the sequencing error and bisulfite conversion probabilities are taken into account.

3.1.3 Extension for base dependent sequencing errors

So far, each possible sequencing error type has the same probability $\varepsilon_j/3$ assigned. We know already that during the sequencing process those errors do not occur with uniform frequencies. For example, C>T substitution errors are, according to different studies, relatively rare [16]. This information can be used in our method to improve the certainty of bisulfite C>T conversions.

We extended the computation of the individual likelihoods (see equations 3.8, 3.9) of our method to incorporate non-uniform error probabilities. This is similar to the method used in the mapping module (see Section 2.3.4). The likelihood $P(D_j|H = C^M)$ is now given by:

$$P(D_j|H = C^M) = \begin{cases} (1-\varepsilon_j)(1-\gamma) + (\varepsilon_j/3)\gamma & \text{if } D_j = C^+ \\ \varepsilon_j \cdot P_{seq}(C|T) + (1-\varepsilon_j)\gamma & \text{if } D_j = T^+ \\ 1-\varepsilon_j & \text{if } D_j = C^- \\ \varepsilon_j \cdot P_{seq}(C|x) & \text{if } D_j = x, \end{cases} \tag{3.12}$$

where $P_{seq}(C|D_j)$ represents the probability that the true base is a C under the assumption that the observed base D_j is caused by a substitution error. The likelihoods for the other haplotypes are extended equivalently.

3.1.4 Incorporating mapping qualities

Base quality values serve as a good indicator for sequencing errors. Nevertheless, they are useless if the whole read is mapped at the wrong genomic location. We additionally make use of mapping qualities to give bases from high scoring and very unique mappings a higher weight than bases from low scoring or less unique mappings. Equivalently to base qualities, the phred-scaled mapping quality Q_M is converted to the probability that the mapping is wrong:

$$\varepsilon_M = 10^{-\frac{Q_M}{10}}. \tag{3.13}$$

The probability that the mapping is right is then used for each observed read base D_j to weigh its likelihood:

$$P(D|G = CT, \beta) = \prod_{j=1}^{n} (1-\varepsilon_{M_j}) \cdot \left(\frac{1}{2} P(D_j|H_1 = C, \beta) + \frac{1}{2} P(D_j|H_2 = T) \right). \tag{3.14}$$

3.1.5 Non-uniform SNP probabilities

We know already that genomic substitutions do not occur with uniform distributions (see Section 1.2.5). Especially when working with bisulfite read sequences, it is crucial to distinguish between C>T conversions and C>T SNPs, which are the most frequent SNPs in mammalian genomes. In Bis-SNP this problem is addressed by taking the prior genotype probability $\pi(G)$ for the posterior probability computation (see equation 3.1) into account. The prior probability computation in done based on the method implemented in SOAPsnp [42], a SNP calling tool for non-bisulfite data. It takes the probability of transitions and transversions as well as the probability of haploid and diploid genotypes into account, given the reference allele. Since this is a well-grounded strategy, we incorporated it into our method. For each possible reference and candidate genotype combination its prior probability is precomputed.

3.2 Methylation level estimation

In the previous section we described the underlying Bayesian model to compute the posterior probabilities of all possible diploid genotypes, wherever required, given the respective methylation level. Finding a precise estimation for the methylation level associated with a genotype can now be formulated as an optimization problem, namely finding the β maximizing the likelihood $P(D|G,\beta)$ (see equation 3.14).

Due to the nature of β, only the maximum in the interval $[0,1]$ is of interest. The naive way for solving this problem is to perform a systematic sampling. Therefore the likelihood $P(D|G,\beta)$ needs to be evaluated for each investigated β, whereby each evaluation requires n multiplications of the β dependent sum in equation 3.10. This can be quite expensive for large n and small sampling intervals. Since we aim for accurate methylation levels, we require small sampling intervals. Especially in the case of RRBS (see Section 1.2.4) with highly covered regions, this has a great impact on the runtime. We examined a range of different strategies to reduce the number of β dependent evaluations and the necessary multiplications for one evaluation.

3.2.1 Gradient based numerical optimization

The maxima of an objective function are represented as roots in its first derivative. Finding these roots thus can be used for optimization. However, due to the form of equation 3.10, the theoretically possible number of roots for $\beta \in \mathbb{R}$ is $n-1$, so that we need a numerical approximation in order to find the roots in the interval $[0,1]$. An efficient strategy able to reduce the number of required evaluations are gradient based numerical optimization methods, such as the Newton's iterative method. In this scenario the method makes use of the first derivative of our objective function, representing the roots, and the second derivative, representing the gradient of the first derivative.

Nevertheless, the derivatives of our objective function need to be evaluated in each step of the Newton iteration. Regarding the original form they are rather complex and require in case of the first derivative already $(n-1)^2$ multiplications of the β dependent

sum. The second derivative is even more complex. In order to address this problem, an alternative representation is needed.

Polynomial evaluation Given equation 3.14 and 3.10, the problem can be expressed as a polynomial function of the variable β. The degree of this function is dependent on the number of mapped bases n. This polynomial function can then be converted to the Horner form, reducing the number of required multiplications for each evaluation. An additional advantage is that the first and second derivatives can be computed on the fly while computing the Horner scheme.

However, it turned out that this method is not (well) qualified for evaluating our objective function due to arising rounding errors. The reason lies in the polynomial coefficients that have to be computed for the Horner form. In case of a large polynomial degree the individual base likelihoods get multiplied in such a way that fairly high and low coefficients are resulting at the same time. This becomes a problem during the evaluation of the Horner scheme, because then very small and large numbers are added and multiplied, causing rounding errors. Especially for the evaluation with non-optimal β values the individual likelihoods can become very small. This leads to enormous deviations from the true β dependent likelihood and thus to wrong genotype and methylation level callings. So even though the Horner algorithm is in general quite stable for $|\beta| < 1$ [5], for our polynomial function the rounding errors accumulate causing impractical results.

Since this case almost only arises in the evaluation of non-optimal β values with very low likelihoods, the main problem is less the existence of such errors, but rather becoming aware of these. Thus one possible solution is to compute simultaneously to the Horner scheme evaluation an error bound [26], as shown in Algorithm 1 for the original polynomial function:

Algorithm 1 Evaluation of Horner scheme with running error bound

Input: Horner coefficients a_i with $i \in 1, ..., n$; methylation level β; given precision p
Output: Likelihood y; error bound eB

```
1: procedure EVALUATE(a, β, p)
2:     y = a_n                                   ▷ Initialization
3:     eB = 1/2 · |y|
4:     for i = n − 1 : −1 : 0 do
5:         y = a_i + β · y          ▷ Recursive evaluation of the polynomial function
6:         eB = |y| + β · eB              ▷ Recursive calculation of the error bound
7:     end for
8:     eB = p · (2eB − |y|)                      ▷ p: Given precision
9: end procedure
```

It could be shown that the most wrong optimization results were caused in the context of erroneous likelihood results being smaller than their computed error bound. The error bound is inversely proportional to the true likelihood and rather low for likelihood

results in the direct neighborhood of the true maximizing β. Thus, in this method, likelihood results smaller than their computed error bound are discarded. Subsequently, a bisection method is used to identify a new valid region and the Newton iteration is performed again. In the majority of cases with a limited number of mapped reads it seems sufficient to only compute the error bound of the original polynomial function. Nevertheless, the Newton method uses the first and second derivatives with deviating error bounds, which need to be checked as well in order to obtain accurate and reliable results. The derivatives display validated error bounds in different regions compared to the original polynomial function, i.e. in the near of its roots, which complicates the subsequent processing. A branch and bound strategy is applied to identify valid regions for the Newton iterative method or, in case of a validated error bound close to a potential optimal β, a non-gradient based bisection method is used to find the maxima in the original polynomial function.

Although the number of non-optimal beta values caused by rounding errors can be reduced significantly with the help of the described error bound, the naive optimization method still leads to more accurate results.

Polynomial evaluation in log space We simultaneously implemented the polynomial method in logarithmic space using the SeqAn class LogProb. This stores logarithmic values internally and multiplications are substituted by additions in order to achieve a higher numerical stability. This method provides results with a comparable accuracy as the naive evaluation method. However, since logarithmic values can only represent positive values while the polynomial coefficients can be negative, the intermediate results need additionally to be checked for zero and negative cases. As a consequence, the runtime increases and exceeds the runtime of the naive method.

Log-likelihood function Due to the described problems, none of the previous mentioned strategies performs satisfactorily. Thus we finally decided to apply the maximum likelihood estimation (MLE) method using the log-likelihood function, which is widely used for estimating parameters of statistical models. The monotonic property of the logarithmic function implies that the logarithm of a likelihood function has its maxima at the exact same locations as the original likelihood function and consequently can be used as the new objective for optimization. Let $\mathcal{L}(G, \beta|D)$ be the original likelihood function given indirectly with equation 3.14, then the log-likelihood function is as follows:

$$\begin{aligned}\log \mathcal{L}(G,\beta|D) &= \log_{10}\left[\prod_{j=1}^{n}(1-\varepsilon_{M_j})\left(\frac{1}{2}\left((1-\beta)P(D_j|H_1=C^U)+\beta P(D_j|H_1=C^M)\right)\right.\right.\\ &\qquad\left.\left.+\ \frac{1}{2}P(D_j|H_2=T)\right)\right]\\ &= \sum_{j=1}^{n}\log_{10}\left[(1-\varepsilon_{M_j})\left(\frac{1}{2}((1-\beta)P(D_j|H_1=C^U)+\beta P(D_j|H_1=C^M))\right.\right.\\ &\qquad\left.\left.+\ \frac{1}{2}P(D_j|H_2=T)\right)\right]\\ &= \sum_{j=1}^{n}\log_{10}(a_j+b_j\beta),\end{aligned} \tag{3.15}$$

with

$$a_j = \frac{1}{2}\cdot(1-\varepsilon_{M_j})\left(P(D_j|H_1=C^U)+P(D_j|H_2=T)\right) \tag{3.16}$$

$$b_j = \frac{1}{2}\cdot(1-\varepsilon_{M_j})\left(-P(D_j|H_1=C^U)+P(D_j|C^M)\right). \tag{3.17}$$

Given the coefficients a_j, b_j for $j \in \{1,...,n\}$ the log-likelihood function $\log \mathcal{L}(G,\beta|D)$ can be evaluated. The coefficients are β independent and can be precomputed. In order to use a gradient based optimization method, the first and second derivatives need to be determined, given by:

$$\log \mathcal{L}(G,\beta|D)' = \frac{1}{\ln 10}\cdot\sum_{j=1}^{n}\frac{b_j}{a_j+b_j\beta} \tag{3.18}$$

$$\log \mathcal{L}(G,\beta|D)'' = \frac{1}{\ln 10}\cdot\sum_{j=1}^{n}\frac{-b_j^2}{a_j^2+2a_jb_j\beta+b_j^2\beta}. \tag{3.19}$$

Finally, Newton's iterative method is used to find the maximizing β. In this approach there is no need to check for zero or negative values, since the original likelihood function takes on only positive values. Moreover, the coefficients computed for this method do not take on values as extremely low and high as in the polynomial case, so that rounding errors are prevented. Compared to the naive method, the evaluation of the log-likelihood function itself is less expensive and more stable due to the additions replacing multiplications. Additionally, derivatives can easily be constructed, allowing for an efficient optimization by the gradient based Newton method [1]. In conclusion, we set this log-likelihood method as the default behavior to estimate the methylation level for each candidate genotype.

[1] The evaluations of $\log \mathcal{L}(G,\beta|D)'$ and $\log \mathcal{L}(G,\beta|D)''$ are more expensive than of $\log \mathcal{L}(G,\beta|D)$, but still practical compared to the alternatives.

3.3 Output of called SNPs and methylation levels

Similarly to the Bis-SNP method (see 3.7), each most likely genotype gets a score assigned, given by the log odds ratio between the posterior probability of the best and the second best genotype. For each genomic position the genotypes determined by our method are verified given user defined threshold regarding their coverage, score and likelihood. Deviations from the reference allele are emitted as SNPs. In case of called genotypes containing a C or G, the corresponding estimated methylation levels are emitted. The output is written into a VCF like format, containing the basic information about the reference name, the genomic position, the reference base and the observed coverage. If required, i.e. in the case of the genotype CG, the entry for one genomic position contains the information about the called SNP, the top strand methylation level and the bottom strand methylation level. In this way redundant outputs are prevented.

3.4 Implementation

In order to avoid loading all given alignments into memory at the same time, the SAM input file must be sorted by genomic position. In this way, a window-wise parsing and analysis of the alignments is possible. This strategy is based on the existing SeqAn tool SnpStore [20], which is designed for conventional SNP calling. The window size can be specified by the user and thus fitted to the coverage and available memory size. Additionally, if only the analysis of certain genomic intervals is required, as for example in the cases of RRBS (see Section 1.2.4), only user defined intervals are processed. For each window or interval, all given pairwise alignments are converted into a multiple alignment. This is done using the existing functionality of the SeqAn *FragmentStore*, an efficient data structure to store read mapping information. In order to allow for a strand specific methylation analysis, all reads originating from the original top strand are internally stored as forward mappings, while reads originating from the bottom strand are stored as reverse mappings. Moreover, the paired-end information is stored in order to maintain the information about the origins, which is necessary in retrieving the appropriate sequencing error probabilities.

For the likelihood optimization step we make use of the Newton-Raphson method already implemented in the Boost C++ Libraries [14].

Similar to the mapping module of this bisulfite workflow, the method described in this chapter is able to incorporate non-uniform SNP and sequencing error probabilities. If the data requires, a uniform model can be selected.

4 Local Multiple Sequence Realignment

In the previous chapters we presented an advanced statistical method for computing column-wise accurate pairwise alignments, which subsequently can be used for position-wise SNP and methylation level callings. However, genomic indels have an important impact on the accuracy of such alignments and thus influence the calling results significantly (see Section 1.2.5).

In the following we present a method for the computation of accurate bisulfite MSAs by taking all reads into account in order to compute a consistent multiple sequence alignment (MSA), see Figure 4.0.1.

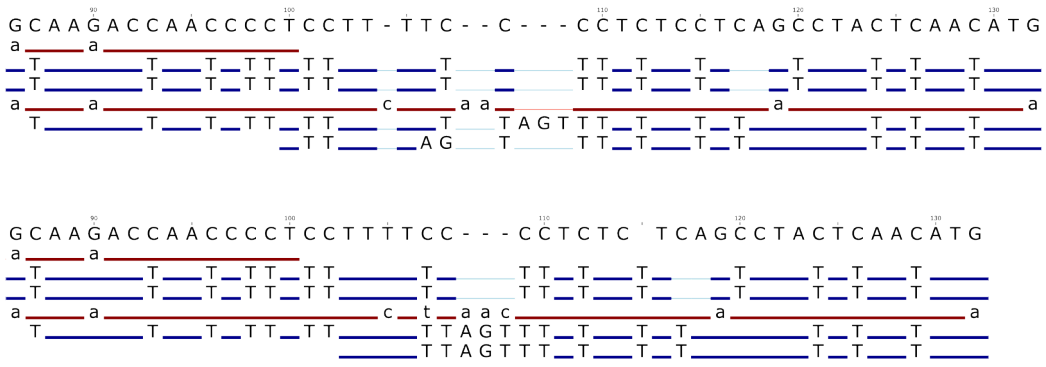


Figure 4.0.1: Multiple sequence alignment (MSA) for bisulfite reads before and after multiple sequence realignment. The blue and red lines represent reads mapped against the top and the bottom strand respectively. Bases, deviating from the reference base are separately displayed (including bisulfite conversions). Gaps in the alignment are denoted with transparent lines. The example illustrates, how gaps can be rearranged to achieve a more consistent layout. For top mapped reads C>T conversions and for bottom mapped reads G>A conversions are allowed.

We use the Anson-Meyers' heuristic (see Section 1.3), in which reads are iteratively realigned against the MSA, until no further improvement is visible.

Our method uses the full four-letter space and distinguishes between reads containing the information about top and bottom strand DNA methylations. Specific scoring functions allow for bisulfite conversions, take C-T ratios into account and incorporate base qualities in a related way as presented in Chapter 2. In this way we avoid losing the accuracy obtained in the upstream alignment computation.

4.1 Scoring scheme

In order to map one read against a given MSA or more precisely against its profile, a scoring function is required. For bisulfite reads, the original Anson-Meyers' scoring function is rather impractical, because in the case of bisulfite C>T conversions the concept of consensus bases would be misleading.

We use a score designed in a similar way as the scores in Section 2.3.4 for the pairwise read-to-reference alignment, taking the base dependent sequencing error probabilities into account. Given a base a and the column χ of the MSA at a specific position, we define the score as:

$$S_{\chi a} = T \cdot logR_{\chi a}, \tag{4.1}$$

where T is again an arbitrary scaling factor. Let $P(\chi a|\mathcal{A}_{\mathcal{BS}})$ denote the target frequency of observing base a aligned against column χ in our bisulfite model. Further, let $P(\chi)$ and $P(a)$ be the background frequencies of the column χ and base a respectively. Then the generalized likelihood ratio $R_{\chi a}$ of column χ and base a given its error probability ε can be computed as follows:

$$R_{\chi a} = \frac{P(\chi a|\mathcal{A}_{\mathcal{BS}})}{P(\chi)P(a)} \cdot (1-\varepsilon) + \sum_{b \in \{A,C,G,T\} \setminus \{a\}} \frac{P(\chi b|\mathcal{A}_{\mathcal{BS}})}{P(\chi)P(b)} \cdot \varepsilon \cdot P(b|a), \tag{4.2}$$

with $P(b|a)$ being the probability that the true base is b under the assumption that the observed base a is caused by a substitution error. However, in this case we do not know the target frequency $P(\chi a|\mathcal{A}_{\mathcal{BS}})$ and the background frequency $P(\chi)$. Applying the axiom of conditional probability, we obtain:

$$R_{\chi a} = \frac{P(a|\chi)}{P(a)} \cdot (1-\varepsilon) + \sum_{b \in \{A,C,G,T\} \setminus \{a\}} \frac{P(b|\chi)}{P(b)} \cdot \varepsilon \cdot P(b|a). \tag{4.3}$$

This formula comes with the great advantage of being independent of the probability to observe χ.

4.2 MSA profile

The remaining challenge now is to determine the posterior probability $P(a|\chi)$ of aligning the base a against the given column χ. Here, the observed base frequencies are brought into the equation. We assume that the observed frequency of base a given n reads somehow reflects the probability of the $(n+1)$th base in this column being an a.

Commonly, sequence profiles are used to store for each position in an MSA the frequency of each possible character, e.g. A, C, G, T. For non-bisulfite read alignments, the posterior probability $P(a|\chi)$ could be estimated by simply using the base frequency of a given by the profile. However, for bisulfite data this is not a trivial task. Since it is

not possible to know at this point whether a read base T originates from a genomic T or a genomic C, it is not possible to make use of the overall base frequencies.[1] A concrete example is displayed in Figure 4.2.1.

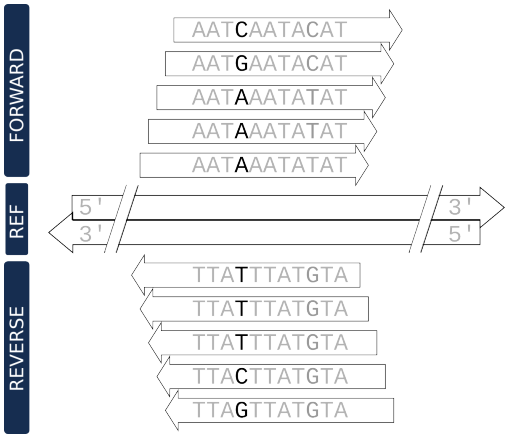


Figure 4.2.1: Difficulties for frequency determination: For bisulfite reads the overall base frequencies are not unambiguous due to possible bisulfite conversions. Assume one wants to determine the frequency of base A at the highlighted position, this would be straightforward for taking into account only reads mapped against the forward strand, whereas the reads mapped against the reverse strand do not contain clear information, since it is not possible to know whether the Ts are originating from a genomic C or T.

We decided to treat reads mapped against the top and the bottom strand separately. Therefore we use an extended profile distinguishing between top and bottom strand base frequencies. Additionally the information about the gap frequency is included. This gives us a profile over the alphabet $\Sigma_{BS} = \{A_{top}, C_{top}, G_{top}, T_{top}, A_{bot}, C_{bot}, G_{bot}, T_{bot}, -\}$, with *top* denoting top bases and *bot* bottom bases respectively. Note that the gap frequency is defined regarding to both strands.

We denote the column of such a profile, representing the different character frequencies of a certain MSA position, with ρ. Given a base a corresponding to a read mapped against the top strand, the posterior probability of mapping it against an alignment column χ is estimated with:

$$P(a|\chi) = \rho(a) \quad \forall a \in \{A_{top}, C_{top}, G_{top}, T_{top}\} \tag{4.4}$$

$$\tag{4.5}$$

[1] Addressing this problem by taking global methylation and bisulfite conversion rates into account would be too inaccurate for a specific position. Another approach could be the estimation of the position specific bisulfite conversion level and of the underlying genotype. But this is not the goal of this module and wrong initial estimations would cause errors.

For aligning gaps, the overall gap frequency is used:

$$P(-|\chi) = \rho(-). \tag{4.6}$$

This estimation is also valid for bases involved in bisulfite conversions, as it can be assumed that the observed C-T ratio regarding one strand (independently of its causes) represents the ratio of the probability to align a C or a T. This method does not guarantee to find the optimal MSA, since not all information is used. It may lead to cases where the top and bottom mapped bases are not optimally aligned respectively to each other. However, for the posterior probability computation regarding gaps, all reads are taken into account in order to ensure a consistent layout between the reads mapped to the different strands. Since multiple sequence realignment reorders gaps within the sequences, while the bases between those gaps stay the same, such inaccuracies are expected to be minor if a minimum coverage is given for both strands.

Additionally, in order to improve the consistency, the reference sequence is realigned as well. This requires a different posterior probability computation, on the one hand because the reference is not strand specific and on the other hand because it does not contain any bisulfite conversions. We use an approximation, by taking for each base to be aligned only those frequencies into account that are not biased through possible bisulfite conversions.

4.3 Gaps

Previously we presented our method for computing the score for mapping one base against an MSA column. Since the main goal of the realignment is to compute a consistent layout regarding genomic indels, accurate modelled gap scores are at least just as crucial.

The score function for introducing a gap into the current read is again correspondingly designed to the pairwise read-to-reference alignment score in Section 2.3.4, taking base dependent sequencing deletion error probabilities into account. But instead of using a simple predefined gap score, the position specific gap posterior probability can be used now. Let the parameter ϕ be again the global deletion error rate, then the score can be computed as follows:

$$R_{\chi-} = \frac{P(-|\chi)}{P(-)}(1-\phi) + \sum_{b\in\{A,C,G,T\}} \frac{P(b|\chi)}{P(b)}\phi \cdot P(b|-), \tag{4.7}$$

with $P(b|-)$ denoting the base dependent deletion error probability. For $P(-)$ a global genomic deletion rate can be used.

Introducing a gap into the profile needs to be handled a bit differently, since there is no MSA column given for this position. Therefore the posterior probability is modelled by using a pseudo-column containing minimal pseudo-frequencies for each base to avoid probabilities of zero. In other words, we model an MSA column consisting mainly of

gaps. For the gap frequency, the number of mapped reads at the previous position is used [2]. Given the posterior probabilities, the likelihood ratio R_{-a} for mapping a base a against a gap in the MSA can now be calculated with:

$$R_{-a} = \frac{P(a|\chi_{pseudo})}{P(a)}(1-\varepsilon) + \varepsilon \cdot P(-|a) + \sum_{b \in \{A,C,G,T\} \setminus a} \frac{P(b|\chi_{pseudo})}{P(b)} \cdot \varepsilon \cdot P(b|a), \quad (4.8)$$

with $P(-|a)$ being the base dependent insertion error probability and ε being the error probability of base a. Since the probability of introducing a gap into the MSA in combination with a sequencing substitution error is extremely small, we omit these summands for the purpose of simplification.

4.4 MSA profile taking base qualities into account

In the previous sections we presented a scoring method for bisulfite read-to-MSA alignments using strand aware profiles. These scores take the sequencing error probabilities of the current read into account. However, so far the error probabilities of the bases contained in the MSA do not have any influence, since only base frequencies are used.

Thus we decided to incorporate the base quality values into the profile. In the following we describe our method for estimating the true base and gap frequencies regarding the underlying bisulfite treated sequences by taking possible sequencing errors into account. Therefore we define an alphabet $\Sigma_{top} = \{A_{top}, C_{top}, G_{top}, T_{top}, -_{top}\}$ for the top strand and use characters of this alphabet instead of bases in order to include gaps. An equivalent alphabet is defined for the bottom strand.

Individual posterior probabilities Prior to the actual frequency estimation, we first need to compute for each sequence position in the MSA the probability that the true underlying character is c under the observed alignment character. In the case that a read base a with a given quality value q is observed, the probability can be computed as follows using the indirectly given error probability ε:

$$P(c|q,a) = \begin{cases} 1-\varepsilon & \text{if } c = a \\ \varepsilon/3 & \text{else if } c \neq -_{top} \\ 0 & \text{else} \end{cases} \qquad \forall a \in \Sigma_{top} \setminus \{-_{top}\}, c \in \Sigma_{top} \quad (4.9)$$

Note that $\sum_c P(c|q,a) = 1$ holds for all possible a. If the observed character is a gap, no base quality is given and simply an error probability of zero is used. For consistency reasons, an infinite pseudo quality value is assigned, so that the following holds:

[2] The more reads are mapped at this position, the lower the probability to insert a gap into the profile.

$$P(c|q,-) = \begin{cases} 1 & \text{if } c = -_{top} \\ 0 & \text{else} \end{cases} \qquad \forall c \in \Sigma_{top} \tag{4.10}$$

Additionally, we make use of the information given by the reference sequence included in the MSA. Since the reference sequence differs from the possible bisulfite treated sequences, we model the pseudo-probability that given a reference base r a bisulfite treated sequence contains the alignment character c. Therefore one sequence is modelled mapping against the top and one mapping against the bottom strand, in order to use this information for both strands. Due to highly confident reference bases without quality values, the assumed error probability is again zero. However, possible bisulfite conversions need to be modelled. For the top strand, this is implemented as follows:

$$P(c|r) = \begin{cases} 1 & \text{if } r \neq C_{top} \;\&\; c = r \\ \gamma & \text{if } r = C_{top} \;\&\; c = C_{top} \\ 1-\gamma & \text{if } r = C_{top} \;\&\; c = T_{top} \\ 0 & \text{else} \end{cases} \qquad \forall c \in \Sigma_{top}, \tag{4.11}$$

where the parameter γ is an estimate for the C>T conversion rate based on the position specific C-T ratio observed in the top strand mapped reads. Equivalently, this is done for the bottom strand.

Frequency estimation Previously we have shown how to compute the posterior probabilities for each possible underlying alignment character. These probabilities are now used to estimate the underlying frequencies.

Assume we observed a base G mapped against the top strand with a given quality value. Instead of only increasing the frequency of G_{top} in the profile, we calculate this for the frequencies of all top bases using their probabilities. Let ρ denote the column of the profile at a certain alignment position and n_{top} the number of reads mapped against the top strand. Given the observed bases x_j with $j \in \{1, ..., n_{top}\}$ and their assigned quality values q_j, the frequency $\rho(a)$ for a top base is computed as follows:

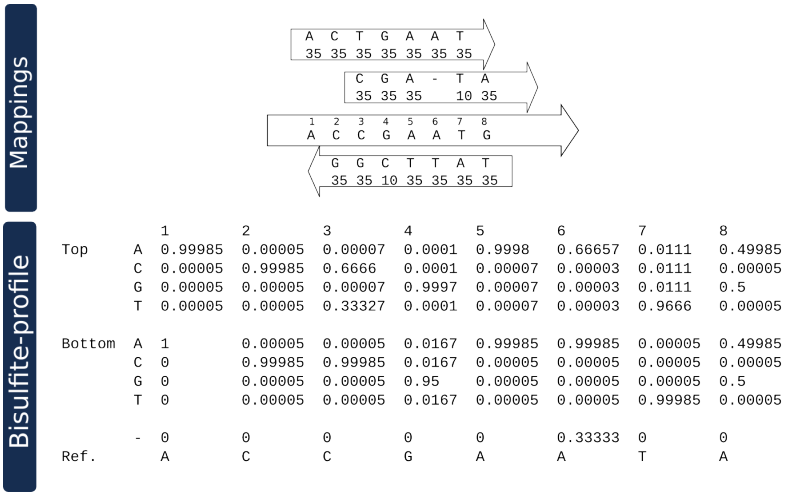
$$\rho(a) = \frac{P(a|r) + \sum_{j=1}^{n_{top}} P(a|q_j, x_j)}{\sum_{c \in \Sigma_{top}} \left(P(c|r) + \sum_{j=1}^{n_{top}} P(c|q_j, x_j) \right)} \quad \forall a \in \Sigma_{top} \backslash \{-_{top}\}. \tag{4.12}$$

Note that for the computation of top base frequencies only alignment characters from top mapped sequences and the reference base are considered. In contrast, for the gap frequency, all reads are taken into account, with n being the total number of mapped

reads:

$$\rho(-) = \frac{P(-|r) + \sum_{j=1}^{n} P(-|q_j, x_j)}{\sum_{c \in \Sigma_{top} \cup \Sigma_{bottom}} \left(P(c|r) + \sum_{j=1}^{n} P(c|q_j, x_j) \right)}. \tag{4.13}$$

Equivalently the bottom character frequencies can be determined.



		1	2	3	4	5	6	7	8
Top	A	0.99985	0.00005	0.00007	0.0001	0.9998	0.66657	0.0111	0.49985
	C	0.00005	0.99985	0.6666	0.0001	0.00007	0.00003	0.0111	0.00005
	G	0.00005	0.00005	0.00007	0.9997	0.00007	0.00003	0.0111	0.5
	T	0.00005	0.00005	0.33327	0.0001	0.00007	0.00003	0.9666	0.00005
Bottom	A	1	0.00005	0.00005	0.0167	0.99985	0.99985	0.00005	0.49985
	C	0	0.99985	0.99985	0.0167	0.00005	0.00005	0.00005	0.00005
	G	0	0.00005	0.00005	0.95	0.00005	0.00005	0.00005	0.5
	T	0	0.00005	0.00005	0.0167	0.00005	0.00005	0.99985	0.00005
	-	0	0	0	0	0	0.33333	0	0
Ref.		A	C	C	G	A	A	T	A

Figure 4.4.1: Bisulfite profile construction (whereby mapping qualities are not taken into): The reference bases are displayed with their position, the read bases are given with their quality values. For the profile construction the mapped bases are taken into account corresponding to their probabilities, separately for top and bottom bases.

Mapping qualities Correspondingly to the calling module of this BS-Seq analysis workflow (see Section 3.1.4), optionally the mapping quality is taken into account. Therefore the mapping quality is converted into a mapping error probability ε_M, which then serves as a weight for the frequency estimation. For top base frequencies, the following formula is used:

$$\rho(a) = \frac{P(a|r) + \sum_{j=1}^{n_{top}} (1 - \varepsilon_M) \cdot P(a|q_j, x_j)}{\sum_{c \in \Sigma_{top}} \left(P(c|r) + \sum_{j=1}^{n_{top}} (1 - \varepsilon_M) \cdot P(c|q_j, x_j) \right)} \quad \forall a \in \Sigma_{top} \backslash \{-_{top}\}. \tag{4.14}$$

4.5 Implementation

The described realignment method is integrated into the SNP and methylation calling module, which was presented in Chapter 3. Respectively, the same window-wise processing is performed (see Section 3.4) to reduce the memory consumption. For the Anson-Meyers' round-robin realignment strategy, we used the method already in SeqAn implemented [55] and adopted it to the specific requirements of our method. However, this method represents the runtime bottleneck of the whole BS-Seq analysis workflow, due to the frequently performed read-to-MSA alignment computations. In order to reduce this limitation, we implemented it using openMP [13], an API that supports shared memory multiprocessing programming. This enables a contig-wise parallel processing, where each thread parses a contiguous part of the input SAM file, corresponding to its assigned contig, and analyses the data window-wise.

A large fraction of the given read mappings is probably already quite accurate. For this reason, only genomic regions with a certain number of indel-containing mappings are selected for the realignment. In order to avoid edge-effects, the connected subset of reads at these regions is taken into account. For the actual alignment computation, correspondingly to the Anson and Meyers approach, a banded alignment function is used, aiming to keep the complexity as low as possible.

A characteristic of the scoring scheme is that the score for introducing a new gap into a read at a position with a high gap frequency is greater than zero. In a free-end gap alignment model, additional gaps would be introduced through all reads in the MSA due to the alignment score maximization. In order to address this problem, we assign a positive score to end-gaps, proportionally to internal gap scores at high gap frequency positions.

5 Annotation Mapping

In the previous chapters we described our method of computing a precise estimation of methylation levels at single-nucleotide resolution. From a users' perspective, beside the position-wise methylation levels also overall methylation rates for specific genomic regions are worth analysing. For this reason, we provide a convenience module to efficiently compute some basic statistics for annotated genomic regions. The previous position-wise called methylation levels are mapped to genomic annotations given in GFF/GTF format. For this purpose we project the methylation levels to methylation states, i.e. methylated, fuzzy methylated and unmethylated. By default Cs with a methylation level lower than 0.25 are defined as 'unmethylated', while Cs with a methylation level higher than 0.75 are defined as 'methylated'. Everything between those values is called 'fuzzy' methylated. For each annotation the rate of each methylation state is determined. Optionally, if a reference sequence is given, this is done for each context (CG, CHG, CHH) separately.

If specified, the same is done for each given annotation type (e.g. exon, intron, UTRs, etc.) to get an overview of the overall methylation rates in this data. The output is in GTF format, containing the methylation rates as additional key-value pairs in the 9th GTF field (see Figure 5.0.1).

```
21      .       exon    9893990 9894300 .       +       .   \
  gene_name "AF254982.2"; transcript_name  "AF254982.2-001"; exon_number "1"; \
  meth_rate_cg "-1"; fuzzy_rate_cg "-1"; \
  meth_rate_chg "0.0909091"; fuzzy_rate_chg "0.181818"; \
  meth_rate_chh "0.016129"; fuzzy_rate_chh "0.0645161"; \
  gene_id "ENSG00000220964"; transcript_id "ENST00000402565";
```

Figure 5.0.1: Example output line of a GTF file containing for each annotation additionally the information about the context dependent methylation rates.

5.1 Implementation

In order to map the position specific methylation levels to genomic annotations, existing data structures of SeqAn are used. The annotations given in GFF/GTF format are loaded into the SeqAn *AnnotationStore*, an efficient data structure designed to store information about annotations with a mapping to other related data structures, e.g. storing the contig sequences. For the actual mapping we make use of interval trees, a commonly used data structure to represent intervals in an hierarchical way. Such interval

trees are already provided by SeqAn. For each reference contig and strand, one interval tree is constructed containing the corresponding annotation intervals. Finally, these interval trees are used to retrieve efficiently all annotations overlapping the methylation level positions.

6 Bisulfite Read Simulation

This chapter introduces the bisulfite read simulation method used to provide read sequences with realistic haplotype variations, methylation patterns, bisulfite conversions and error distributions. The simulated haplotype variations and methylation levels can be written out. In this way, we are able to perform a precise benchmarking of the developed mapping and analysis modules.

6.1 Mason by Holtgrewe et al.

We make use of the existing SeqAn tool Mason [30], an advanced non-bisulfite read simulator. In the following we present only the main parts being important for the subsequent bisulfite extension, assuming the directional protocol. Mason first simulates haplotypes based on a given reference sequence, containing SNPs and indels corresponding to user defined probabilities. Next, genomic fragments of a certain length are randomly simulated over the given sequence. For each fragment one read is simulated starting at one of the ends. The orientation is picked randomly, the read length can be specified by the user. In case of reverse orientation, the complementary sequence is assigned. For paired-end reads it is ensured, that both reads are simulated from opposite strands.

Matches, mismatches, deletions and insertions are assigned to each read position corresponding to given probabilities. Mismatches follow additionally a sequencing technology dependent empirical error distribution, for Illumina reads with an increasing error probability towards the end. Based on this error distribution and on the current error state, additionally quality values are assigned to each read position.

Both genomic variations and sequencing errors are applied independently of the bases. The substituting and inserted bases are randomly chosen from an uniform base distribution.

6.2 Extension for bisulfite read simulation

For the bisulfite extension Mason represents already a good starting point, since its individual simulation steps somehow reflect the order of the real biological processes followed by the significant steps performed during the actual sequencing protocol. This allows us to easily extend the model correspondingly to the directional BS-Seq protocol.

First, given a reference sequence, the methylation levels for each genomic C position are simulated dependent on the genomic context. This is done for each C position in the top and in the bottom strand (seen as a G in the reference) independently. During the haplotype construction, these methylation levels are adjusted to the new positions,

removed in case of deletions and de-novo assigned in case of inserted Cs. Next, for each fragment the bisulfite treatment is simulated. Therefore first the methylation states are randomly picked corresponding to the position specific methylation levels. For unmethylated Cs it is then randomly chosen given a user defined rate, whether the bisulfite C>T conversion is applied or not. The sequencing errors are applied subsequently. This corresponds to the order of the biological process and prevents conflicts between errors and bisulfite conversions.

As the default setting, reads corresponding to the directional protocol are simulated (see Section 1.2.3). Thus in the case of single-end reads a read either comes from the original top strand or bottom strand containing C>T conversions. For paired-end reads the right reads originate from the reverse complement strands and thus contain G>A instead of C>T conversions.

6.3 Methylation level distribution

Each genomic C is assigned a context dependent methylation level. Since these methylation levels are naturally beta distributed (see Section 1.2), we use a corresponding model in our method. More precisely, each context is assigned its own beta distribution, which can be specified by the user. For each genomic C a methylation level is then randomly chosen following the respective distribution (see Figure 6.3.1).

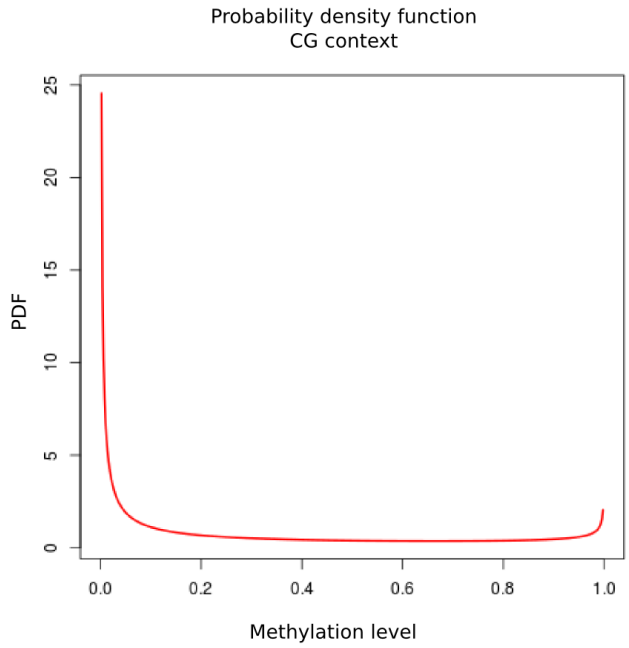


Figure 6.3.1: Methylation level simulation: For genomic Cs in CG context the methylation levels are beta distributed, with $Beta(0.22, 0.715)$ as the default setting.

6.4 Base dependent sequencing error probabilities

Sequencing errors have a large impact on the accuracy of NGS analysis tools. As described in Section 1.2.6, sequencing error frequencies are not distributed uniformly over all possible types. For the developing and benchmarking process of our method it is crucial to simulate such errors as realistic as possible. This is especially important, since one goal of our method is to reduce the bias caused by such errors.

We extended Mason to optionally use base dependent error probabilities for substitution as well as indel errors.

7 Results

In this chapter we describe the results we obtained with our method, using different types of data and different settings of the individual workflow modules. Additionally we compared it to the standard tools for BS-Seq data analysis. In order to enable an accurate benchmarking, we used simulated reads providing a gold-standard. Therefore, we know the true genomic origins of our reads, the simulated SNPs and the simulated methylation levels for each genomic C position.

7.1 Simulated data

We used the method described in Chapter 6 in order to simulate realistic Illumina bisulfite reads of length 100 bp. For all experiments, the human chromosome 21 reference sequence was used to simulate two different haplotypes, over which the reads were randomly distributed. If not specified further, the settings described in the following were applied.

The default SNP rate was set to 0.005 and the indel rate to 0.001. 7.9 million single-end reads were simulated in order to achieve an average genomic coverage of approximately 16, i.e. a coverage of 8 per strand. We assigned quality values of 32 and 11 for correct and erroneous bases respectively, which are realistic values for real data [50]. Note, that no position dependent error curves were applied. The substitution error rate was set to 0.005 and the insertion and deletion error rates to 0.001. However, we simulated reads with base dependent sequencing error probabilities, using the values obtained by previous studies that analysed error distributions [16][50]. Table 7.1 and 7.2 display the sequencing substitution and indel error probabilities, which were used for the simulation and furthermore in the subsequent modules, if required.

From \ To	A	C	G	T	$\sum$
A	-	0.14	0.05	0.05	0.25
C	0.13	-	0.02	0.04	0.19
G	0.04	0.08	-	0.12	0.24
T	0.08	0.15	0.09	-	0.33
$\sum$	0.25	0.38	0.16	0.21	-

Table 7.1: Sequencing substitution error probabilities dependent on original and erroneous base type.

	A	C	G	T
Ins	0.43	0.065	0.065	0.43
Del	0.42	0.075	0.075	0.42

Table 7.2: Sequencing indel error probabilities for different bases.

7.2 Benchmarking method

Read mapping

In order to evaluate the read mapping results of our workflow, we made use of the well-defined SeqAn benchmarking tool Rabema [31], developed originally for non-bisulfite reads. In short, Rabema reads one SAM file containing the gold-standard alignments and one containing the mapping results in order to compute the fraction of correctly mapped reads among others. We applied some minor modifications to Rabema, preventing conflicts caused by bisulfite conversions. This enables the comparison of the mapping results to the gold-standard obtained from the read mapping simulation tool.

In the following we evaluate the recall (sensitivity) and the precision. As recall we define the fraction of reads that could be mapped correctly, i.e. the simulated mapping location was found. The precision is defined as the fraction of reads that were mapped correct of all uniquely mapped reads.

This method only evaluates the mapping location in the genome, but not the alignment itself at single-base resolution [1]. Therefore we additionally use the benchmarking methods described in the following in order to get an idea about column-wise accuracy.

SNP and methylation calling

The calling results have a twofold usage. First they serve for the direct benchmarking of the SNP and methylation calling module. Second, they indirectly serve as a measurement for the column-wise accuracy of the underlying alignments, as they depend on correctly mapped bases. For the SNP calling evaluation the following cases are considered:

False negative A simulated SNP was not called.

False positive Nothing was simulated, but a SNP was called.

Wrong called A SNP was simulated, but the wrong SNP type was called.

Right called The correct SNP type was called.

True negative No SNP was simulated and no SNP was called.

[1] Rabema does provide accurate statistics about the alignment edit-distances, but here we need a measurement regarding the position-wise original mappings.

We will have a direct look at the counts of the specific cases and again investigate the recall and the precision. The recall is defined as:

$$\text{recall} = \frac{\text{No. of right calls}}{\text{No. of right calls} + \text{No. of wrong calls} + \text{No. of false negatives}}. \tag{7.1}$$

In order to validate the recall, one has to consider that some regions of the genome might be only poorly covered by reads. SNPs located in these regions are hence not called, so that the recall in these regions is limited. The precision is defined as follows:

$$\text{precision} = \frac{\text{No. of right calls}}{\text{No. of right calls} + \text{No. of false positives} + \text{No. of wrong calls}}. \tag{7.2}$$

Furthermore, the called methylation levels are compared position-wise to the original simulated methylation levels. The deviations are then used for validation.

7.3 Parameter settings

In the following we describe the default settings used in our methods, if no further specifications are given. We performed the three-letter mapping using RazerS3 allowing an error rate of 4% and applying the default settings in all other respects.

For our four-letter pairwise realignment we used in all experiments the threshold of zero for the minimal mapping quality, since RazerS3 does not provide mapping qualities. Further, we used the same error rate threshold of 4% for the raw three-letter alignments (reads above this threshold would be discarded) and a threshold of 5% for the final four-letter alignment. Additionally the simulated deletion error rate was handed over. The parameters for the bisulfite conversion rate and average methylation rate were correspondingly adjusted to the current simulation settings.

The calling was conducted using the maximum likelihood estimation method combined with the Newton optimization. As a minimum coverage for calling we used 6. The minimum mapping quality was set to 1 and the genotype likelihood ratio score threshold was set to 10.

7.4 Existing tools used for comparison

We compare our mapping method against Bismark [38], since it seems to be the most widely used tool out of the three-letter mapping category. It uses Bowtie2 as an external mapper and thus base qualities are taken into account for the scoring function of mismatches (see Section 1.3). We changed the default settings regarding Bowtie2 to allow 1 mismatch (instead of 0) in the seed region.

We additionally compare our method against a tool out of the wild-card mapping category, naturally coming along with a slightly higher sensitivity. Therefore we have chosen Last [24], as we use a similar approach for incorporating base qualities. The alignment score threshold was set to 120, which was recommended by the authors.

we compare the SNP calling and methylation calling module of our workflow against Bis-SNP [46], since it is the only tool we are aware of that combines methylation calling with genotyping. We set the bisulfite conversion rate parameter of Bis-SNP to simulate rate. The threshold for the genotype score was set to 10, the remaining parameters where not adjusted. Bis-SNP provides an additional script for recalibrating base qualities using known SNPs as input. We do not make use of this here, since our tool is developed for de-novo genotyping and we aim for comparable results.

7.5 Experimental results of the core workflow

In order to evaluate the performance of our read mapping and calling methods, we tried various settings for the bisulfite read simulation and for the actual analysis. We will present the impact of the different models used in our method and compare our results to other tools.

The following tests were conducted using a simplified bisulfite setting (bisulfite conversion rate: 0.9999, mean methylation rate: 0.2 for each context, standard deviation: 0.02). The influence of this rates will be investigated in detail in Section 7.5.6.

7.5.1 Impact of the four-letter verification module

In order to get an impression about the importance of the four-letter verification, we compared the intermediate results obtained from the three-letter mapping against the final verified results, using the described default settings. In this setup we can compare the read mappings directly using the Rabema benchmark. Additionally, we run our calling method using once the raw three-letter alignments (with original four-letter sequences) and once the verified mappings. In order to prevent wrong mappings from biasing the result, we used only uniquely mapped reads for the calling.

In Table 7.3 we present the recall and precision respective to the read mapping and SNP calling.

		three-letter mapping	four-letter post-processing
mapping	recall	96.97 %	94.09 %
	precision	97.88 %	99.80 %
calling	recall	80.01 %	87.46 %
	precision	94.16 %	99.19 %

Table 7.3: Recall and precision with and without the four-letter pairwise realigning and verification step. SNP calling: Only uniquely mapped reads were processed.

The results show that the four-letter verification leads to very high precision rates ($> 99\%$), for the mapping and the calling. In contrast, in the three-letter scenario the mapping precision is noticeably lower. This is due to a high fraction of reads that

could not be mapped uniquely and correctly, as the three-letter space increases the ambiguity and may lead to unique three-letter hits that are not unique or even not best in the full four-letter space. However, the calling results prove that the three-letter alignments are relatively poor, since the recall and the precision are clearly improved by verification. Although, this could be caused by wrong mappings or column-wise alignment inaccuracies.

7.5.2 Influence of genomic coverage on SNP and methylation calling

In order to know how to interpret the subsequent results and how exact position-wise analyses can be expected, we will examine the influence of the genomic read coverage. We simulated different numbers of reads leading to different coverages of approximately 12X and 20X.

	≈ 12X	≈ 20X
false negatives	153 913	11 761
false positives	308	1 261
wrong calls	922	1 002
right calls	159 449	301 521
recall	50.73 %	95.88 %
precision	99.23 %	99.26 %

Table 7.4: SNP calling results dependent on the genomic coverage (X).

The results in Table 7.4 show that the SNP recall given the coverage 12X is fairly low. This is probably caused by low covered regions and positions with ambiguous genotypes due to bisulfite conversions. The same holds for the methylation level estimation. Figure 7.5.1 displays the position-wise called methylation levels plotted against the simulated methylation levels.

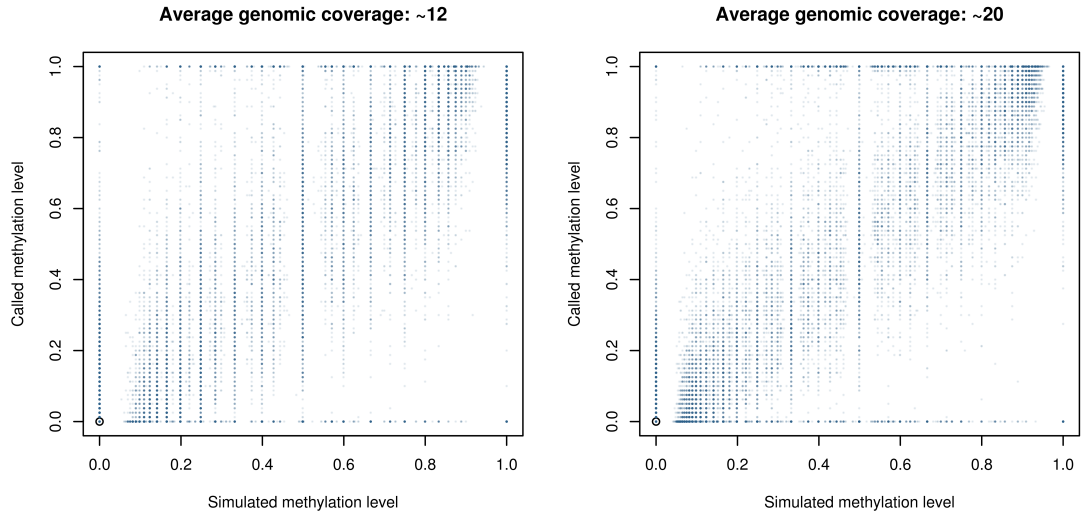


Figure 7.5.1: Comparison of simulated and called methylation levels for different genomic coverages, displayed for the last 200 000 methylation sites of chromosome 21. The transparency indicates the frequency. The accumulation in the bottom left and top right corners represent the bimodal methylation level distributions.

As expected, the higher the coverage the more accurate are the callings. In this case the coverage has an even greater influence, since few reads may not represent the real fraction. The subsequent experiments are performed with the a coverage of approximately 16X.

7.5.3 Influence of base qualities

The mapping and calling methods presented in this thesis make use of the information given by base qualities. In order to evaluate the impact of these, we additionally simulated reads with equal qualities for all bases. Thus, no information about the probability that a base might be wrong is given. In order to point out the impact, we simulated reads with double the sequencing substitution error rate of 0.01. Further we decreased the calling score threshold to 5, in order to illustrate the influence also on low scoring positions. The results in Table 7.5 show that in the case of reads with equal base qualities more false positive SNPs were called, compared to reads with the default quality values of 32 for correct bases and 11 for erroneous bases. In other words, error probabilities can prevent wrong base-calls from causing false SNP calls.

	32 | 11	32 | 32
false negatives	47 152	199 999
false positives	896	2 909
wrong calls	1 203	1 287
right calls	265 929	265 720
recall	84.61 %	84.54 %
precision	99.22 %	98.44 %

Table 7.5: SNP calling results dependent on base qualities (match quality| mismatch quality).

Correspondingly, the erroneous bases or their probability can influence the methylation level estimates, indirectly by the genotype call and directly by biasing the C-T ratio. Figure 7.5.2 displays a slightly increased number of correct called methylation levels, to be more precise an increase of 1 % (or 121 841 sites). Note, that for the methylation level call deviations in the range between 0.25 - 1.0 no noticeable differences were observed.

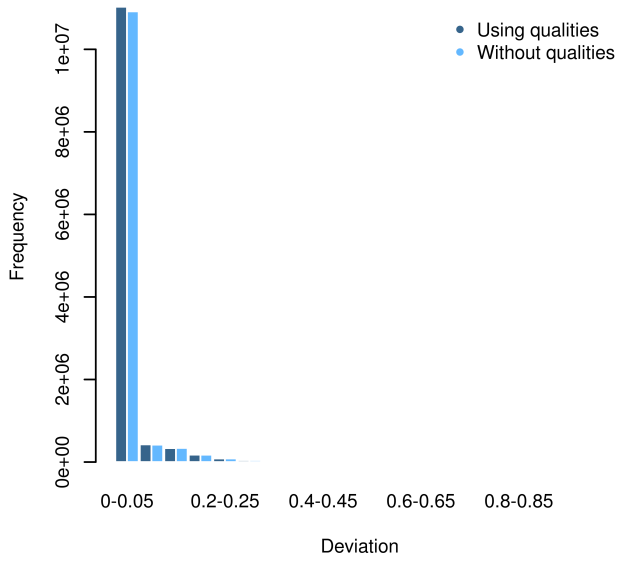


Figure 7.5.2: Deviation of methylation level calls from the simulated methylation levels dependent on base qualities.

7.5.4 Influence of base dependent sequencing errors

Previously we have seen how error probabilities influence the analysis. In the following we will illustrate the impact of the non-uniform sequencing error model used in our methods. Therefore we tested our individual modules using a uniform error distribution model and compared this to using a non-uniform model, while analysing reads containing base dependent sequencing errors. Again a sequencing substitution error rate of 0.01 was used. Since the majority of bases displays a fairly low error probability, we can expect to see only little influence on the overall mapping results and methylation level estimates. Thus we focus here on the SNP calling results, as the analysis at positions with heterozygous SNPs is particularly prone to errors and therefore serves as a good measurement. We decreased the calling score threshold again to 5.

In Table 7.8 the results are shown for the different models.

	a)	b)	c)
false negatives	49 199	49 173	47 152
false positives	865	867	896
wrong calls	1 244	1 244	1 203
right calls	263 841	263 866	265 929
recall	83.95 %	83.96 %	84.61 %
precision	99.21 %	99.21 %	99.22 %
called methylation sites	12 282 245	12 282 440	12 284 186

Table 7.6: SNP calling results and callable methylation sites using a model a) without taking base dependent sequencing errors into account in both modules, b) taking base dependent sequencing errors into account for the four-letter alignment and c) additionally taking base dependent sequencing errors into account for the calling.

The base depend sequencing error probabilities do not have a big influence for the four-letter alignment. Only a few additional SNPs could be called right. The reason is probably that the genomic indel and indel error rates are rather low, so that the base mappings are relatively unambiguous already in the majority of cases and could not be improved.

Interestingly, for the calling method the non-uniform model does have a greater influence. The recall is significantly increased to 84.61 % and higher number of methylation sites could be called. Thus, the respective scores were probably below the calling threshold in the uniform model due to errors. To sum up, base dependent error probabilities do not have a major impact, but can avoid uncertainties caused by errors.

7.5.5 Trade-off between recall and precision for SNP calling

We tested our method using various sensitivity settings and compared the SNP calling results regarding recall and precision. Therefore we modified the threshold for the genotype likelihood ratio score (3, 6, 9, 12), which is required for calling. Figure 7.5.3 illustrates the trade-off between recall and precision.

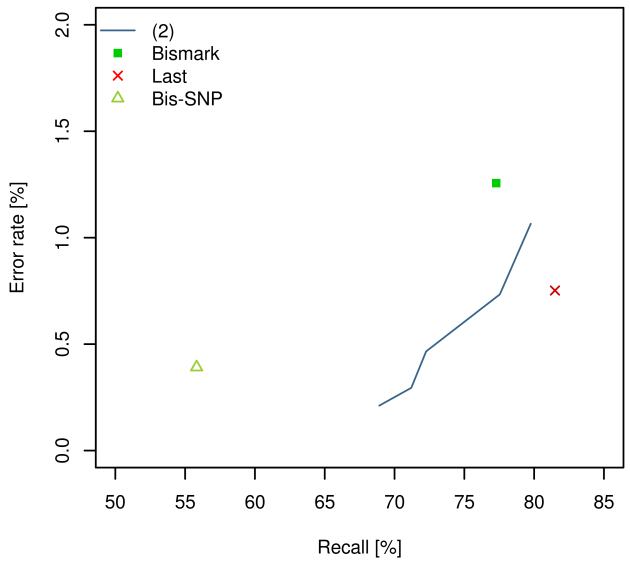


Figure 7.5.3: Accuracy of SNP calling results using different sensitivity settings for our calling method (2). Additionally results from other tools are displayed.
(Error rate = 100% - precision)

The comparison to results of other tools provides an first impression about the performance of our method. The Bis-SNP method was tested using the same set of alignments computed with our bisulfite mapping method and thus serve for a comparison of the SNP calling methods. We can see that for the used error rate, our method provides a significantly higher recall rate than Bis-SNP. However, we believe that the Bis-SNP results can be further optimized by using different parameters as score thresholds and for filtering fake SNPs.

Additionally, we tested our SNP calling method using the alignments of the other mapping tools in order to examine the influence of the alignments on our results. As seen in Figure 7.5.3, the recall for a given error rate is best when using Last alignments. Nevertheless, our method outperforms Bismark using the described settings.

7.5.6 Comparison to existing tools using different methylation rates

In order to ensure that our method is able to analyse reads coming from highly methylated as well as unmethylated regions, we tested it on reads with different methylation rates and compared it to other mapping and calling methods. Therefore we simulated read sets with a mean methylation rate of 0.05, 0.50 and 0.95 for all three contexts. Additionally, we simulated reads with a mean methylation rate of 0.80, 0.10 and 0.05 respectively to the CG, CHG and CGG context, in order to provide reads with realistic methylation rates for comparison.

Mapping We tested the three mapping methods for all four read sets and evaluated the mapping and subsequent SNP calling results.

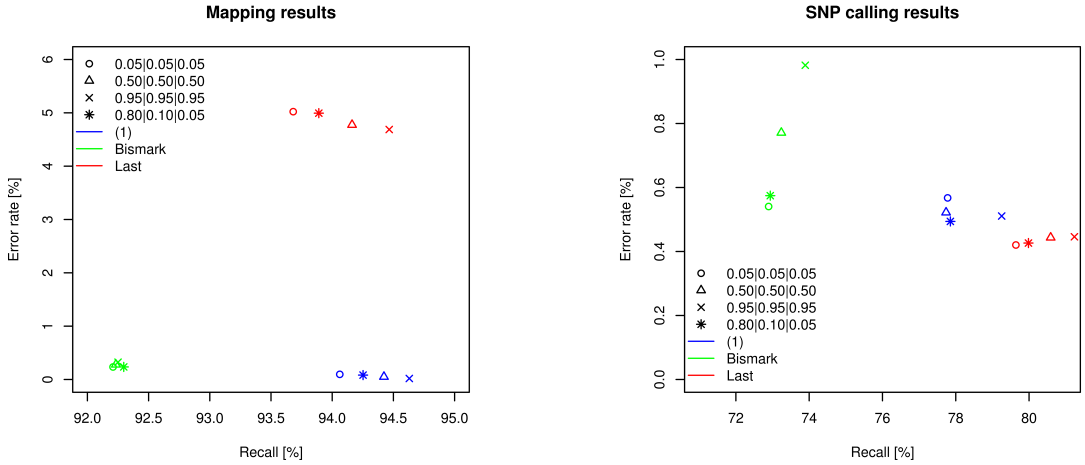


Figure 7.5.4: Recall and error rates for various context dependent methylation rates (CG|CHG|CGG) obtained by our mapping method (1) compared to other mapping methods.

The mapping results in Figure 7.5.4 show that, for our method and for the Last method, the higher the methylation rate, the higher the recall and the lower the error rate. This is expected, since higher methylations rates decrease the ambiguity in the four letter-space and lead to a higher mapping efficiency. In contrast, the Bismark mapping results do not differ significantly for different methylation rates, as reads are mapped only in three-letter space.

The SNP calling results of Last and our method show similar characteristic. Surprisingly, for Bismark alignments the error rate is increased for reads coming from highly methylated regions. We assume that this is caused by wrong read C to reference T mappings, which are allowed in the three-letter alignment and not verified further by Bismark. For low methylation rates such Cs are most likely bisulfite converted and do not cause wrong SNP calls subsequently.

Beside the methylation rate dependencies, the results illustrate the performance of the three tools compared to each other. Clearly, for the given data and settings, our method has a significantly higher recall rate than Bismark. We expected this, since Bismark only computes alignments that are unique in the three-letter space and we have seen in Section 7.5.1 that the precision for such three-letter mappings is relatively low. Consequently the SNP calling results show a higher recall rate for alignments computed by our method.

However, the results for the Last method were not expected. For the mapping they display a consistently higher error rate with a slightly lower recall, whereas the SNP calling results outperform all other results, both in terms of error rate and recall. Probably, this outcome is caused by wrong mappings which are considered for the mapping benchmark, but not for the calling due to low mapping qualities. Furthermore, the calling results indicate that the column-wise accuracy of the Last alignments is relatively high.

Additionally, we evaluated the called methylation levels for the reads with non-uniform methylation rates (0.80|0.10|0.05) in order to compare our method against Bismark and Last. The results are presented in Figure 7.5.5.

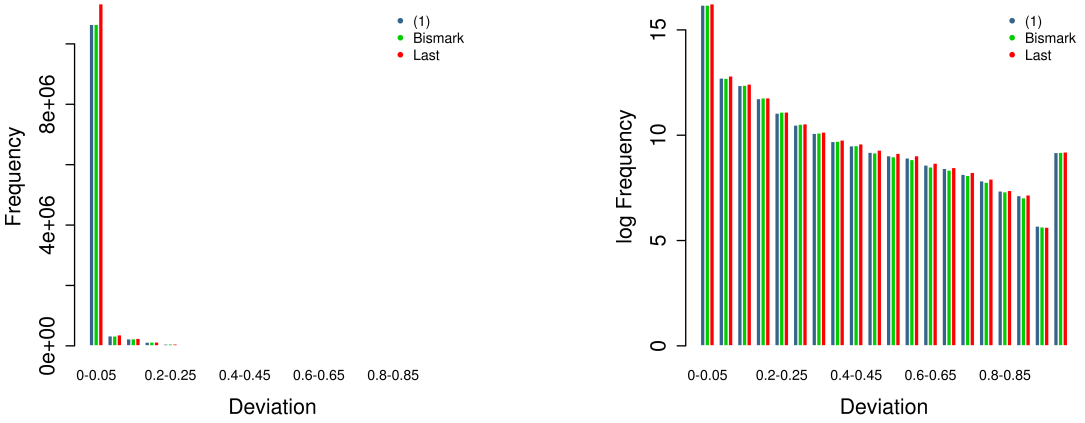


Figure 7.5.5: Differences between simulated and called methylation levels obtained by the calling module performed on alignments of our bisulfite mapping method (1) versus alignments of Bismark and Last. (0.80|0.10|0.05)

We can see that all mapping methods allow for accurate methylation level estimations in the majority of positions. Note that a high fraction of genomic positions is either fully methylated or unmethylated throughout all reads, which benefits highly accurate methylation level calls. However, the Last method leads to a greater number of right called levels, which correlates to the SNP calling results.

Furthermore we evaluated the methylation level calls of our method dependent on the

overall methylation rates. Our method performs quite well for all tested methylation rates, which can be seen in Figure 7.5.6. It holds that the higher the methylation rate, the more methylation levels can be called in total and the higher the fraction of correct calls.

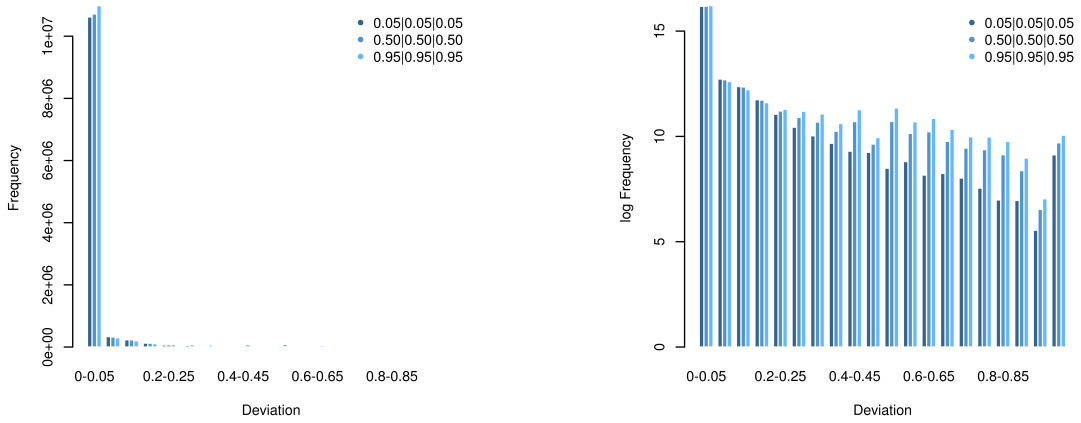


Figure 7.5.6: Deviation of called methylation levels from simulated methylation levels in our method dependent on the simulated methylation rates (CG|CHG|CGG).

The general drawback of wild-card mapping tools such as Last are the biased methylation rates. Due to their four-letter mapping strategy, reads from highly methylated regions show a greater mapping efficiency. Thus the overall genome methylation rates get overestimated. We compared the context dependent methylation rates obtained by the calling module using our mapping method against the Last method.

	simulated	(1)	Last
CG	80.0 %	80.17 %	79.93 %
CHG	10.0 %	10.20 %	10.09 %
CHH	5.0 %	5.42 %	5.37 %

Table 7.7: Average called methylation levels for the three contexts using our mapping method (1) compared to using Last alignments.

Surprisingly, the average methylation rates obtained by using the Last method are not overestimated significantly, as shown in Table 7.7. One reason might be that the methylation levels are simulated for each position independently. This means the methylation levels are equally distributed over the genome rather than building highly or low methylated genomic regions that cause differences in the mapping efficiency.

As the verification in our method takes place in the four-letter space as well, it might also cause minor overestimations. Reads from methylated regions get a higher mapping score in average, since they show a higher uniqueness, and are less often discarded. This bias could be avoided by using only uniquely mapped reads from the three-letter mapping, but this would lower the recall as well.

SNP calling We compared our SNP calling results against the Bis-SNP method, while using again the four read sets with different methylation rates. For both methods the alignments obtained by our mapping method were processed. The parameters regarding sensitivity were adjusted for both methods in order to achieve a comparable precision.

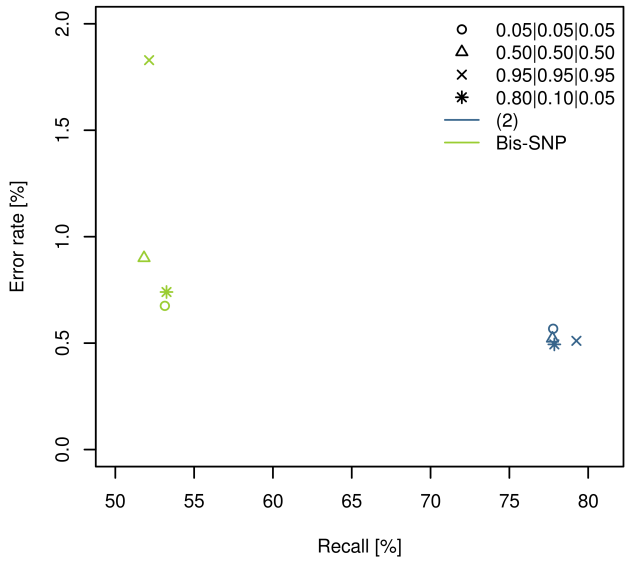


Figure 7.5.7: Accuracy of SNP calling results obtained by our method (2) and by Bis-SNP for different context dependent methylation rates (CG|CHG|CGG).

Figure 7.5.7 illustrates the impact of the different methylation rates on recall and precision. While the results of our method show only minor changes, with low methylation rates causing the highest error rate, the Bis-SNP results differ significantly. The reads with the highest methylation rate cause the highest error rate. One reason are the simulated base dependent error frequencies, which are not taken into account by Bis-SNP. Bisulfite conversion caused Ts are in the case of sequencing substitution errors most frequently called as Cs, which influences the methylation level estimation, but does not lead to erroneous SNP calls. A high methylation rate prevents genomic Cs from being converted. Cs on the other hand are most frequently called as As in case of sequencing

errors. Since our method explicitly makes use of base dependent error probabilities, the results remain unbiased, whereas the Bis-SNP precision decreases for higher methylation rates.

7.5.7 Detailed investigation on SNP and methylation level calling

To better understand the varying results between the Bis-SNP method and our calling method, we performed further tests with a fixed methylation rate (CG:0.8, CHG:0.1, CHH:0.05) and evaluated the SNP and methylation level calls. Since the number and accuracy of methylation level calls highly depends on the parameter defining the minimal genotype likelihood ratio score that is required for the calling, we adjusted the settings to achieve roughly the same number of called methylation sites for both methods. In this way, the accuracy of the methylation level calls can be compared.

In Figure 7.5.8 we present the deviations between the called methylation levels and the simulated levels for both methods. The results show that the Bis-SNP method could achieves a small additional fraction of highly accurate calls (with a deviation between 0.0 - 0.05), which occur in our method with a deviation between 0.05 and 0.2. Methylation levels with a deviation greater than 0.2 occur with rather similar frequencies in both methods. This indicates that the deviations are caused by the data and most probably cannot be further reduced. However, we expect that the observed inaccuracies can be decreased by further parameter optimization in our method.

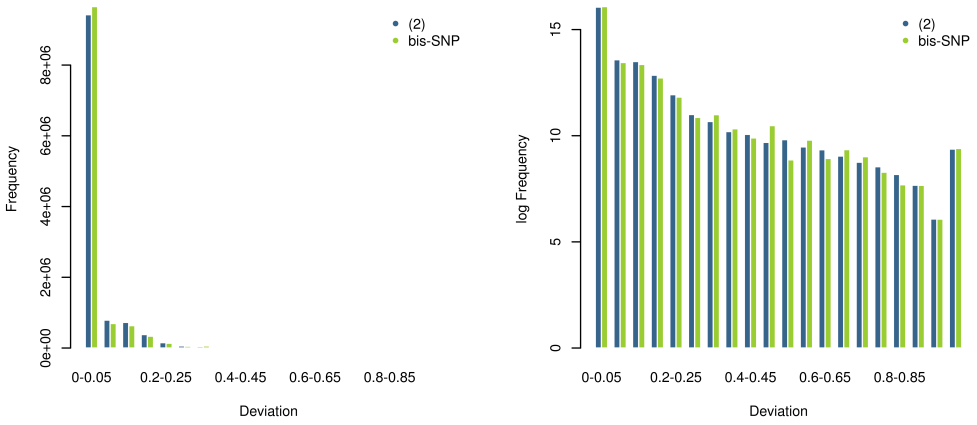


Figure 7.5.8: Differences between simulated and called methylation levels in our method (2) compared to Bis-SNP.

Further, we investigated the SNP calling results of both methods, using a score threshold of 5 for our method. In Table 7.8 we can see that the recall of Bis-SNP is significantly

lower than for our method, while its precision is slightly lower too. We assume that the rather low recall is due to not optimal chosen parameters. For example, Bis-SNP applies a filtering method to discard fake SNPs, based on the average base quality and the coverage (>120) among others, that may be optimized. However, we would assume that a less strict filtering simultaneously would decrease its precision.

	(2)	Bis-SNP
false negatives	69 582	153 432
false positives	495	525
wrong calls	717	778
right calls	244 122	174 750
recall	77.64 %	53.12 %
precision	99.51 %	99.26 %

Table 7.8: SNP calling results of our method (2) and Bis-SNP.

In order to get an impression about the difference between the two methods, we present the number of overlapping results in Figure 7.5.9. The results show that a fairly high fraction of the SNPs called correctly by Bis-SNP is called by our method as well. Surprisingly, the erroneous calls in the two methods are rather distinct.

Figure 7.5.9: Overlap between SNP calling results from Bis-SNP and our method (2). Left: The number of true positive SNP calls. Right: The number of false positives and wrong calls.

7.6 Runtime performance

The BS-Seq analysis workflow was developed aiming for highly accurate results. Nevertheless, the running time is an important factor for analysing real biological data. Therefore we investigated the running time of our tool and compared it to the running

times of the related methods. All benchmarks were conducted on a machine with 2 Intel Xeon X5570 @2.67GHz (Quad Core) processors and 72GB RAM.

Bisulfite read mapping We first run the workflow with the default settings of the previous experiments and compared it against Bismark and Last. As Bismark runs internally two instances of Bowtie2 in parallel (2*1x), one for each strand, we did the same for the Last mapping module [2] and for our three-letter mapping using RazerS3. We run RazerS3 allowing for reads having multiple mappings (all). The results are presented in Figure 7.6.1.

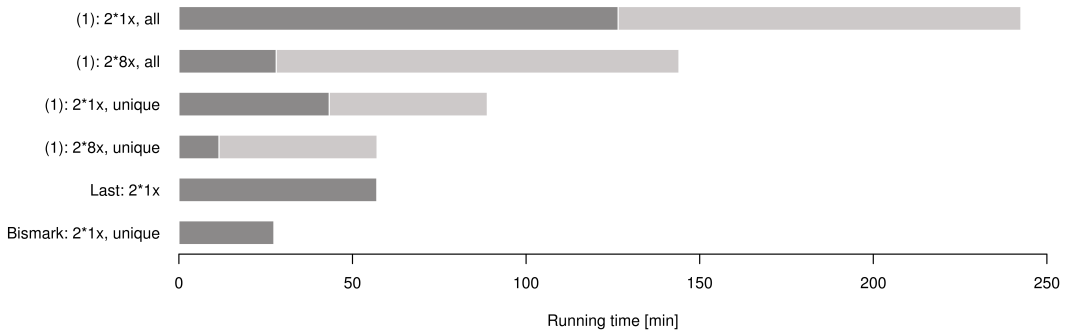


Figure 7.6.1: Running times of the mapping module (1) using different settings in comparison to Bismark and Last. For our method the running time consist of the time used for three-letter mapping with RazerS3 (dark grey) and the four-letter post-processing (light grey). The displayed times illustrate where further improvements may be possible. '2*1x' denotes that two instances of the tool were run in parallel, each on one thread. 'all' and 'unique' means the three-letter mapping tool allows for reads with multiple hits or for only unique hits respectively. For all tools two instances run in parallel, one for the top and one for the bottom strand. Index constructions and three-letter conversions are not included in the measurements.

Bismark, that is based on Bowtie2 for the three-letter mapping, is the by far fastest tool, followed by Last. Bismark has a clear time advantage, since it only computes bisulfite alignments that are unique in the three-letter space. In contrast, our method with default settings computes the best 100 alignments under a given error threshold. Consequently, the running time is longer for RazerS3 and the subsequent four-letter verification, since a significant higher number of reads needs to be realigned and verified. In order to understand the impact of the number of reads, we tested the RazerS3 setting unique hits. This reduced the running time for both individual modules to less than

[2] Last provides individual modules for the mapping and post-processing.

half. Nevertheless, Last and Bismark were still faster.

However, since we did not optimize our method respectively to the running time so far, there is room for further improvements. To get an impression, how parallelisation may improve the performance, we additionally tested our mapping method by running RazerS3 on eight threads in parallel for each strand, which radically reduced the required time. We tried to run Bismark and Last in parallel using 2x8 and 2x4 threads, but an increased memory footprint, exceeding our resources, slowed down both tools.

SNP and methylation calling We compared the running time of our calling method against the Bis-SNP tool, that is mainly written in Java. In Figure 7.6.2 we can see that our method is more than eight times faster than Bis-SNP.

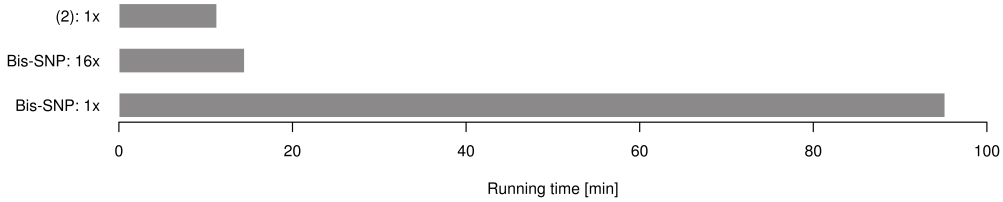


Figure 7.6.2: Running times of our calling method (2) in comparison to Bis-SNP.

The presented results indicate, that further runtime improvements might be possible, which will be discussed at the end of this thesis.

7.7 Experimental results of multiple sequence realignment

In the following we illustrate the potential of the multiple sequence realignment method for reducing the bias caused by genomic indels. We exemplary tested the strategy on a 300 000 bp subset of the human chromosome 21, containing no masked repeat regions. Since the realignment is done locally using already mapped reads, the performance of this method is dependent on the accuracy of the mapped reads, but independent of the reference size. We simulated 80 000 bisulfite reads, with an increased genomic indel rate of 0.02 and a maximal indel length of two bp. The read mapping was performed with equivalent settings to the previous runs.

We present results obtained with our BS-Seq analysis workflow using different settings. No comparison to Bis-SNP is done, as the Bis-SNP realigning method requires known indels as input, which would lead to not comparable results.

7.7.1 Comparison to pairwise alignment computation

In order to illustrate the influence of the multiple sequence realignment, we performed the SNP and methylation calling with and without prior realignment. We used a relatively low threshold of 5 for the minimal calling score, in order to see the influence of the realignment also on positions with ambiguous genotypes.

Due to the frequent genomic indels, the bisulfite read mapping recall dropped down to 63.01%, while the precision remained with 98.98% relatively high. The low mapping recall causes poorly covered genomic regions, such that an overall lower calling recall and precision can be expected. Table 7.9 represents the SNP calling results.

	Pairwise aligning	After local MSR
not called	1246	1192
false positive	240	250
wrong called	44	56
right called	1482	1524
recall	54.32 %	56.11 %
precision	83.91 %	83.28 %

Table 7.9: SNP calling results obtained by using pairwise alignments and after performing a local multiple sequence realignment (MSR).

As shown in Table 7.9, the multiple sequence realignment significantly increases the recall. One drawback is, that the precision is slightly reduced at the same time. However, we expect that this can be improved further by parameter optimization regarding the sensitivity.

For a more detailed evaluation of the column-wise alignment accuracy, we examined the methylation level calls. Furthermore, we tested the method additionally on a smaller read set of 40 000 reads, to illustrate how the realignment affects the calling at poorly covered regions.

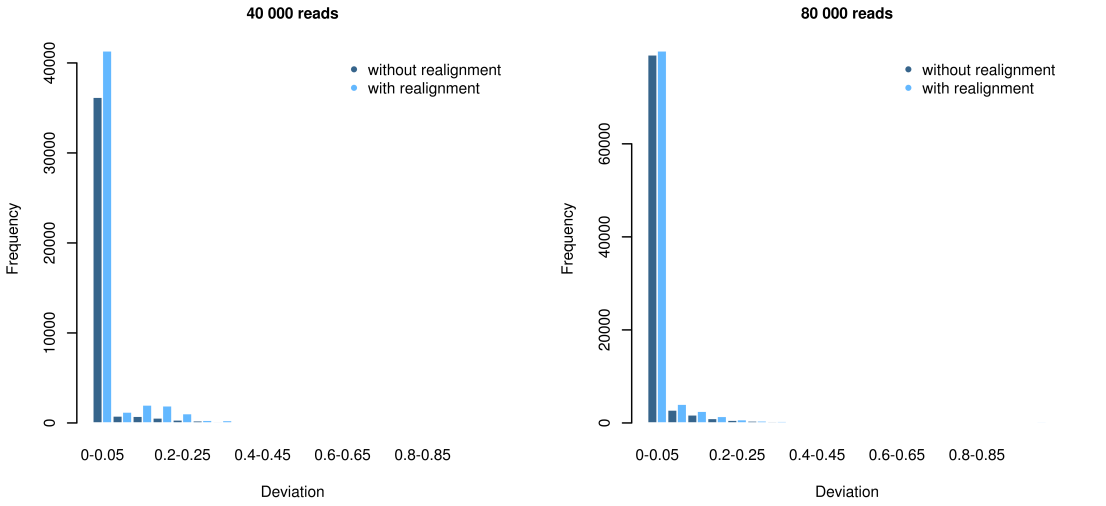


Figure 7.7.1: Deviations of called methylation levels before and after multiple sequence realigning for different genomic coverages.

As can be seen in Figure 7.7.1, the realignment increases the number of highly accurate methylation level calls. Particularly in the case of the lower coverage, the improvement is significant. In total, approximately 20% more methylation sites could be called. This might be caused by indel regions where inconsistently introduced gaps reduce the base coverage at specific genomic positions or wrong base mappings reduce the score and thus prevent the calling. Beside the highly accurate methylation level calls, also calls with a relatively low deviation between 0.05 and 0.25 are called more frequently. These cases can be caused by lower covered indel regions, resulting in less accurate methylation level estimates.

To conclude, the results illustrate the benefit of a multiple sequence realignment for reads containing genomic indels. We used the read mapping tool RazerS3 for the three-letter mapping, which is not designed explicitly for indel reads. Since our BS-Seq analysis workflow can be combined with any external read mapping tool that provides a valid SAM format output, it is possible to use other read mapping tools that explicitly take indels into account. In this way, we expect the mapping and calling recall to increase again to a satisfactory value.

7.7.2 Influence of base dependent sequencing errors

In order to evaluate the impact of non-uniform sequencing errors frequencies for the multiple sequence realignment, we tested our method with and without taking such non-uniform error probabilities into account for the alignment of the current read against the MSA. In Figure 7.7.2 we can see, that the base dependent sequencing error model slightly increases the number of accurately called methylation levels.

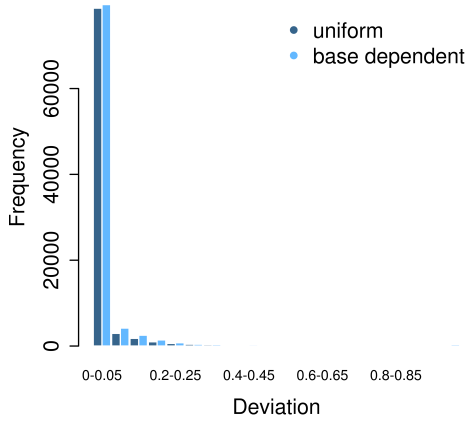


Figure 7.7.2: Deviations between called and simulated methylation levels: Influence of taking base dependent sequencing error probabilities into account for the alignment of a read against a given MSA (using qualities of current read only).

8 Conclusion and Outlook

In this thesis we presented a workflow for BS-Seq data analysis, from bisulfite read mapping to accurate methylation level calling at single-nucleotide resolution. We developed an advanced statistical method making use of base qualities, base dependent error frequencies and BS-Seq specific parameters. The full four-letter space is explored and genomic variations are considered for the methylation analysis.

We implemented a bisulfite read mapping method that handles three-letter alignments of an external read mapping tool. Therefore reads with multiple three-letter hits can be processed. A local pairwise realignment in four-letter space using a statistical model is performed, in order to improve the column-wise accuracy and to determine the true best alignment. We could show that this method radically improved the results compared to the raw three-letter mappings. The comparison to related bisulfite read mapping tools revealed, that our method performs better than the examined three-letter mapping tool Bismark [38], but slightly underlies Last [23] in terms of recall and precision. Simultaneously mapping qualities are computed for the output that serve as a measurement of the uniqueness.

The mapping method provides column-wise accurate alignments, that are analysed in the subsequent calling module. We developed a method based on work from Liu et al. [46] to combine genotype and methylation level calling, that allows for the distinction between genomic SNPs and bisulfite conversions. The algorithm was extended to enable the additional methylation level optimization. Information given by base qualities, base dependent error frequencies and bisulfite conversion rates are taken into account. Regarding the methylation level calls, our method performs comparable to the original method Bis-SNP. However, the SNP calling results indicate the advantage of our method regarding the trade-off between recall and precision. The results obtained by our method provide a well founded base for downstream analyses that require single-nucleotide resolution, such as detection of allele-specific methylations or deviation methylations in cancer cells.

Additionally, we implemented a method for local multiple sequence realignment, to reduce the bias caused by genomic indels. The full four-letter space is used to compute a consistent layout of all reads. This is done by considering alignment gaps for both strands, while distinguishing between the top and bottom DNA strand for base mappings. Our algorithm computes sequence profiles taking error probabilities into account. It iteratively performs a statistical realignment of reads against the MSA using its profile. Experimental results on small datasets illustrated the improvement of the column-wise accuracy revealed by this realignment for reads containing genomic indels.

For convenience, we provided an optional module to merge the information of position-specific estimated methylation levels and given genomic annotations, computing some

basic statistics. Moreover, we developed a bisulfite read simulation method based on Mason [30], able to synthesize reads with accurate BS-Seq specific characteristic and base dependent error rates, allowing for comprehensive benchmarks.

Outlook The methods presented in this thesis are a good foundation for the analysis of BS-Seq data. Nevertheless, there are several aspects that can be optimized and are worth further investigation. First of all, the different methods depend highly on the used parameters, such as sensitivity thresholds, fixed gaps costs, pseudo-frequencies and error thresholds just among others. We could not explore the full parameter space within the scope of this project, but we expect an improved performance by optimizing the settings. Furthermore, we implemented a model for non-uniform SNP frequencies, using different probabilities for transitions and transversions in both the alignment and the calling method. The influence of such specific SNP probabilities was not examined so far, as our read simulation tool uses uniform frequencies. Particularly for the SNP calling, it would be interesting to evaluate the influence.

The pairwise four-letter realignment uses a model not distinguishing between different contexts. In general, the computation of a well-adjusted base mapping score for the bisulfite cases is not trivial, if the methylation probability is taken into account. This is due to the bimodal distributed methylation levels that might cause biases for a relatively high fraction of differently methylated positions. However, it would be worth testing the model with context dependent methylation rates by incorporating neighbor bases into the score.

Another interesting point is the dependency of the SNP and methylation level results on the conducted three-letter mapping strategy, particularly on the number of allowed hits per read. Using unique hits only causes a low mapping precision and thus not optimal calling results, too many allowed hits cause a long running time. Moreover we do not know, if adjusting such settings might influence the overall overestimation of methylation rates. Therefore, it might be possible to find a more suitable trade-off by more detailed analyses.

Regarding the multiple sequence realignment strategy, we definitely think it is worth trying to optimize the scoring function. So far, we could show that the score for aligning one read against an MSA lead already to more accurate alignments around indel regions. However, the strand specific score, distinguishing between top and bottom reads, for mapping one base against one MSA column may lead to inconsistent mappings regarding the two different DNA strands. We propose to try a scoring method that takes the information of the other strand into account. As it is not possible to use frequencies due to possible unknown bisulfite conversions, a score, indicating whether the bases from the other strand support this current base or not, could be used. Differently weighted combinations could be tested. Additionally, it probably would be useful to incorporate base dependent sequencing error probabilities into profile computation.

Beside the mentioned optimizations regarding the accuracy, there is scope for runtime improvements respective to the read mapping module. We precomputed already base, strand and quality value dependent scoring tables in order to save the computation

time. Nevertheless, we expect that the running time can be reduced significantly by optimizing the band size for the pairwise alignment computation, adjusting the number of allowed hits per read, precomputing further tables and most promising by parallelizing the program. This can be done fairly simple, since the reads are processed independently of each other.

Finally, we integrated the presented methods into a workflow engine called KNIME (Konstanz Information Miner) [2], which displays one well-qualified option and is shown in Figure 8.0.1. In doing so we improved the usability and made it more attractive to scientists with no computer science background. However, an extensive testing with end users that could reveal possible pit falls and communication problems should be addressed in the future.

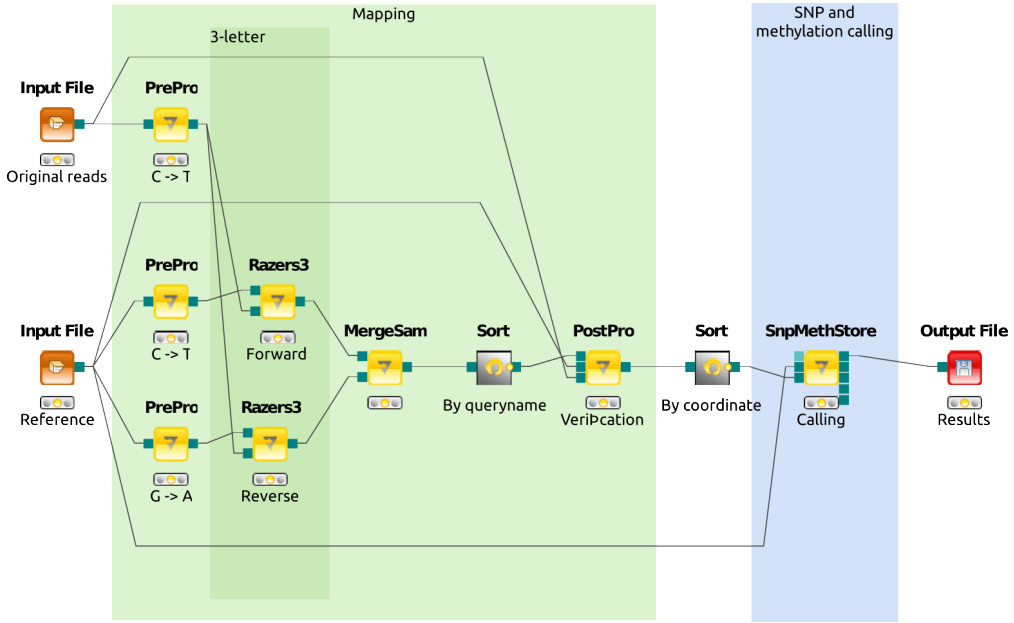


Figure 8.0.1: The core BS-Seq analysis workflow in KNIME for single-end reads.

Bibliography

[1] Eric L Anson and Eugene W Myers. Realigner: a program for refining dna sequence multi-alignments. *Journal of Computational Biology*, 4(3):369–383, 1997.

[2] Michael R Berthold, Nicolas Cebron, Fabian Dill, Thomas R Gabriel, Tobias Kötter, Thorsten Meinl, Peter Ohl, Kilian Thiel, and Bernd Wiswedel. Knime-the konstanz information miner: version 2.0 and beyond. *ACM SIGKDD Explorations Newsletter*, 11(1):26–31, 2009.

[3] *Bis-SNP User Guide v001*. http://epigenome.usc.edu/publicationdata/bissnp2011/BisSNP-UserGuide-latest.pdf, 2013.

[4] Jane Bradbury. Human epigenome project—up and running. *PLoS biology*, 1(3):e82, 2003.

[5] C Sidney Burrus, James W Fox, Gary A Sitton, and S Treited. Horner's method for evaluating and deflating polynomials, 2003.

[6] Aniruddha Chatterjee, Peter A Stockwell, Euan J Rodger, and Ian M Morison. Comparison of alignment software for genome-wide bisulphite sequence data. *Nucleic acids research*, 40(10):e79–e79, 2012.

[7] Pao-Yang Chen, Shawn J Cokus, and Matteo Pellegrini. Bs seeker: precise mapping for bisulfite sequencing. *BMC bioinformatics*, 11(1):203, 2010.

[8] Pao-Yang Chen, Suhua Feng, JW Joo, Steve E Jacobsen, and Matteo Pellegrini. A comparative analysis of dna methylation across human embryonic stem cell lines. *Genome Biol*, 12(7):R62, 2011.

[9] Jung K Choi, Jae-Bum Bae, Jaemyun Lyu, Tae-Yoon Kim, and Young-Joon Kim. Nucleosome deposition and dna methylation at coding region boundaries. *Genome Biol*, 10(9):R89, 2009.

[10] Netta Mendelson Cohen, Ephraim Kenigsberg, and Amos Tanay. Primate cpg islands are maintained by heterogeneous evolutionary regimes involving minimal selection. *Cell*, 145(5):773–786, 2011.

[11] Shawn J Cokus, Suhua Feng, Xiaoyu Zhang, Zugen Chen, Barry Merriman, Christian D Haudenschild, Sriharsa Pradhan, Stanley F Nelson, Matteo Pellegrini, and Steven E Jacobsen. Shotgun bisulphite sequencing of the arabidopsis genome reveals dna methylation patterning. *Nature*, 452(7184):215–219, 2008.

[12] Jeffrey M Craig and Nicholas C Wong. *Epigenetics: a reference manual*. 2011.

[13] Leonardo Dagum and Ramesh Menon. Openmp: an industry standard api for shared-memory programming. *Computational Science & Engineering, IEEE*, 5(1):46–55, 1998.

[14] B Dawes and D Abrahams. Boost c++ libraries (2012).

[15] Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo del Angel, Manuel A Rivas, Matt Hanna, et al. A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nature genetics*, 43(5):491–498, 2011.

[16] Juliane C Dohm, Claudio Lottaz, Tatiana Borodina, and Heinz Himmelbauer. Substantial biases in ultra-short read data sets from high-throughput dna sequencing. *Nucleic acids research*, 36(16):e105–e105, 2008.

[17] Andreas Döring, David Weese, Tobias Rausch, and Knut Reinert. Seqan an efficient, generic c++ library for sequence analysis. *BMC bioinformatics*, 9(1):11, 2008.

[18] Pan Du, Xiao Zhang, Chiang-Ching Huang, Nadereh Jafari, Warren Kibbe, Lifang Hou, and Simon Lin. Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis. *BMC bioinformatics*, 11(1):587, 2010.

[19] Melanie Ehrlich. Dna hypomethylation in cancer cells. 2009.

[20] Anne-Katrin Emde. *Next-Generation Sequencing Algorithms: From Read Mapping to Variant Detection*. PhD thesis, Freie Universität Berlin, 2012.

[21] Manel Esteller and James G Herman. Cancer as an epigenetic disease: Dna methylation and chromatin alterations in human tumours. *The Journal of pathology*, 196(1):1–7, 2002.

[22] Martin C. Frith. Dnemulator. http://www.cbrc.jp/dnemulator/.

[23] Martin C Frith, Ryota Mori, and Kiyoshi Asai. A mostly traditional approach improves alignment of bisulfite-converted dna. *Nucleic acids research*, 40(13):e100–e100, 2012.

[24] Martin C Frith, Raymond Wan, and Paul Horton. Incorporating sequence quality data into alignment improves dna read mapping. *Nucleic acids research*, 38(7):e100–e100, 2010.

[25] Hongcang Gu, Zachary D Smith, Christoph Bock, Patrick Boyle, Andreas Gnirke, and Alexander Meissner. Preparation of reduced representation bisulfite sequencing libraries for genome-scale dna methylation profiling. *Nature protocols*, 6(4):468–481, 2011.

[26] Nicholas J Higham. *Accuracy and Stability of Numberical Algorithms*. Number 48. Siam, 1996.

[27] Emily Hodges, Andrew D Smith, Jude Kendall, Zhenyu Xuan, Kandasamy Ravi, Michelle Rooks, Michael Q Zhang, Kenny Ye, Arindam Bhattacharjee, Leonardo Brizuela, et al. High definition profiling of mammalian dna methylation by array capture and single molecule bisulfite sequencing. *Genome research*, 19(9):1593–1605, 2009.

[28] Robin Holliday. The inheritance of epigenetic defects. *Science*, 238(4824):163–170, 1987.

[29] Robin Holliday. Epigenetics: a historical overview. *Epigenetics*, 1(2):76–80, 2006.

[30] Manuel Holtgrewe. Mason–a read simulator for second generation sequencing data. *Technical Report FU Berlin*, 2010.

[31] Manuel Holtgrewe, Anne-Katrin Emde, David Weese, and Knut Reinert. A novel and well-defined benchmarking method for second generation read mapping. *BMC bioinformatics*, 12(1):210, 2011.

[32] Michelle D Johnson, Michael Mueller, Laurence Game, and Timothy J Aitman. Single nucleotide analysis of cytosine methylation by whole-genome shotgun bisulfite sequencing. *Current Protocols in Molecular Biology*, pages 21–23, 2012.

[33] Peter A Jones. The dna methylation paradox. *Trends in Genetics*, 15(1):34–37, 1999.

[34] Samuel Karlin and Stephen F Altschul. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Sciences*, 87(6):2264–2268, 1990.

[35] Antoine Kerjean, Annick Vieillefond, Nicolas Thiounn, Mathilde Sibony, Marc Jeanpierre, and Pierre Jouannet. Bisulfite genomic sequencing of microdissected cells. *Nucleic acids research*, 29(21):e106–e106, 2001.

[36] Szymon M Kiełbasa, Raymond Wan, Kengo Sato, Paul Horton, and Martin C Frith. Adaptive seeds tame genomic sequence comparison. *Genome research*, 21(3):487–493, 2011.

[37] Felix Krueger. Sherman - bisulfite-treated read fastq simulator. http://www.bioinformatics.babraham.ac.uk/projects/sherman/.

[38] Felix Krueger and Simon R Andrews. Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. *Bioinformatics*, 27(11):1571–1572, 2011.

[39] Peter W Laird. Principles and challenges of genome-wide dna methylation analysis. *Nature Reviews Genetics*, 11(3):191–203, 2010.

[40] Julie A Law and Steven E Jacobsen. Establishing, maintaining and modifying dna methylation patterns in plants and animals. *Nature Reviews Genetics*, 11(3):204–220, 2010.

[41] Heng Li, Jue Ruan, and Richard Durbin. Mapping short dna sequencing reads and calling variants using mapping quality scores. *Genome research*, 18(11):1851–1858, 2008.

[42] Ruiqiang Li, Yingrui Li, Xiaodong Fang, Huanming Yang, Jian Wang, Karsten Kristiansen, and Jun Wang. Snp detection for massively parallel whole-genome resequencing. *Genome research*, 19(6):1124–1132, 2009.

[43] Ryan Lister and Joseph R Ecker. Finding the fifth base: genome-wide sequencing of cytosine methylation. *Genome research*, 19(6):959–966, 2009.

[44] Ryan Lister, Ronan C O'Malley, Julian Tonti-Filippini, Brian D Gregory, Charles C Berry, A Harvey Millar, and Joseph R Ecker. Highly integrated single-base resolution maps of the epigenome in¡ i¿ arabidopsis¡/i¿. *Cell*, 133(3):523–536, 2008.

[45] Ryan Lister, Mattia Pelizzola, Robert H Dowen, R David Hawkins, Gary Hon, Julian Tonti-Filippini, Joseph R Nery, Leonard Lee, Zhen Ye, Que-Minh Ngo, et al. Human dna methylomes at base resolution show widespread epigenomic differences. *nature*, 462(7271):315–322, 2009.

[46] Yaping Liu, Kimberly D Siegmund, Peter W Laird, Benjamin P Berman, et al. Bis-snp: Combined dna methylation and snp calling for bisulfite-seq data. *Genome Biol*, 13(7):R61, 2012.

[47] Jinyu Wu Luke You. *BSSim User manual, BSSim: Bisulfite sequencing simulator for next-generation sequencing.* http://122.228.158.106/BSSim/BSSim

[48] Frank Lyko, Sylvain Foret, Robert Kucharski, Stephan Wolf, Cassandra Falckenhayn, and Ryszard Maleszka. The honey bee epigenomes: differential methylation of brain dna in queens and workers. *PLoS biology*, 8(11):e1000506, 2010.

[49] Alexander Meissner, Tarjei S Mikkelsen, Hongcang Gu, Marius Wernig, Jacob Hanna, Andrey Sivachenko, Xiaolan Zhang, Bradley E Bernstein, Chad Nusbaum, David B Jaffe, et al. Genome-scale dna methylation maps of pluripotent and differentiated cells. *Nature*, 454(7205):766–770, 2008.

[50] André E Minoche, Juliane C Dohm, Heinz Himmelbauer, et al. Evaluation of genomic high-throughput sequencing data generated on illumina hiseq and genome analyzer systems. *Genome Biol*, 12(11):R112, 2011.

[51] Christian Otto, Peter F Stadler, and Steve Hoffmann. Fast and sensitive mapping of bisulfite-treated sequencing data. *Bioinformatics*, 28(13):1698–1704, 2012.

[52] Brent Pedersen, Tzung-Fu Hsieh, Christian Ibarra, and Robert L Fischer. Methylcoder: software pipeline for bisulfite-treated sequences. *Bioinformatics*, 27(17):2435–2436, 2011.

[53] Mattia Pelizzola and Joseph R Ecker. The dna methylome. *FEBS letters*, 585(13):1994–2000, 2011.

[54] Vardhman K Rakyan, Thomas Hildmann, Karen L Novik, Jörn Lewin, Jörg Tost, Antony V Cox, T Dan Andrews, Kevin L Howe, Thomas Otto, Alexander Olek, et al. Dna methylation profiling of the human major histocompatibility complex: a pilot study for the human epigenome project. *PLoS biology*, 2(12):e405, 2004.

[55] Tobias Rausch, Sergey Koren, Gennady Denisov, David Weese, Anne-Katrin Emde, Andreas Döring, and Knut Reinert. A consistency-based consensus algorithm for de novo and reference-guided sequence assembly of short reads. *Bioinformatics*, 25(9):1118–1124, 2009.

[56] Frederick Sanger, Steven Nicklen, and Alan R Coulson. Dna sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467, 1977.

[57] Robert Shoemaker, Jie Deng, Wei Wang, and Kun Zhang. Allele-specific methylation is prevalent and is contributed by cpg-snps in the human genome. *Genome research*, 20(7):883–889, 2010.

[58] Andrew D Smith, Wen-Yu Chung, Emily Hodges, Jude Kendall, Greg Hannon, James Hicks, Zhenyu Xuan, and Michael Q Zhang. Updates to the rmap short-read mapping software. *Bioinformatics*, 25(21):2841–2842, 2009.

[59] Andrew D Smith, Zhenyu Xuan, and Michael Q Zhang. Using quality scores and longer reads improves accuracy of solexa read mapping. *BMC bioinformatics*, 9(1):128, 2008.

[60] Thanh H Vu, Tao Li, Danielle Nguyen, Binh T Nguyen, Xiao-Ming Yao, Ji-Fan Hu, and Andrew R Hoffman. Symmetric and asymmetric dna methylation in the human igf2–h19 imprinted region. *Genomics*, 64(2):132–143, 2000.

[61] David Weese, Manuel Holtgrewe, and Knut Reinert. Razers 3: faster, fully sensitive read mapping. *Bioinformatics*, 28(20):2592–2599, 2012.

[62] David M Woodcock, Celine B Lawler, Martha E Linsenmeyer, Judith P Doherty, and William D Warren. Asymmetric methylation in the hypermethylated cpg promoter region of the human l1 retrotransposon. *Journal of Biological Chemistry*, 272(12):7810–7816, 1997.

[63] Yuanxin Xi and Wei Li. Bsmap: whole genome bisulfite sequence mapping program. *BMC bioinformatics*, 10(1):232, 2009.

List of Figures

List of Tables