2

DNA Methylation Profiling from High-Throughput Sequencing Data

Michael Hackenberg, Guillermo Barturen and José L. Oliver
Dpto. de Genética, Facultad de Ciencias, Universidad de Granada,Granada, Lab. de Bioinformática, Inst. de Biotecnología, Centro de Investigación Biomédica, Granada, Spain

1. Introduction

The methylation of a cytosine at the carbon 5 position (5meC) is a common epigenetic mark in eukaryotic cells that is normally found in CpG and CpHpG (H=A,C,T) sequence contexts. The inactivation of one of the X-chromosomes in female cells, the allele-specific expression of imprinted genes or the role in the embryonic development which is shown by the early lethality of Dnmt3a and Dnmt3b deficient mice are only some of the important functions that DNA methylation plays (Chen and Li, 2004; Dodge, et al., 2005; Karpf and Matsui, 2005; Okano, et al., 1999). The methylation of the gene promoter region is commonly associated with silenced transcription; however recently it was shown that the DNA methylation in the gene body of transcribed genes is increased in both animals and plants (Hellman and Chess, 2007; Jones, 1999; Zhang, et al., 2006). Furthermore, CHG methylation relates to the silencing of transposons in plants (Miura, et al., 2009) . Given all these facts, it is clear that the DNA methylation pattern along the genome sequence carries valuable biological information and is crucial for our understanding on gene expression and developmental control. Furthermore, it can reveal how aberrant epigenetic changes might lead to dysregulation of gene expression and to the development of diseases such as cancer (Costello, et al., 2000).

A broad panoply of techniques to detect DNA methylation has been developed. The DNA methylation is erased by PCR and not detected by hybridization as the methyl-group is located within the major groove and not at the hydrogen bonds. Therefore, virtually all techniques rely on a methylation dependent pretreatment of the DNA before hybridization, amplification or sequencing. The three main classes of pretreatments are: digestion by methyl-sensitive endonucleases, methyl-sensitive immunoprecipitation and bisulfite conversion (reviewed by (Laird, 2010)). An important impulse for epigenetic research was the adoption of DNA microarrays to methylation profiling (Estecio, et al., 2007). This technique was initially used together with a methyl-sensitive digestion of the DNA, however meanwhile also immunoprecipitation and bisulfite conversion variants do exist (Bibikova and Fan, 2009). Microarrays usually (there are arrays for individual CpGs which cover several thousand sites) allow to obtain information of the "mean" methylation values of a given region, however the methylation pattern cannot be revealed at a single base pair

resolution. The generation of whole genome, single-base-pair resolution methylation maps became feasible just recently with the advent of Next-Generation Sequencing (NGS) or High-Throughput Sequencing (HTS) methods like those from Illumina, Roche 454 and Solid (applied biosystems) to mention just the 3 with the highest diffusion (Shendure and Ji, 2008). These techniques are frequently termed whole-genome shotgun bisulfite sequencing and have been applied already in several methylome projects (Bock, et al., 2011; Laurent, et al., 2010; Lister, et al., 2009). The denatured genomic DNA is treated with sodium bisulfite which leads to the deamination of cytosines, preserving however methylcytosines. After sequencing the treated DNA, the methylation state of the individual cytosines can be profiled directly from the aligned sequence reads: an unconverted cytosine indicates methylation while a thymine instead of a cytosine will reveal an unmethylation. In order to obtain sufficient coverage, the genome is (re)sequenced at a typical coverage of 15x which obviously implies a notable bioinformatics challenge. Limitations and main bias effects of genome-wide DNA methylation technologies, as well as the factors involved in getting an unbiased view of a methylome have been recently reviewed (Robinson, et al., 2010).

In this chapter we will review the common steps in the analysis of whole genome single-base-pair resolution methylation data including the pre-processing of the reads, the alignment and the read out of the methylation information of individual cytosines. We will specially focus on the possible error sources which need to be taken into account in order to generate high quality methylation maps. Several tools have been already developed to convert the sequencing data into knowledge about the methylation levels. We will review the most used tools discussing both technical aspects like user-friendliness and speed, but also biologically relevant questions as the quality control. For one of these tools, NGSmethPipe, we will give a step by step tutorial including installation and methylation profiling for different data types and species. We will conclude the chapter with a brief discussion of NGSmethDB, a database for the storage of single-base resolution methylation maps that can be used to further analyze the obtained methylation maps.

2. Analysis workflow

A general analysis workflow to convert the sequencing data into methylation maps can be divided into 3 parts: (i) pre-processing of the reads, (ii) alignment and (iii) the profiling of the methylation states from the alignments. In all three steps several error controls can be applied which are therefore discussed in a separate section below. Some of these steps are shared by virtually all of the so far published tools, others however are unique to a single or few applications. In this section we will give the theoretical background and in the next section we will compare the different implementations into the software tools.

2.1 Preprocessing of the reads

The pre-processing of the reads can be grouped into: (i) elimination or manipulation of low quality reads, and (ii) preparation of the reads for the alignment step. It is known that Illumina sequence reads loose quality towards the 3' end. In order to only use the high quality part of the read, Lister *et al* (Lister, et al., 2009) proposed to trim the read to before the first occurrence of a low quality base call (PHRED score <= 2). Another step which might increase the alignment accuracy is the removal of the adapter sequences. If the DNA

fragment is shorter than the read length, parts of the 3' adapter will also be sequenced. The adapter sequence will however not align to the genome which might lead to missed or incorrect mapping of the read. Depending on the alignment algorithm the adapter removal is a mandatory step before mapping the reads to a reference genome. For example, some programs perform a seed alignment, i.e. not the whole read is aligned but only a given subsequence of the 5' part of the read: the seed. If the adapter does not extend into the seed, those reads can be aligned even without removing the adapter. However, there are also methods that align the whole read. Those algorithms will very likely fail to map reads that contain adapter sequences (see section 3 for more details).

Like mentioned before, and explained in detail within the next section, many methods are based on alignments using a 3-letter alphabet. Those programs need to manipulate the reads before the alignment step replacing all remaining cytosines (those that are not converted by the bisulfite) by a thymine (see Figure 1).

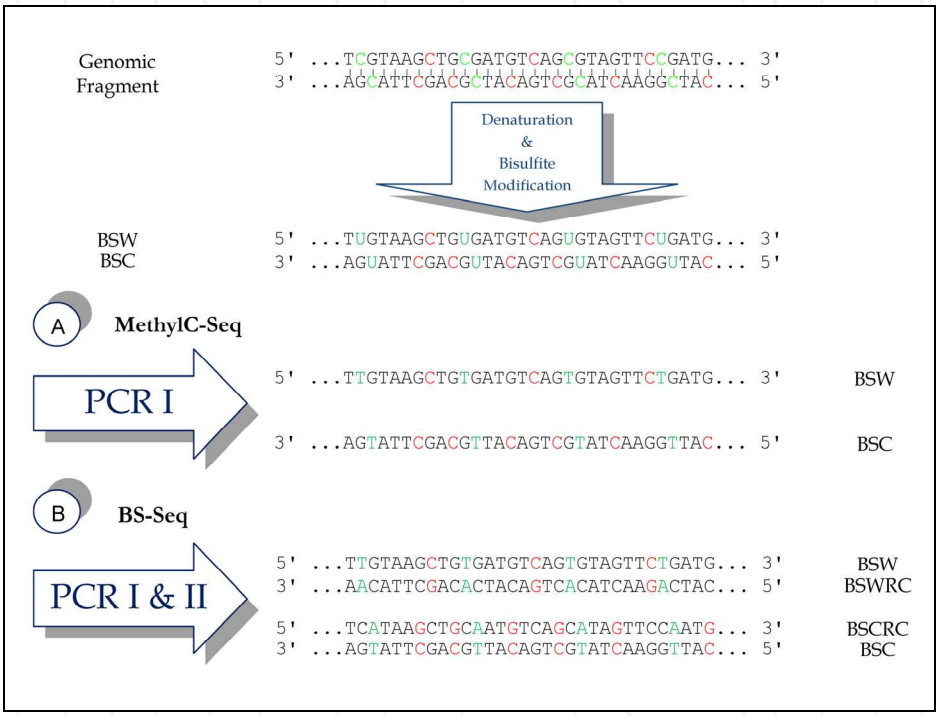


Fig. 1. Whole genome bisulfite sequencing: MethylC-Seq and BS-seq. After denaturing and bisulfite treatment, the genome DNA will lose the strand complementarity, as unmethylated cytosines are converted to uracils (green coloured cytosines). During the PCR, the uracils will be substituted by thymines. Here we show an illustration for the reads from MethyC-Seq (directional) and BS-Seq (non-directional). A) The MethylC-Seq protocol generates the library in a directional manner, resulting in either BSW (Bisulfite Watson) or BSC (Bisulfite Crick) reads. B) The BS-Seq protocol performs two consecutive PCRs which yields BSW and BSC reads, as well as their reverse complementary strands (BSWRC and BSCRC).

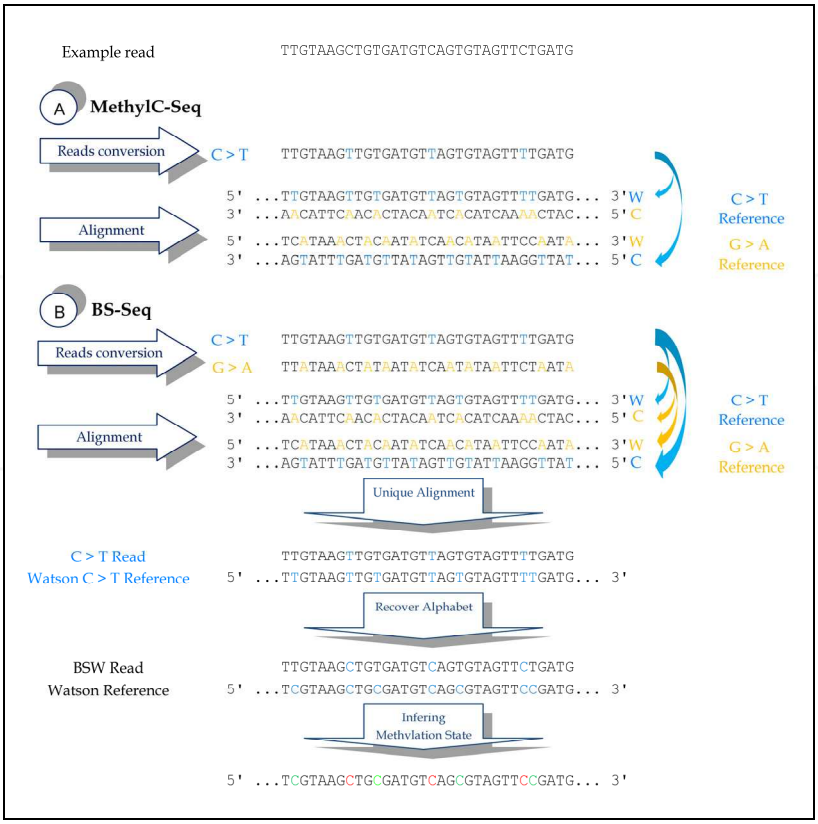


Fig. 2. Methylation profiling from bisulfite-treated reads. In order to retrieve the methylation information, the bisulfite treated reads must be first mapped to the reference genome. During the experimental protocol the unmethylated cytosines are converted to thymines. One way to deal with this reduced sequence complexity is to convert both, the reads and the reference sequence into a 3-letter alphabet. In the MethylC-Seq protocol (A), the reads come from the Watson (BSW) and Crick (BSC) bisulfite treated DNA fragments. Therefore, all cytosines of the input reads will be converted to thymines and will be tried to align to the C to T converted reference (blue arrows). For the BS-Seq protocol (B), the reads can map to the Watson and Crick C to T converted reference (reads BSW and BSC), as well as to the Watson and Crick G to A converted reference (reads BSWRC and BSCRC). Therefore, the reads must be converted into two different alphabets: C to T changed reads that will be mapped to the C to T converted reference (blue arrows) and G to A changed reads that will be mapped to the G to A converted reference (yellow arrows). If a best unique alignment exists for a read, both, the original read and reference sequences are recovered which allows directly to read out the methylation information: a C/T mismatch indicates an unmethylated cytosine (green colored), a cytosine in both the read and the reference indicates methylation (red colored). In case of reverse complement reads (G to A changed), a guanine in the reference and an adenine in the read indicates unmethylation and a guanine match between the read and reference sequence allows to infer methylation.

2.2 Alignment

After the treatment of the genomic DNA with sodium bisulfite, the sequence complexity is reduced as all cytosines with the exception of methylcytosines are converted into thymine. In mammal genomes methylation normally occurs exclusively in a CpG sequence context which only accounts for approx. 1.8% of all dinucleotides and therefore the majority of all cytosines is converted to T. This reduced complexity complicates the alignment process, i.e. the detection of the genome loci where the read originates from. A common philosophy in all developed programs and protocols is to just accept unique alignments, i.e. just one genome position exists to which the read aligns under a given set of parameters. In many cases it might occur that one and the same read has a unique alignment in a 4-letter presentation but exhibits mappings to several positions with the same quality in a 3-letter alphabet. In such cases, the information carried by the read is lost. The existence of sequencing errors complicates the correct read mapping further. An additional challenge in bisulfite read mapping is the increased search space. This is because the Watson and Crick strands of bisulfite treated DNA sequences are not complementary to each other as the bisulfite just acts on cytosines. As a consequence, both bisulfite Watson (BSW) and Crick (BSC) have their own reverse complementary strands, BSWRC and BSCRC. Frequently used experimental protocols are MethylC-Seq (Lister, et al., 2008), RRBS (Reduced Representation Bisulfite Sequencing, (Meissner, et al., 2005)) and BS-seq (Cokus, et al., 2008) (reviewed by ((Lister and Ecker, 2009))). Without going into the experimental details, the relevant difference is that in the BS-seq protocol two subsequent PCRs are performed which leads to the generation of four read types, BSW, BSWRC, BSC and BSCRC. On contrary, MethylC-Seq only generates BSW and BSC reads (see Figure 1). This difference has immediate consequences on the bioinformatics analysis (see figure 2).

After the bisulfite treatment, the reads cannot be aligned simply to the reference sequence as converted cytosines would lead to mismatches. In theory, it would be possible to increase the number of allowed mismatches in order to align the reads to the genome, however this would entail some serious drawbacks: (i) a high number of allowed mismatches will make the alignments less specific, i.e. a higher number of incorrect mappings will be the consequence leading to the incorrect inference of the methylation states, (ii) it would be virtually impossible to profile CpG dense regions (CpG islands) as those are frequently un-methylated presenting a high number of converted cytosines. Another possibility would be to generate for each read all combinations of possible T/C conversions which seems however computationally too demanding.

Given that we can rule out the increase of the allowed mismatches as a solution, two general approaches have been proposed to align bisulfite treated sequence reads. First, an alignment matrix can be used which gives the same weights to all matches and C/T mismatches (cytosine in the reference sequence and T in the read) and second, a three letter alphabet can be used to adapt the reference sequence to the reduced sequence complexity of bisulfite treated reads (see Figure 2 for an illustration). In theory, the first method could be slightly more accurate due to a higher sequence complexity, however it seems also clearly slower (see section 3.3). In the 3-letter alphabet approach, as a direct consequence of the lost of complementarity of bisulfite treated DNA, two different reference genomes must be prepared: 1) substituting all Cs by Ts and 2) substituting all Gs by As. Depending on the

concrete experimental protocol, the reads need to be aligned only against the Watson strand from the C/T reference and the Crick strand from the G/A reference (MethylC-Seq) or against both strands of the two reference genomes.

2.3 Post-processing and output

Once the reads are aligned to the reference genome, the methylation information of the individual cytosines can be read out. In order to do so, both, the reads and the reference sequences need to be converted back to a 4-letter alphabet (see Figure 2). Methylcytosines are then indicated by C/C matches while unmethylated cytosines are given by a T/C mismatch in the alignment (also G/G and G/A in the case of BS-Seq). The methylation level of a given cytosine position in the genome is determined by the information of all reads that overlap this position. The methylation level of a cytosine is given simply by the number of methylcytosines divided by the total number of reads that map to the position. In this way, the methylation level is a number between 0 (completely unmethylated) and 1 (completely methylated). There are at least two reasons which can lead to intermediate methylation levels: (i) usually a cell population is used to extract DNA and intermediate values can indicate fluctuations at a given position between the individual cells and (ii) allele specific methylation is a well known phenomenon in imprinting. Bisulfite treatment together with sequencing has the advantage that the methylation level of each individual cytosine can be assessed. That means that, unlike in many other techniques not only the methylation levels of CpGs can be determined but also other sequence contexts like CHG or CHH. These sequence contexts will be particularly interesting in plants or embryonic stem cells (Lister, et al., 2009).

2.4 Quality control

There are several error sources which can compromise the quality of the methylation maps which ideally should be taken into account. This error sources include (i) wrong alignment of the reads, (ii) existence of Single Nucleotide Variants (SNV), (iii) sequencing errors, (iv) bisulfite failure.

2.4.1 Incorrect alignments

The reduced sequence complexity of bisulfite treated reads and the existence of sequencing errors can lead to wrong alignments, i.e. the read is aligned to a genome position where it does not originate from. This is particularity true for highly repetitive DNA sequences which are frequently CpG rich (like Alu retro-transposons) and methylated. When single end reads are used in the experimental assay, wrong mappings cannot be detected and the only way to control the number of incorrect alignments is the appropriate choice of the alignment parameters. The search for the best alignment is a highly parameterized task and each alignment algorithm has its own set of parameters depending whether seed alignment is possible, the base call probabilities are considered, etc. This is probably the main reason why no large scale comparison exists in order to fix the best parameters for the different algorithms. Apart from that, it is clear that the number of allowed mismatches and the minimum alignment length (length of the seed alignment in the methods were it applies) are crucial in the mapping accuracy. A higher number of allowed mismatches and a short alignment seed will permit to map more reads to the genome, however, the number of

incorrect alignments will also increase (high sensitivity, low specificity). On the other hand, very strict parameters will impede the alignment of many "valid" reads and therefore many genome regions will be failed to be profiled. On contrary, the usage of paired-end or mate reads bears the big advantage that a considerable number of the wrong mappings can be detected and removed. In the paired-end technique, both ends of the DNA fragment are sequenced. Normally, the approximate fragment length distribution is known and therefore a narrow window on the genome can be established to which both reads must map. In this way, if the two mate reads are independently mapped to the genome and the best alignment is on different chromosomes or distanced far away on the same chromosome, these mappings can be eliminated as at least one of the two alignments will be incorrect.

2.4.2 Detection of SNVs

One frequently occurring type of variation are the Single Nucleotide Variants (SNVs) which are variations in just one nucleotide between the reference sequence and the sequenced genome. Many of these SNVs might be SNPs as their frequency is higher than 1% in the population. Nevertheless, we will call them SNVs as this is a more general and population genetics independent concept. Over two third of all SNPs are known to occur at the cytosine in a CpG sequence context (Tomso and Bell, 2003). These SNPs have usually two alleles, C and T. Given this high number of C/T SNPs, it can be supposed that the percentage of unknown C/T SNPs and C/T SNVs will be very similar. In the case of a C/T variation, normally the sequenced genome carries a thymine while in the reference genome a cytosine is annotated. If the presence of this sequence variant is unknown or ignored, the inference would be that the cytosine annotated in the reference genome is unmethylated. The correct conclusion however would have been that no cytosine exists in the genome, and therefore no methylation state can be detected. One possibility to take into account the existence of variation is to query a SNP database eliminating all positions with C/T alleles. The disadvantage is that also valuable information is lost when the sequenced genome carries the same allele as the reference genome. Another possibility is to detect the SNVs directly by means of the sequencing data which is possible for positions with a sufficiently high coverage. A C/T variation would manifest on the complementary DNA strand as an adenine, while bisulfite deamination would not affect the guanine on the complementary strand as the bisulfite is applied to denatured DNA (Weisenberger, et al., 2005). Currently, the detection of SNVs is exclusively implemented in the NGSmethPipe program while all other algorithms would interpret such C/T SNVs erroneously as unmethylated cytosines.

2.4.3 Bisulfite failure

Incomplete bisulfite conversion can be caused either by incomplete denaturing before applying the bisulfite treatment or reannealing during the bisulfite conversion. In any case, if the bisulfite has not acted the cytosines would remain unconverted within the read independently of its methylation state. If such reads are not detected, all cytosines would be inferred to be methylated. Lister et al. proposed to use non-CpG contexts to detect reads that are likely not bisulfite converted. This protocol proposes to discard those reads with more than 3 methylated cytosines in a non-CpG context. This measure should work fine in organisms or cell types were non-CpG methylation is virtually absent, however it might discard many valuable information when real non-CpG methylation exists like in embryonic stem cells.

2.4.4 Sequencing errors

Another source of incorrect profiling of the methylation state is the erroneously calling of a thymine instead of a cytosine. Such sequencing errors would be incorrectly interpreted as an unmethylated cytosine. Each base of a read has assigned a sequencing quality in form of a Phred Score which can be interpreted as the probability of the base to be incorrectly called. Therefore, in theory a probabilistic approach could be applied in order to control for the incorrect profiling due to sequencing errors (see section 3.6.2).

3. Bioinformatics tools

Several sofware applications and protocols have been developed so far. In this section we will discuss 9 tools: mrsFAST (Hach, et al., 2010), RMAP (Smith, et al., 2009), SOCS-B (Ondov, et al., 2010), BS-seeker (Chen, et al., 2010), BSMAP (Xi and Li, 2009), BRAT (Harris, et al., 2010), MethylCoder (Pedersen, et al., 2011) , Bismark (Krueger and Andrews, 2011) and NGSmethPipe (Barturen, et al., Submitted)). In table 1, the availability and some basic features are displayed.

Basically, we can distinguish two types of tools: (i) bisulfite alignment tools that perform the pre-processing and alignment but do not report methylation levels and (ii) full pipeline tools that perform all necessary steps from the pre-processing over the alignment to the methylation profiling and error control. We will concentrate the discussion mainly on the full pipeline tools as those will be the choice of many users with little or no bioinformatics background, mentioning the bisulfite aligners for advanced users. Apart from the available software packages, several protocols have been used and proposed (Bock, et al., 2010; Cokus, et al., 2008; Gu, et al., 2010; Harris, et al., 2010; Lister, et al., 2008; Lister, et al., 2009). We will mention those protocols whenever any of the implemented analysis steps have been proposed before in any of these works. Note finally, that several other programs like Methyl-Analyzer (Xin, et al., 2011) have been developed to generate methylation maps at a single-base-pair resolution for non-bisulfite data. Many of the analysis steps are shared by the tools; however, the concrete implementation might vary if other parameter sets are applied. Therefore, we will not discuss each tool within a separate section but analyze the differences directly within the section on the different analysis steps.

3.1 Implementation

The programming language is highly related to many features as the installation process and the speed. Most of the programs discussed here are implemented in C(++) and need to be compiled locally which might require the help of a system administrator in order to install the program. On the other hand, all tools that rely on an external alignment program are based on an interpreted scripting language like Perl (Bismark, NGSmethPipe) or Python (BS-seeker, MethylCoder, Methyl-Analyzer) which are normally installed on a standard Linux distribution. This implies a relatively easy installation or set-up process. Some of these tools rely on additional Perl or Python modules which can be easily installed from the command line. Therefore, the installation process is very similar for all of the Perl/Python based tools. Roughly, the set-up includes the following steps: (i) download the perl or python scripts to a local directory, (ii) "install" the alignment program (for example, Bowtie is delivered as binary files which just need to be downloaded), (iii) install additional

Program	Availability	Language	Sequence space / Color space	Multi-threads	Scope
mrsFAST	http://mrsfast.sourceforge.net/	C	yes/no	No	BS align
RMAP	http://www.cmb.usc.edu/people/andrewds/rmap/	C++	yes/no	No	BS align
SOCS-B	http://solidsoftwaretools.com/gf/project/socs/	C++	no/yes	Yes	BS align
BS Seeker	http://pellegrini.mcdb.ucla.edu/BS_Seeker/BS_Seeker.html	Python	yes/no	yes*	BS align
BSMAP	http://code.google.com/p/bsmap/	C++	yes/no	Yes	BS align
BRAT	http://compbio.cs.ucr.edu/brat/	C++	yes/no	No	Full
MethylCoder	https://github.com/brentp	Python/C	yes/yes	No	Full
Bismark	http://www.bioinformatics.bbsrc.ac.uk/projects/bismark/	Perl	yes/no	yes*	Full
NGSmethPipe	http://bioinfo2.ugr.es/NGSmethPipe	Perl	yes/no	Yes	Full

Table 1. The availability and basic features of the different software tools. The asterisk indicates those programs that have multi-threading in the alignment process through Bowtie, but all pre and post-processing steps are single-threaded. The column scope refers to whether the program reports the methylation levels of the individual cytosines or if only the bisulfite alignment is performed leaving the read out of the methylation information to the user.

modules if applies and (iv) prepare the reference genomes. For the last point, all programs provide scripts that take a multifasta file or a directory name as input yielding the two 3-letter reference genomes as output. Currently, all programs are tested only on Unix, Linux and/or Mac OSX platforms and no Windows support is available.

3.2 Input data and scope

The programs differ quite notably in the accepted input data and the number of implemented features (see table 1-3). Currently just one program, Methyl-Coder can be used for both, sequence space (Illumina, Roche 454) and color space input (SOLiD). With the exception of SOCS-B (SOLiD), all other tools presented here can only handle sequence space input. Another important difference is the availability to process the data from the different library preparation protocols (BS-seq, MethylC-Seq) and single/paired end reads. While all tools implement the MethylC-Seq protocol for single reads, mrsFAST, RMAP and Methyl-Coder do not implement BS-Seq. Methyl-Coder supplies a script that allows to convert BS-seq data into MethylC-Seq ("*tagged_reads_prep.py*"). Paired end support for directional reads is currently available by all tools with the exception of RMAP, SOCS-B and BS-Seeker while paired-end support for non-directional reads is only available in Bismark.

Program	Aligner	Method	Seed	Q	Single (non-dir)	PE (dir)	PE (non-dir)	MP	strand merge
mrsFAST	mrsFAST	3-letter	Yes	No	No	Yes	No	No	No
RMAP	RMAP	Matrix	Yes	Yes	No	No	No	No	No
SOCS-B	SOCS	Matrix	No	Yes	Yes	No	No	No	No
BS Seeker	Bowtie	3-letter	No	No	Yes	No	No	No	No
BSMAP	SOAP	4-letter	Yes	No	Yes	Yes	No	No	No
BRAT	BRAT	2-letter	Yes	No	Yes	Yes	No	Yes	No
Methyl-Coder	bowtie/ GSNAP	aligner dependent	Yes/ Yes	Yes/ No	No	Yes/ Yes	No	Yes	No
Bismark	Bowtie	3-letter	Yes	Yes	Yes	Yes	Yes	Yes	Yes
NGSmethPipe	Bowtie	3-letter	Yes	Yes	Yes	Yes	No	Yes	Yes

Table 2. Alignment method and output features. The 'Seed' column indicates whether a seed alignment methods is used, the Q column indicates whether the Phred Score base call quality values are used in the alignment process. The column PE (directional) and PE (non-directional) refer to the capability to perform paired end alignments for the two main sequencing protocols: MethylC-Seq (directional) and BS-Seq (non-directional). The columns 'MP' and 'strand merge' refer to the output options indicating whether the methylation profiling (MP) step is performed and if a strand merged output for palindromic sequences is possible. Note that, all programs that perform methylation profiling can report the methylation values for different sequence contexts (CpG, CHG, CHH).

3.3 Alignment

A huge number of short read alignment programs and methods have been developed so far (review by (Horner, et al., 2010)). Some of the tools discussed here implement its own alignment algorithm like mrsFast, BRAT or RMAP while others are based on an external aligner. Those that are based on an external aligner implement the conversion into a 3-letter alphabet as the mapping to a bisulfite treated 4-letter alphabet is only possible when manipulating internal parameters which are not accessible. In theory, every alignment method could be used to map the bisulfite treated reads to the 3-letter reference sequences. Currently, one of the most successful programs is Bowtie (Langmead, et al., 2009) which is used by Bismark, BS-seeker and NGSmethPipe. Methyl-Coder allows to chose between Bowtie and GNASP (Pedersen, et al., 2011). Although most full pipeline programs are based on Bowtie alignments, this does not at all imply that the programs produce identical or even similar results. There are basically two reasons: First, the alignment of short reads is a highly parameterized process and each program uses a different set of default parameters which in some cases cannot be changed and second, the programs differ greatly in the amount of implemented quality control features. The alignment parameters are crucial in order to control for the number of wrong alignments which is especially true for single end reads. Like mentioned before, relaxed parameters will lead to a high coverage (many cytosines can be profiled) however, a high number of incorrect alignments can also be expected leading to wrong methylation profiling. On the contrary, strict parameters might lead to a low coverage discarding a considerable amount of valuable information. Despite of this importance, the tools usually recommend some default parameters, however without basing them on a strict

analysis. The first study trying to fix the best parameters for single end reads was carried out by (Barturen, et al., Submitted). These authors, in order to detect the best parameter set for the NGSmethPipe tool, used a golden standard generated by paired end reads to measure the alignment accuracy as a function of the seed length and number of mismatches.

Some alignment programs use the Phred Scores that are assigned to each base (see table 2) to improve the mapping accuracy. The reasoning is the following: mismatches of low quality bases are more likely due to sequencing errors while mismatches of high quality bases are more likely due to incorrect alignments. To take this fact into account, Bowtie calculates the sum of the Phred scores of all mismatches discarding an alignment if this sum (e value) is higher than a given threshold (70 by default, which corresponds to more than 2 high quality mismatches).

Finally NGSmethPipe performs a different 'best alignment' detection compared to other programs. The method is similar to the one used in the miRanalyzer tool (Hackenberg, et al., 2011) and functions as follows: (i) using Bowtie, a seed alignment (40 bp by default) with a given number of allowed mismatches in the seed (1 MM by default) is performed and the N best alignments (N depends on the number of possible read orientations, obtaining for each possible orientation up to 5 alignments) are obtained (--best --strata), (ii) the maximal N alignments are extended until the next mismatch occurs, (iii) if only one unique longest alignment exists (non-ambiguous mapping) after the extension step, this alignment is retained and extended as long as the 'global' number of allowed mismatches is not exceeded. Usually, just unique seed alignments are used for the methylation profiling and the seed alignment extension method proposed here has the advantage that it can disambiguate some read mappings leading to a higher number of mapped reads without compromising the quality.

3.4 Speed

Given that a resequencing experiment can easily produce up to 3.000 million sequence reads of length that currently vary between 36 - 100 nt, it is clear that the alignment speed is an important issue. Most of the CPU is consumed by the alignment process and by some specific output options like the strand joining which is however not implemented by all programs. A sound comparison of the speed performance for all 9 tools is currently impossible. The reason is that a comparison is only meaningful if it was performed on the same platform and CPU configuration which was not carried out for all these tools together. The Bismark authors for example compared their tool with the performance of BS Seeker on a set of reads containing approx. 15 million reads taken from SRR020138 (Lister, et al., 2009). The number of aligned reads and CPU time is 9,633,448 (64.2%)/42 min and 9,664,184 (64.4%)/29 min for Bismark and BS Seeker respectively. The BS Seeker authors (Chen, et al., 2010) compared the speed of their tool to RMAP showing that BS Seeker is over 10 times faster. Note that, the huge BS Seeker speed increase over RMAP is due to the faster Bowtie alignment algorithm while the slightly higher speed performance compared to Bismark (also based on Bowtie) is due to the different parameter set and choice of unique alignments. Finally, mrsFast was reported to be around twice as fast as Bowtie (Hach, et al., 2010). In summary, no final conclusion on the speed performance can be drawn as every author compared their tools to just one or very few other tools. It seems however, that the Bowtie based full pipeline programs and mrsFast might be the fastest tools currently available.

There are at least two other factors that influence the performance which are the availability of multi-threading (using more than one CPU) and memory issues. Right now, many of the programs discussed here do not support multi-threading. For example, mrsFast performs very well but it does run on only one CPU and it can be therefore easily be outperformed by programs based on Bowtie which supports multi-threading. Currently only NGSmethPipe and BSMAP are completely parallelized while Bismark for example uses the multi-threading capacity of Bowtie but all pre and post-processing steps are only single threaded. Finally, NGSmethPipe allows to adapt the memory usage to the resources of the user (see section 4).

3.5 Output

There are two important output features that a software tool should ideally implement: (i) the methylation levels for each cytosine in a given sequence context for each strand and (ii) an output for palindromic sequence contexts (like CpG) in which the information from both strands is merged together. For example, it has been observed that hemi-methylation (strand specific methylation of a palindromic sequence) is quite uncommon, and therefore the methylation levels from both strands can be merged together assigning the methylation level to the position of the C in the plus strand. Right now, only Bismark and NGSmethPipe implement the strand merge for palindromic sequence contexts. Furthermore, NGSmethPipe gives in the output the difference of the methylation levels between the two strands. In this way, hemi-methylation can be detected easily. Finally, NGSmethPipe is the only program that detects SNVs which are also reported in BED file format.

3.6 Quality control

The most disregarded aspect in the methylation profiling is the strict control of the different error sources. Most of the available programs do not implement quality controls like bisulfite failure check, removal of low quality reads or bases and the detection of sequence variation that can lead to a wrong inference of the methylation state. Currently, NGSmethPipe is the only algorithm that takes into account all these error sources.

In order to detect SNVs, the information from both strands needs to be considered. If NGSmethPipe detects a C/T mismatch after reconverting the alignments to a 4-letter alphabet, the information from the other strand is accessed: if the complementary base is an 'A' this means that the C/T mismatch is caused by an SNV, if on the other strand is a 'G', this means that in the genome we have an C:G pair and that the C/T mismatch is caused by the conversion of an unmethylated cytosine into a thymine. This is the theoretical background, however the experimental data will be influenced by fluctuations and noise, and therefore it is possible to find for one and the same position both, C/C and C/T on the forward strand and A and G on the reverse strand. Therefore, we need to parameterize this model reporting a SNV if a given threshold is surpassed.

NGSmethPipe first checks if at least one read exists that contains a C/T conversion for a given position. For these positions, the program calculates the "SNV-fraction" as the number of reads that have not got a G on the complementary strand (an A, C or T might indicate the existence of a SNV) divided by the total number of reads that map to the position on the complementary strand. We define a C/T conversion as caused by an SNV if the "SNV-fraction" is above a given threshold which is set by default to 0.7 in the tool.

Program	Trim reads	Trim tags	Remove adapter	Clonal reads	Bisulfite failure	Base call errors	SNVs
mrsFAST	no	No	No	No	No	No	No
RMAP	no	No	No	No	No	No	No
SOCS-B	yes	No	No	No	No	No	No
BS Seeker	no	Yes	Yes	No	Yes	No	No
BSMAP	yes	No	Yes	No	No	No	No
BRAT	yes	No	No	Yes	No	No	No
MethylCoder	no	No	No	No	No	No	No
Bismark	no	No	No	No	No	No	No
NGSmethPipe	yes	Yes	Yes	No	Yes	Yes	Yes

Table 3. Quality control features. The table indicates if the following quality control features are implemented into the programs: *trim reads*: the low quality end of the input reads is removed; *trim tags*: the tags in the BS-seq protocol are detected and removed, if the tag is not detected the read is trimmed on both ends by the tag size; *remove adapter*: detect the adapter sequence and remove it from the read; *clonal reads:* eliminate duplicated reads; *bisulfite failure:* detect those reads for those the bisulfite didn't acted; *Base call errors:* discard the information for low quality base calls; *SNVs:* detect single nucleotide variation and discard these positions.

3.6.1 SNV detection

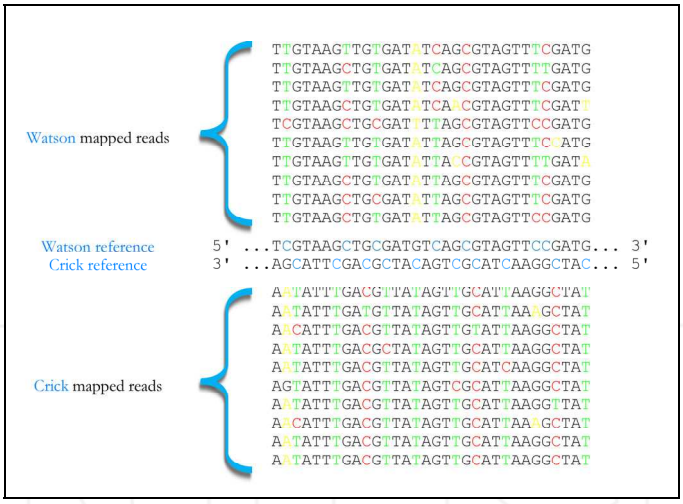


Fig. 3. SNV detection. The figure shows an illustration of the SNV detection model and methylation profiling for reads from a MethylC-Seq protocol. Above the reference sequences the reads that map to the Watson strand are shown and below those that map to the Crick strand. Nucleotides colored in red mean potentially methylated positions (C/C matches); the green ones, potentially unmethylated positions (C/T mismatch and G's in the opposite strand) and the yellow one, mismatches to the reference sequence. If a position has both, yellow and green columns it is a good candidate to be detected as a SNVs. Table 4 shows the results inferred from this diagram.

Context	Start	Watson Level	Crick level	Merged Level	Watson SNV fraction	Crick SNV fraction	Result
CG (+/-)	2	0.1	0.2	0.15	0	0.9	rejected
CTT (-)	4	-	0	-	0	-	unMeth
CWG (+/-)	8	0.6	0.9	0.75	0	0	interMeth
CG (+/-)	11	0.2	0.1	0.15	0	0	unMeth
CAT (-)	13	-	0	0	1	-	rejected
CWG (+/-)	17	0.3	0.1	0.2	0.2	0	unMeth
CG (+/-)	20	1	0.9	0.95	0	0	meth
CTA (-)	22	-	0.1	-	0	-	unMeth
CCG (+)	27	0.2	-	-	-	0.2	unMeth
CG (+/-)	28	0.8	0.9	0.85	0.1	0	meth
CAT (-)	30	-	0	-	0.2	0	unMeth

Table 4. Inference from example in figure 3.As it has been explained in section 2.4.2, the existence of SNVs are an important error source in the profiling of methylation values. The table shows, both the methylation profiling and the detection of SNVs from the example in figure 3. Context, refers to the sequence context in the reference genome and in brackets their orientations are given. Start indicates to the coordinate of the first base in the Watson strand. Watson level and Crick level are the methylation levels for the two strands (fraction between methylcytosines and total number of reads). The merged level is the global fraction of methylcytosines for palindromic sequence contexts taking into account the cytosines of both strands. Watson SNV and Crick SNV are the fractions of mismatches (generally A/G mismatches) in the reverse complementary strand position. The last column shows the inferred methylation state (in red we mark the value on which the inference is based). No methylation level is assigned ('reject') if the 'SNV fraction' is above the threshold of 0.7, methylation ('meth') if the level is above 0.8, unmethylation ('unmeth')for levels below 0.2 and intermediate methylation between 0.2 and 0.8.

3.6.2 Sequencing errors

The contribution of the individual bases can be controlled as a function of their base call quality. Each base has assigned a Phred Score which varies between 0-93 in Sanger format and 0-64 in illumina 1.3+ format although in NGS sequencing no higher values than 60 in Sanger format and 40 in illumina 1.3+ format are normally achieved. The Phred Score has a very easy interpretation. For example, a score of 10 indicates a probability of 0.1 that the base call is wrong, while a Phred Score of 20 corresponds to a probability of 0.01 etc. The Phred Score threshold can be seen as an upper limit of wrongly inferred methylation states. For example, when setting Q>=20 (just accept bases with a probability less than 0.01 to be incorrectly called), less than 1% of all inferred methylation states are incorrect.

3.6.3 Bisulfite failures

NGSmethPipe implements the method proposed by Lister *et al.* to detect those reads where the bisulfite probably failed to act. While Lister et al discard all reads with more than 3 methylated cytosines in a non-CpG context, NGSmethPipe allows to set two different

thresholds: (i) the absolute number of methylcytosines in a non-CpG context can be set as threshold and (ii) the fraction of methylcytosines and total number of non-CpG cytosines can be used with the advantage that the fraction is independent of the read length. As default parameter, a fraction of 0.9 is used.

3.6.4 Paired end incorrect alignment detection

Like mentioned before, by means of paired end reads the number of incorrect alignments can be lowered drastically. Some of the bisulfite aligners do not support paired end reads, however, all of the full pipeline programs discussed here do (see table 2).

4. Brief guideline to analyze the data with NGSmethPipe

NGSmethPipe is a program to analyze bisulfite sequencing data generated by means of BS-Seq and MethylC-Seq protocols. It is implemented in Perl and can be easily set-up as it needs no compilation. The modular structure allows both, running just one of the different sub-tasks or a full analysis by means of a meta-script. The main focus of NGSmethPipe is on the quality control of the generated methylation maps, implementing several quality related features that are currently only available in NGSmethPipe.

4.1 Main features

NGSmethPipe implements several exclusive features including both, aspects regarding the quality of the inferred methylation levels and technical issues like full multithreading (parallelization of the process) and memory scaling (the memory needs of the program can be adjusted to the resources of the user). The main properties of NGSmethPipe can be summarized as follows:

- The program implements three important quality filters: (i) putative bisulfite failures can be detected, (ii) the number of false inferences on the methylation state can be controlled by means of the Phred Scores (probability of a sequencing error), (iii) SNV (single nucleotide variants) can be detected and removed
- Usage of a "seed extension" method applied to the Bowtie alignments allowing to map a higher number of reads without compromising the mapping quality
- Extensive output options including all possible cytosine sequence contexts (CG, CHG and CHH; where H is A, T or C) and the possibility to join the information from both strands (detection of hemi-methylation)
- Complete statistics of the whole process, including aligned reads, discarded reads, discarded positions, chromosome data coverage, etc
- The memory and CPU needs can be adapted to the user's computer resources
- Single and paired end input is accepted
- Fastq input files are accepted in zip, gzip, bzip2 or uncompressed

4.2 Installation

Right now, NGSmethPipe is tested only for Linux platforms. In this section we will assume that the user has a Linux workstation with an installed Perl interpreter (by tying 'perl -v' in a

terminal it can be checked easily if Perl is installed). The installation or set-up process can then be done in 3 simple steps:

1. Install two Perl modules that are needed. This can be done typing with super-user rights on the command line:
   - perl MCPAN –e "install Bundle::Thread"
   - perl -MCPAN -e "install IO::Uncompress::AnyUncompress"
2. Download the bowtie aligner http://bowtie-bio.sourceforge.net/index.shtml and extract the binary files to a folder (for example '/home/user/bowtie').
3. Download the NGSmethPipe files from http://bioinfo2.ugr.es/NGSmethPipe/downloads/NGSmethPipe.tgz and uncompress them (tar xzvf NGSmethPipe.tgz). For the rest of this chapter we will assume that the directory is '/home/user/NGSmethPipe'. After uncompressing the tar file, the user should see 4 Perl scripts which are explained in the next section.

4.3 Structure of the program

NGSmethPipe is composed of 4 scripts which can be launched individually or as a pipeline by means of a meta-script. The scripts have the following functions:

- 'NGSmethPipeIndex.pl': Generation of two 3-letter reference sequences in bowtie format: one changing C by T in the Watson strand (C to T reference in figure 2) and one changing A by G in the Watson strand G/A Watson conversion (G to A reference in figure 2)
- 'NGSmethPipeAlign.pl': Pre-processing and alignment of the reads
- 'NGSmethPipeRatios.pl': Methylation profiling and quality control
- NGSmethPipe.pl: This is the meta-script; it launches internally the other 3 scripts described above

4.4 Running NGSmethPipe

In this section, we will show how to prepare the reference sequences and how to perform the methylation profiling for different data sets.

4.4.1 Building the reference sequences

Before NGSmethPipe can be used, the reference genome must be downloaded in fasta format (for example from the UCSC Genome Bioinformatics site http://hgdownload.cse.ucsc.edu/goldenPath/hg19/chromosomes/) into a local directory (for example, 'home/user/sequences/hg19'). Note that in the current version, all chromosomes, contigs or scaffolds must be in single-fasta file format. No multi-fasta files are currently supported. The reference sequences can then be prepared using the following command:

perl NGSmethPipeIndex.pl seqDir=/home/user/sequences/hg19 bowtieDir=/home/user/bowtie

The only optional parameter is the number of threads (number of CPUs) to be used (p=4 by default). The output of the process consists of 12 bowtie index files and is also written into the input directory (seqDir). Bowtie generates for each reference 6 files. Here we obtain 12

files as we need two different 3-letter references: one changing C by T and one substituting G by A. Note that the meta script checks if converted reference sequences do exist and therefore, this step can be skipped when performing a full data analysis by means of the meta-script.

4.4.2 Download the example data

On the NGSmethPipe webpage (http://bioinfo2.ugr.es/NGSmethPipe/downloads/Examples.tgz), some examples can be downloaded to test the program. The test data contains both, the input reads and the reference sequences. After downloading (we will assume that the download directory is /home/user/Examples) and uncompressing with "tar xzvf NGSmethPipe.tgz", the user can see three folders: h1_exampleChr22 which contains sequencing data from the MethylC-Seq protocol for a region on the human chromosome 22, wt_shoots_example which contains reads from BS-Seq for *Arabidopsis thaliana* and wa09fibro_exampleChr21 which contains pair-end reads for a region in human chromosome 21.

4.4.3 Running the example data

In general, NGSmethPipe can be launched in two different ways: (i) by means of command line parameters or (ii) by using a configuration script. The parameter syntax is the same for both ways: parameter=<value>. An example for the configuration script can be seen within the three test data folders. To analyze MethylC-Seq data, only three mandatory parameters do exist, 'seqDir' (the folder with the reference genome single fasta files), 'inDir' (the input directory with the read files in (compressed) fastq format and 'bowtieDir' (the folder with the Bowtie binary files). The program can be launched giving the parameters on the command line:

perl NGSmethPipe.pl seqDir=/home/user/Examples/h1_exampleChr22/seqDir/ inDir=/home/user/Examples/h1_exampleChr22/inDir/ bowtieDir=/home/user/bowtie/

or by means of the configuration script:

perl NGSmethPipe.pl /home/user/Examples/h1_exampleChr22/NGSmethPipeConfigFile_h1.dat

If the first parameter on the command line is a file, NGSmethPipe will read and treat it automatically as a configuration file. These two commands will launch a full analysis with default parameters by consecutively launching the 3 scripts. Each of the scripts will write its own output files which are explained in detail in the section 4.4.5.

4.4.4 Important parameters

Multithreading: Two parameters, '*p*' and '*maxChunk*' allow to adapt the memory and CPU requirements of the process to the local resources. The number of CPUs can be controlled by '*p*' (p=6 will generate 6 threads) and the memory needs can be fine-tuned by means of the '*maxChunk*' parameter. By default, 10000 reads are processed by each thread which leads to a memory of approx. 2500Mb per thread for the human genome. The more reads are processed by a thread, the higher will be the addressed memory, but the speed will increase also as the hard disk access times decrease. It is important to adjust these parameters in

order to exploit the available resources: number of processors and random access memory (the available RAM of the computer). For example, a high number of threads could increase the speed, however it will also increase the overall memory usage and together with a high value for 'maxChunk' it could exhaust the available random memory.

perl NGSmethPipe.pl seqDir=/home/user/Examples/h1_exampleChr22/seqDir/ inDir=/home/user/Examples/h1_exampleChr22/inDir/ bowtieDir=/home/user/bowtie/ p=6 maxChunk=20000

BS-Seq protocol: For non-directional reads we need to provide the sequences of the forward and reverse tags which can be indicated by the '*fw*' and '*rc*' parameters respectively. Example reads can be found in the folder wtshoots_example (extracted from (Pellegrini, et al., 2010)).

perl NGSmethPipe.pl seqDir=/home/user/Examples/wtshoots_example/seqDir/ inDir=/home/user/reads/Examples/wtshoots_example/inDir/ bowtieDir=/home/user/bowtie/ fw=TCTGT rc=TCCAT

Pair-end protocol: In the case of pair-end reads the user has to provide the file suffix for each of the two mate files by means of 'm1' and 'm2'. NGSmethPipe will search within the 'inDir' directory for files with these suffixes running Bowtie in pair-end mode. Two important parameters implemented in Bowtie are the minimum and maximum insert sizes of the pairs, 'I' and 'X' respectively. Example reads can be found in the folder wa09finro_exampleChr21 (extracted from (Li, et al., 2010)).

perl NGSmethPipe.pl seqDir=/home/user/Examples/wa09fibro_exampleChr21/seqDir/ inDir=/home/user/reads/Examples/wa09fibro_exampleChr21/inDir/ bowtieDir=/home/user/bowtie/ m1=_1 m2=_2 I=0 X=500

Alignment Options: Like mentioned in section 3.3, the alignment parameters of NGSmethPipe have been chosen so that the percentage of incorrect alignments is lower than 1%. Nevertheless, if the user needs more coverage (accepting a higher percentage of false mappings and the corresponding consequence on the methylation levels), several alignment parameters can be manipulated: '*l*' sets the Bowtie seed length, 'n' sets the maximum number of mismatches within the seed region and '*m*' sets the maximum number of allowed mismatches in the whole alignment (the seed is extended until m+1 is reached).

perl NGSmethPipe.pl seqDir=/home/user/Examples/h1_exampleChr22/seqDir/ inDir=/home/user/Examples/h1_exampleChr22/inDir/ bowtieDir=/home/user/bowtie/ l=25 n=2 m=4

Quality control: Regarding the quality control, two parameters are of special interest: (i) the minimum Phred Score quality value of a "valid" base call (default:' minQ'=20) and (ii) the maximum number of nonCpG methylcytosines in order to detect bisulfite failure reads (default: ''methNonCpGs'=0.9). The 'methNonCpGs' parameter can take both, values between 0 and 1 and integers. If a value between 0 and 1 is detected, it is interpreted as the fraction between methylated non-CpG contexts and total number of non-CpGs while integers are taken as 'number of methylated non-CpGs'. In the example below, the 'minQ' parameter is set to 40. This means that all positions with less than a Phred Score of 40 (probability of an erroneous base call less than 0.0001) are ignored. Setting the methNonCpGs to 3, the program will discard all reads with more than 3 methylated non-CpG contexts like proposed by (Lister, et al., 2009).

perl NGSmethPipe.pl seqDir=/home/user/Examples/h1_exampleChr22/seqDir/inDir=/home/user/Examples/h1_exampleChr22/inDir/ bowtieDir=/home/user/bowtie/ minQ=40 methNonCpGs=3

Output parameters: Several output parameters do exist. If the user is only interested in a particular sequence context (CpG, CHG, CHH), it can be set with the 'pattern' parameter (for example pattern=CG). Next, the user can choose between reporting the methylation for each cytosine on both strands or merging the methylation levels of the two cytosines that belong to a palindromic sequence context (CpG, CWG). By default the strand merge is not performed, but it can be set with 'uniStrand=Y'. Finally, the output can also be reported in BED or WIG format (see http://genome.ucsc.edu/FAQ/FAQformat.html) for further analysis in a Genome Browser (see section 5). The example below, launches a full analysis for CpGs (pattern=CG), reporting the merged methylation levels (uniStrand=Y) and a BED file (bedOut=Y).

perl NGSmethPipe.pl seqDir=/home/user/Examples/h1_exampleChr22/seqDir/inDir=/home/user/Examples/h1_exampleChr22/inDir/ bowtieDir=/home/user/bowtie/ pattern=CG uniStrand=Y bedOut=Y

4.4.5 Analyzing the output

When a full analysis is launched, each of the scripts will write its own output files.

NGSmethPipeIndex: Bowtie indexes will be stored in the fasta sequence directory, specified by 'seqDir'. The script will create two genomic indexes: in the first one, cytosines are converted to thymines and in the second one guanines to adenines. The indexes files have **.ebwt* extension, which is the output extension used by Bowtie.

NGSmethPipeAlign: The output of this script is stored into the 'outDir' folder, or into the reads directory by default. The alignments are reported in files with **.align* extension. The pair-end mode output will be the same as single-end mode, except a mate identifier at the end of the read id (/1 for the #1 mates and /2 for the #2 mates). The output file has 6 columns:

- ID: original identifier of the read
- Strand: the strand where the read maps (+ or -)
- Chromosome: chromosome where the read maps (chr1, chr2, chrX, etc…)
- Start position: start position of the read in the chromosome (0-based). The coordinate refers to the Watson-strand
- Read: The sequence of the read with its original alphabet, and with an asterisk (*) on the mismatch positions. In case of Crick-strand reads, the sequence returned is the reverse complement of the original sequence
- Quality line: Encoded Phred Quality Scores

Additionally, a log file is written giving the information about the used parameters, running time, and the number of processed and aligned reads.

NGSmethPipeRatios: This script extracts the methylation levels, either for each strand separated or merged together for palindromic sequence contexts. The output files are named after the analyzed pattern (methylation context): for example CG.output. The output file format depends on whether the uniStrand option is set or not.

- uniStrand=N
  - Chromosome: the chromosome
  - Start position: start position of the methylation context (1-based), in the positive strand
  - End position: end position (1-based)
  - Strand: the strand
  - Number of reads: number of reads covering the cytosine position
  - Methylation ratio: methylation level (number of methylcytosines divided by the number of reads mapped to the position)
- uniStrand=Y (by default)
  - Chromosome: the chromosome
  - Start position: start position of the methylation context (1-based), in the positive strand
  - End position: end position (1-based)
  - Number of reads: number of reads covering the context, on both strands
  - Methylation ratio: methylation level (number of methylcytosines divided by the number of reads mapped to the position)
  - Meth difference: the absolute difference between the methylation levels on each strand
  - Number of reads on the Watson strand
  - Methylation level on the Watson strand
  - Number of reads on the Crick strand
  - Methylation level on the Crick strand

The corresponding log file (RatiosCGStats.log) stores the number of reads discarded by the bisulfite check, the number of positions discarded by means of the Q value threshold and the coverage (% of positions covered) as a function of chromosome, strand and context.

5. Storage, browsing and data mining tools for methylation data: NGSmethDB

Powered by the emergence of whole-genome shotgun bisulfite sequencing techniques, many epigenetic projects are currently on the way (Bernstein, et al., 2010). In order to make this data available to all researchers several aims exist: (i) the data should be processed following the same protocol in order to make them comparable among each other, (ii) the results need to be stored in easily accessible databases including data browsing and download, (iii) the databases should implement basic data mining tools in order to compare different data sets retrieving only the relevant information. In order to attend these needs we developed NGSmethDB (Hackenberg, et al., 2011), a database for the storage, browsing and data mining of single-base-pair resolution methylation data which has the following main features:

- Based on the GBrowse (Stein, et al., 2002), the stored methylation data can be easily browsed and analyzed in a genomic context together with other annotations like RefSeq genes, CpG islands, Conserved elements, TFBSs, SNPs, etc
- The user can easily upload methylation data and analyze it together with publically available data in the database

- The raw data can be downloaded for all data sets for different read coverages (1, 5, 10)
- The implemented data mining tools allow for example to retrieve unmethylated cytosines or differentially methylated cytosines in a user defined set of tissues

5.1 Scope

At the end of August 2011, the database holds data for 3 species (human, mouse and *Arabidopis*) and 26 unique tissues. The data are available for 3 different levels of "read coverage": 1, 5 and 10. The read coverage is the number of reads that contribute to the methylation level of a cytosine. Finally, the DB hosts two different sequence contexts, CpG dinucleotids and CHG

5.2 Browsing

A web browser interface is set up by means of the GBrowse genome viewer that is connected through a MySQL backend to NGSmethDB. Features of the browser include the ability to scroll and zoom through arbitrary regions of a genome, to enter a region of the genome by searching for a landmark or performing a full text search of features, as well as the ability to enable and disable feature tracks and change their relative order and appearance. The user can also upload private annotations to view them in the context of the existing ones at the NGSmethDB web site. Apart from the methylation data, other functional annotations are available like CpG islands (Hackenberg, et al., 2006), RefSeq genes (Pruitt, et al., 2007) and Repetitive Elements (RepeatMasker track from UCSC) together with the chromosome sequences and the local G+C content. The methylation information of a given context is represented by the coordinate of the cytosine on the direct strand. To display the methylation values of the cytosines we use a color gradient from white (methylation value = 0, unmethylated in all reads) to red (methylation value = 1, methylated in all reads).

Directly with the NGSmethPipe BED output file (for example CG.bed), the user can upload his methylation results and compare them with other datasets in its genomic context. A brief example is shown in the Figures 4 and 5, for the pair-end reads example (wa09fibro_exampleChr21).

Fig. 4. Uploading BED files.The user can easily upload the bed files output from the NGSmethPipe to the NGSmethDB. In the 'Custom track' tab there are three options to upload tracks, selects the 'From a file' option and choose for example the CG.bed file from the 'outDir' or 'inDir' by default. For more information, see the help link.

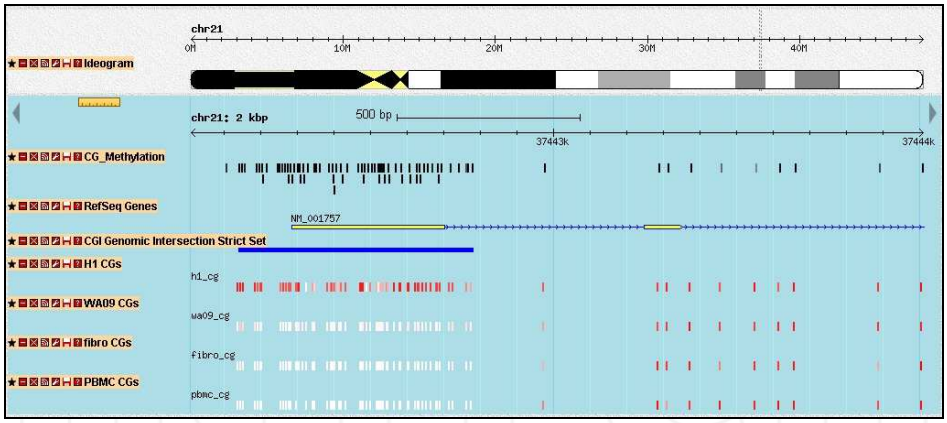


Fig. 5. Uploaded track example.The figure shows the pair-end wa09fibro example in grey scale: from black (methylated CpGs) to white (unmethylated CpGs). A region from the human chromosome 21 is depicted, where the CpG island overlapping the NM_001757 gene promoter is methylated in the uploaded track and H1 track, while it remains unmethylated in the rest of the displayed tissues. Note that the tracks from the database are shown in a different format (from red (methylated) to white (unmethylated)) and therefore the uploaded track can be easily distinguished. Once the track has been uploaded, the user can modify its features with the configure button at the left side of the track.

5.3 Data mining

Currently, the data can be accessed in 5 different ways. (i) Dump download, (ii) Retrieve unmethylated contexts, (iii) Retrieve differentially methylated contexts, (iv) Get methylation states of promoter regions and (v) Retrieve methylation data for chromosome region.

The 'Dump download' option shows first an overview of current database content, including a short description of the tissue, the genome coverage in %, a link to PubMed, and raw data files for #read ≥ 1, #read ≥ 5 and #read ≥ 10 coverage. The files show the chromosome, chromosome-start and chromosome-end coordinates, the sequence methylation context (either CpG or CWG), the number of reads and the cytosine methylation ratio.

The 'Retrieve unmethylated contexts' tool can be used to retrieve all unmethylated cytosines in a given set of tissues. The user has to select the sequence context (CG or CWG), the read coverage, the threshold for unmethylation (often a threshold of 0.2 is used, i.e. all cytosines with values ≤0.2 are considered to be unmethylated) and the tissues. The tool will detect all cytosine contexts showing lower methylation ratios than the chosen threshold in all selected tissues. The provided output file holds the chromosome, chromosome start- and end-coordinates and the methylation values in all selected tissues. Note that this tool can be also used to retrieve all CpGs which are present in every single analyzed tissue by setting the threshold to one. In doing so, cytosines with methylation data in all tissues will be reported regardless of its methylation state, i.e. cytosines that are not covered by at least the number of chosen coverage threshold (1, 5 or 10) in any of the analyzed tissues will not be reported in the output.

By means of the 'Retrieve differentially methylated contexts' tool all differentially methylated cytosine contexts can be determined in a given set of tissues. All parameters of the 'Retrieve unmethylated contexts' (see above) are available here, plus one additional parameter: the threshold for the methylation value which defines whether a cytosine is considered to be methylated (often a threshold of 0.8 is used, i.e. all cytosines with higher values than ≥0.8 are considered to be methylated). We define a cytosine as differentially methylated if it is unmethylated in at least one tissue and methylated in at least one other tissue. The tool reports those differentially methylated cytosine contexts that are either methylated or unmethylated in all analyzed tissues, i.e. those contexts that show intermediate methylation in only one tissue will not be reported.

Another tool allows depicting the methylation states of all cytosine contexts within the promoter region of RefSeq genes. The promoter region is defined from 1.5 kb upstream of the Transcription Start Site (TSS) to 500 bp downstream of the TSS. The output is displayed by default as an overview table that summarizes the fluctuation along the promoter as well as over the different tissues.

Finally, all methylation values for a selected set of tissues can be retrieved for a given chromosomal region, once the user provides the start and end chromosome coordinates.

6. Conclusions and outlook

The price for high-throughput DNA sequencing is dropping constantly and consequently the number of available whole genome shotgun bisulfite sequencing data is increasing at a very high rate. This implies a strong need for bioinformatics applications able to deal with this vast amount of data, converting them into high quality, single-base pair resolution methylation maps. In this chapter we review the most important bioinformatics tools that are currently available comparing them by several aspects including both, technical and biological issues. The quality control is still rather disregarded by many tools and currently NGSmethPipe is the program that implements the highest number of quality related features. This application is based on Bowtie with optimized alignment parameters and a seed extension method. It implements several quality control features like the detection of SNVs, the deletion of bisulfite failure reads, and the consideration of the base call qualities in the methylation profiling step. Furthermore, NGSmethPipe delivers output files that can be uploaded directly into the Genome Browser of the NGSmethDB database, thus allowing the user to analyze the custom data within the context of other tissues or genomic elements. Future directions will be to populate the NGSmethDB database by means of the NGSmethPipe application adding also other relevant data like histone methylation, expression or disease data.

7. Acknowledgment

This work was supported by the Ministry of Innovation and Science of the Spanish Government [BIO2010-20219 (M.H.), BIO2008-01353 (J.L.O.)]; 'Juan de la Cierva' grant (to M.H.) and Basque Country 'Programa de formación de investigadores' grant (to G.B.).

8. References

Barturen, G., *et al.* (Submitted) NGSmethPipe: A tool to generate high-quality methylation maps.

Bernstein, B.E., *et al.* (2010) The NIH Roadmap Epigenomics Mapping Consortium, *Nature biotechnology*, 28, 1045-1048.

Bibikova, M. and Fan, J.B. (2009) GoldenGate assay for DNA methylation profiling, *Methods in molecular biology (Clifton, N.J*, 507, 149-163.

Bock, C., *et al.* (2011) Reference Maps of human ES and iPS cell variation enable high-throughput characterization of pluripotent cell lines, *Cell*, 144, 439-452.

Bock, C., *et al.* (2010) Quantitative comparison of genome-wide DNA methylation mapping technologies, *Nature biotechnology*, 28, 1106-1114.

Cokus, S.J., *et al.* (2008) Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning, *Nature*, 452, 215-219.

Costello, J.F., *et al.* (2000) Aberrant CpG-island methylation has non-random and tumour-type-specific patterns, *Nature genetics*, 24, 132-138.

Chen, P.Y., Cokus, S.J. and Pellegrini, M. (2010) BS Seeker: precise mapping for bisulfite sequencing, *BMC Bioinformatics*, 11, 203.

Chen, T. and Li, E. (2004) Structure and function of eukaryotic DNA methyltransferases, *Curr Top Dev Biol*, 60, 55-89.

Dodge, J.E., *et al.* (2005) Inactivation of Dnmt3b in mouse embryonic fibroblasts results in DNA hypomethylation, chromosomal instability, and spontaneous immortalization, *J Biol Chem*, 280, 17986-17991.

Estecio, M.R., *et al.* (2007) High-throughput methylation profiling by MCA coupled to CpG island microarray, *Genome Res*, 17, 1529-1536.

Gu, H., *et al.* (2010) Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution, *Nature methods*, 7, 133-136.

Hackenberg, M., Barturen, G. and Oliver, J.L. (2011) NGSmethDB: a database for next-generation sequencing single-cytosine-resolution DNA methylation data, *Nucleic acids research*, 39, D75-79.

Hackenberg, M., *et al.* (2006) CpGcluster: A distance-based algorithm for CpG-island detection, *BMC Bioinformatics*, 7, 446.

Hackenberg, M., Rodriguez-Ezpeleta, N. and Aransay, A.M. (2011) miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments, *Nucleic acids research*, 39, W132-138.

Hach, F., *et al.* (2010) mrsFAST: a cache-oblivious algorithm for short-read mapping, *Nature methods*, 7, 576-577.

Harris, E.Y., *et al.* (2010) BRAT: bisulfite-treated reads analysis tool, *Bioinformatics (Oxford, England)*, 26, 572-573.

Harris, R.A., *et al.* (2010) Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications, *Nature biotechnology*, 28, 1097-1105.

Hellman, A. and Chess, A. (2007) Gene body-specific methylation on the active X chromosome, *Science*, 315, 1141-1143.

Horner, D.S., *et al.* (2010) Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing, *Briefings in bioinformatics*, 11, 181-197.

Jones, P.A. (1999) The DNA methylation paradox, *Trends Genet*, 15, 34-37.

Karpf, A.R. and Matsui, S. (2005) Genetic disruption of cytosine DNA methyltransferase enzymes induces chromosomal instability in human cancer cells, *Cancer research*, 65, 8635-8639.

Krueger, F. and Andrews, S.R. (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications, *Bioinformatics (Oxford, England)*, 27, 1571-1572.

Laird, P.W. (2010) Principles and challenges of genomewide DNA methylation analysis, *Nat Rev Genet*, 11, 191-203.

Langmead, B., *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, *Genome biology*, 10, R25.

Laurent, L., *et al.* (2010) Dynamic changes in the human methylome during differentiation, *Genome Res*, 20, 320-331.

Li, Y., *et al.* (2010) The DNA methylome of human peripheral blood mononuclear cells, *PLoS Biol*, 8, e1000533.

Lister, R. and Ecker, J.R. (2009) Finding the fifth base: genome-wide sequencing of cytosine methylation, *Genome Res*, 19, 959-966.

Lister, R., *et al.* (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis, *Cell*, 133, 523-536.

Lister, R., *et al.* (2009) Human DNA methylomes at base resolution show widespread epigenomic differences, *Nature*, 462, 315-322.

Meissner, A., *et al.* (2005) Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis, *Nucleic acids research*, 33, 5868-5877.

Miura, A., *et al.* (2009) An Arabidopsis jmjC domain protein protects transcribed genes from DNA methylation at CHG sites, *EMBO J*, 28, 1078-1086.

Okano, M., *et al.* (1999) DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development, *Cell*, 99, 247-257.

Ondov, B.D., *et al.* (2010) An alignment algorithm for bisulfite sequencing using the Applied Biosystems SOLiD System, *Bioinformatics (Oxford, England)*, 26, 1901-1902.

Pedersen, B., *et al.* (2011) MethylCoder: software pipeline for bisulfite-treated sequences, *Bioinformatics (Oxford, England)*, 27, 2435-2436.

Pellegrini, M., *et al.* (2010) Conservation and divergence of methylation patterning in plants and animals, *Proceedings of the National Academy of Sciences of the United States of America*, 107, 8689-8694.

Pruitt, K.D., Tatusova, T. and Maglott, D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins, *Nucleic acids research*, 35, D61-65.

Robinson, M.D., *et al.* (2010) Protocol matters: which methylome are you actually studying?, *Epigenomics* 2, 587-598.

Shendure, J. and Ji, H. (2008) Next-generation DNA sequencing, *Nature biotechnology*, 26, 1135-1145.

Smith, A.D., *et al.* (2009) Updates to the RMAP short-read mapping software, *Bioinformatics (Oxford, England)*, 25, 2841-2842.

Stein, L.D., *et al.* (2002) The generic genome browser: a building block for a model organism system database, *Genome Res*, 12, 1599-1610.

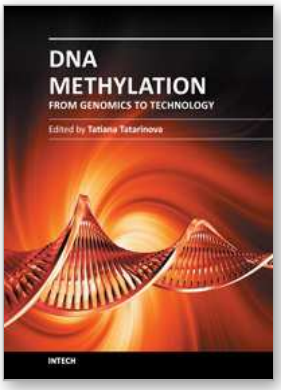
Tomso, D.J. and Bell, D.A. (2003) Sequence context at human single nucleotide polymorphisms: overrepresentation of CpG dinucleotide at polymorphic sites and suppression of variation in CpG islands, *Journal of molecular biology*, 327, 303-308.

Weisenberger, D.J., *et al.* (2005) Analysis of repetitive element DNA methylation by MethyLight, *Nucleic acids research*, 33, 6823-6836.

Xi, Y. and Li, W. (2009) BSMAP: whole genome bisulfite sequence MAPping program, *BMC Bioinformatics*, 10, 232.

Xin, Y., Ge, Y. and Haghighi, F.G. (2011) Methyl-Analyzer--whole genome DNA methylation profiling, *Bioinformatics (Oxford, England)*, 27, 2296-2297.

Zhang, X., *et al*. (2006) Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis, *Cell*, 126, 1189-1201.



DNA Methylation - From Genomics to Technology

Edited by Dr. Tatiana Tatarinova

ISBN 978-953-51-0320-2
Hard cover, 400 pages
Publisher InTech
Published online 16, March, 2012
Published in print edition March, 2012

Epigenetics is one of the most exciting and rapidly developing areas of modern genetics with applications in many disciplines from medicine to agriculture. The most common form of epigenetic modification is DNA methylation, which plays a key role in fundamental developmental processes such as embryogenesis and also in the response of organisms to a wide range of environmental stimuli. Indeed, epigenetics is increasing regarded as one of the major mechanisms used by animals and plants to modulate their genome and its expression to adapt to a wide range of environmental factors. This book brings together a group of experts at the cutting edge of research into DNA methylation and highlights recent advances in methodology and knowledge of underlying mechanisms of this most important of genetic processes. The reader will gain an understanding of the impact, significance and recent advances within the field of epigenetics with a focus on DNA methylation.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Michael Hackenberg, Guillermo Barturen and José L. Oliver (2012). DNA Methylation Profiling from High-Throughput Sequencing Data, DNA Methylation - From Genomics to Technology, Dr. Tatiana Tatarinova (Ed.), ISBN: 978-953-51-0320-2, InTech, Available from: http://www.intechopen.com/books/dna-methylation-from-genomics-to-technology/dna-methylation-profiling-from-high-throughput-sequencing-data

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821