Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Sequence Analysis

gemBS - high throughput processing for DNA methylation data from Bisulfite Sequencing

Angelika Merkel [1,2*], Marcos Fernández-Callejo [1,2], Eloi Casals [1,2], Santiago Marco-Sola [3], Ronald Schuyler [4], Ivo G. Gut [1,2] and Simon C. Heath [1,2]

[1]Centro Nacional de Análisis Genómico (CNAG-CRG), Centre de Regulacio Genómico (CRG), 08028 Barcelona, Spain; [2]Universitat Pompeu Fabra (UPF), 08002 Barcelona, Spain; [3]Universitat Autònoma de Barcelona, Bellaterra 08193, Spain and [4]Department of Immunology and Microbiology, University of Colorado, Aurora, Colorado 80045, USA.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: DNA methylation is essential for normal embryogenesis and development in mammals and can be captured at single base pair resolution by whole genome bisulfite sequencing (WGBS). Current available analysis tools are becoming rapidly outdated as they lack sensible functionality and efficiency to handle large amounts of data now commonly created.

Results: We developed gemBS, a fast high-throughput bioinformatics pipeline specifically designed for large scale BS-Seq analysis that combines a high performance BS-mapper (GEM3) and a variant caller specifically for BS-Seq data (BScall). gemBS provides genotype information and methylation estimates for all genomic cytosines in different contexts (CpG and non-CpG) and a set of quality reports for comprehensive and reproducible analysis. gemBS is highly modular and can be easily automated, while producing robust and accurate results.

Availability: gemBS is released under the GNU GPLv3+ license. Source code and documentation are freely available from www.statgen.cat/gemBS.

Contact: angelika.merkel@cnag.crg.es

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Whole genome sequencing of bisulfite converted DNA (WGBS) is considered the current gold standard for studying DNA methylation at base pair resolution, but due to high costs its application has been lagging behind other more cost-efficient platforms (such as microarrays). Recent decreases in sequencing costs have led to a rise in genomic BS-sequencing applications and the creation of significant amounts of large data sets, frequently as part of large epigenomic consortia such as BLUEPRINT, ENCODE or NIH ROADMAP. However, most available tools (see Krueger et al. (2012) for a review) lack the functionality to process diverse BS-Seq dataset sensibly and do not scale efficiently for high-throughput analysis.

DNA methylation analysis from bisulfite treated DNA poses particular challenges (Laird, 2010; Bock, 2012). Firstly, bisulfite treatment initially converts un-methylated cytosines to uracils, which are replaced by thymines during PCR amplification. This results in four (potentially different) sequences that need to be aligned to a reference genome. Although, powerful tools for short read alignment exists, the increased rate of mismatches and low sequence complexity of bisulfite converted sequences prevent a straightforward implementation. Secondly, cytosine methylation status is derived from read counts of converted and non-converted cytosines. Genetic variants, low base call qualities and conversion failures can lead to misinterpretation of the observed counts (i.e. methylation levels) and have to be accounted for.

Here we describe gemBS, an efficient and scalable pipeline for high performance BS-mapping and accurate variant and methylation calling.

Table 1: Overview of requirements and features of gemBS compared with other popular BS-Seq analysis tools.

	GEMBS	BISMARK	BSMAP	BWA-meth	Novoalign	Bis-SNP	MethylExtract
General							
Memory required (RAM)	48GB	16GB	8-16GB, 26GB	8-16GB	16GB	10GB	?
Multi-threading	yes	yes	yes	yes	yes	yes	yes
Language	C, Python	Perl	C++	Python	?	Java, Perl	Perl
Distribution	GitHub (GNU GPL v3), Singularity, Docker	Github (GNU GPL v3)	Github (GNU GPL v3)	Github (MIT license)	Novocraft (free trial)	SourceForge (MIT license)	Github (GNU GPL v3)
Supported data types							
RRBS	yes	yes	yes	yes	yes	yes	yes
WGBS	yes	yes	yes	yes	yes	yes	yes
PBAT	yes	yes	-	-	-	?	?
NOMe-seq	-	-	-	-	-	yes	yes
Directional/non-directional library	yes/ yes	yes/ yes	yes/ yes	yes/ -	yes/ yes	yes/ yes	yes/ yes
Single end/ paired end	yes/ yes	yes/ yes	yes/ yes	yes/ yes	yes/ yes	yes/ yes	yes/ yes
Base space/ colour space	yes/no	yes/ no	yes/yes	?	yes/ yes	yes/no	yes/no
Functionality							
BS-alignment	GEM3	bowtie2	BSMAP (SOAP)	bwa mem	Novoalign (SOAP)	*(Bwa-meth, BSMAP, Novoalign)*	*(Bismark)*
BS- and nonBS alignment combined	GEM3	-	-	-	-	-	-
Spike-in support	yes	yes	yes	yes	yes	-	-
UMI support	-	-	-	-	-	-	-
Adapter trimming	-	TrimGalore	yes	*(TrimGalore)*	yes	-	-
5'/3' end trimming (M-bias)	BScall	TrimGalore	yes	*(TrimGalore)*	yes	yes (5' end, 3'end)	yes
Remove duplicates	BScall		yes	-	yes	*Picard-Tools*	yes, (Picard-Tools)
Collapse overlapping PE reads	BScall		yes	-	yes	yes	yes
Variant calling	BScall	-	-	*(biscuit)*	-	yes (GATK)	yes
Methylation calling	BScall	MethylationExtractor	methratio.py	*(methylDackel)*	NovoMethyl	yes	yes
Input/ output							
FASTQ/FASTA	yes	yes	yes	yes	yes	-	-
Standard input	yes	-	-	-	-	-	-
Alignments in TXT/BAM	yes/no	yes/ yes	yes/ yes	yes/ yes	yes/ yes	-	-
BAM/SAM sorted by readID	yes/yes	-	yes	-	yes	yes	yes
Genotype and methylation calls	yes (bcf)	-	-	-	-	yes (vcf)	yes (vcf)
CpG and nonCpG both strands /strand specific	yes/ yes (txt,bed)	?/ yes	yes (txt)	-	yes/ yes (txt)	yes/ yes (bed)	yes/ yes (txt)
Visualization methylation and coverage	yes (bigWig,bedGraph)	yes (bedgraph)	-	-	-	yes (.wig, bedgraph)	Yes (.wig, bigWig)
Summary reports	yes (json, html)	yes (html,txt)	-	*(methylDackel)*	yes(txt)	yes	yes

BISMARK, BSMAP, BWA-meth and Novoalign are predominantly used for aligning BS data and methylation calling. Bis-SNP and MethylExtract perform genotype and methylation calling from already aligned data. (Italics indicate third party applications recommended by the authors of the software for extended analysis).

It can be used to analyze DNA methylation in CpG and non-CpG context and allows for variant calling from BS-Seq data. We compare gemBS against some commonly used tools and test its performance on a variety of BS-Seq data types. Additionally, we demonstrate how gemBS can accurately call SNPs and how this is influenced by sequencing coverage.

2 Methods

gemBS is a versatile pipeline that allows for fast and reproducible analysis while providing up-to-date functionality (see Table 1 for a comparison of features between gemBS and other popular tools).
gemBS is capable of analyzing data derived from diverse protocols, such as WGBS, RRBS and PBAT, for single and paired-end sequencing and directional/non-directional libraries, while providing support for spike-in sequences. If available, gemBS can additionally process genomic data, for example to increase power during genotype calling (see below). Outputs are provided in standard data formats for seamless downstream analysis and summary reports allow for quality checks (see Figure 1). gemBS' two core components are GEM3, a high performance short read aligner and BScall, a methylation aware variant caller. Both are embedded in a highly efficient framework together with standard sequence analysis tools (samtools, bcftools) (http://samtools.sourceforge.net/) and a small database (sqlite3) that tracks the failure/completion of individual gemBS tasks.

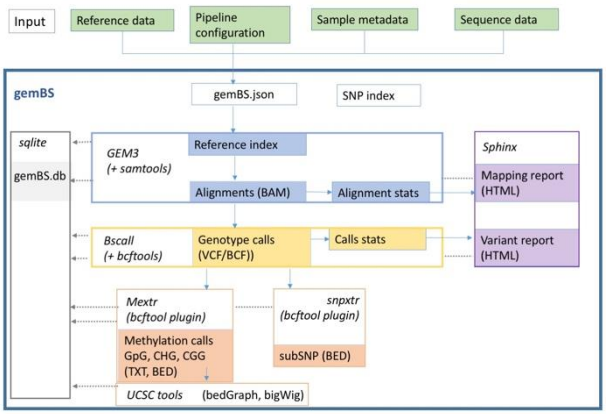


Fig. 1. gemBS workflow and components. (Software in italics)

GEM3 similarly to other '3 letter' aligners, such as Bwa-meth (Pedersen *et al.*, 2014), Novoalign (http://www.novocraft.com/) or Bismark (Krueger and Andrews, 2011), uses an *in silico* conversion approach for mapping. Bisulfite treatment and PCR create four types of reads: Crick+, Crick-, Watson+, Watson-. Crick+ and Watson- correspond to the original strands after bisulfite conversion and are C-depleted, while Crick- and Watson+ are the respective complementary strands and are G-depleted.
In directional sequencing, the first read in paired end sequencing (and the only read in non-paired sequencing) stems from either the Crick+ or Watson- strands and is therefore C-depleted. The second read for paired end sequencing is derived from the complementary strand (Crick- or Watson+) respectively and is G-depleted. For non-directional sequencing, the first read (or only read) can be from any of the four strands and forms a mixture of reads that are C-depleted and G-depleted (with the second read, if present, being from the strand complementary to the first read).

For directional sequencing, GEM3 converts all remaining (non-converted) C's in the first or only read to T's, and all remaining G's in the second read (if present) to A's prior to mapping. For non-directional non-paired sequencing, GEM3 uses the proportion of C's and G's in the reads to assess whether the read is C or G depleted and performs the conversion step accordingly. For non-directional paired-end sequencing, GEM3 uses the base proportions in both reads to assess whether the read pair is C depleted for read 1 / G depleted for read 2 or G depleted for read 1 / C depleted for read 2, and performs the conversion accordingly.
An alternative mapping approach *to in silico* conversion is used by so called '4 letter aligners', such BSMAP (Xi and Li, 2009). Reads are simply mapped against the original reference allowing for either a C or T as match for any potential cytosine.

After successful alignment, BScall performs genotype and methylation calling from the mappings (produced by GEM3 or an alternative bisulfite aware mapper). Variant calls can be derived from matched genome sequencing data, SNP arrays or public databases such as dbSNP (www.ncbi.nlm.nih.gov/projects/SNP/index.html), but it is more sensible (and cost efficient) to call variants directly from BS-Seq data (Barturen *et al.*, 2013; Liu *et al.*, 2012).
BScall uses Bayesian inference to jointly infer the most likely genotype and methylation levels while taking into account sequence quality, bisulfite under- and over-conversion and the observed bases similar to Bis-SNP (Liu *et al.*, 2012) (see Supplement for a detailed description of the model). The program reports all potential cytosines and non-reference homozygous calls together with the conditional methylation estimates (= unconverted and converted bases). Conversion rates can be either set as fixed values (based on previous experience) or estimated from reads mapping to control sequences.
Prior to the calling step, depending on the sequencing protocol (defined in the parameters set by the user), BScall removes duplicated reads, collapses overlapping paired-end reads and trims read-end cytosines that are known be to artificially generated during library preparation. If external processing tools have been used to manipulate the read data previously (e.g. TrimGalore for trimming (www.bioinformatics.babraham.ac.uk/projects/trim_galore/) or Picard tools for marking duplicates (https://broadinstitute.github.io/picard/)), BScall will honor the set flags and process the alignments accordingly.
Once the genotypes are called, they are further processed by two utilities 'mextr' and 'snpxtr' (see Figure 1).'mextr' extracts all homozygous cytosine methylation in different contexts (CpG, CHG, CHH) and in a variety of formats (txt, bed, bedGraph, bigWig). 'snpxtr' extract a set of variants for a given list of genomic positions.

3 Results

3.1 Performance of GEM3 and BScall compared to other analysis tools

We compared GEM3 against the above mentioned aligners and found that GEM3 is 4 and 72 times faster than BSMAP and Novoalign but similarly sensitive in aligning 85% of all sequenced bases with high quality (97% prior to filtering and trimming) (see Table 2, Figure 2A, Table S2). Bwa-meth and Bismark, in comparison, are both slower and slightly less sensitive.
GEM3 achieves its fast processing mostly thanks to its efficient algorithms (Marco-Sola *et al.*, 2012; Marco-Sola, 2017), but also by avoiding unnecessary processing steps and the generation of intermediate files.

Table 2. Processing times (CPU hours) for BS mapping and calling tools

	Mapper				Caller*		
Depth	Bwa-meth	Novo align	Bismark	GEM3	Bis-SNP	Methyl Extract	BScall
27x	278.7	2419.6	354.3	33.5	266.8	204.9	6.2
58x	612.5	5050.9	732.6	69.2	419.5	372.5	8

*Calls performed on GEM3 alignments

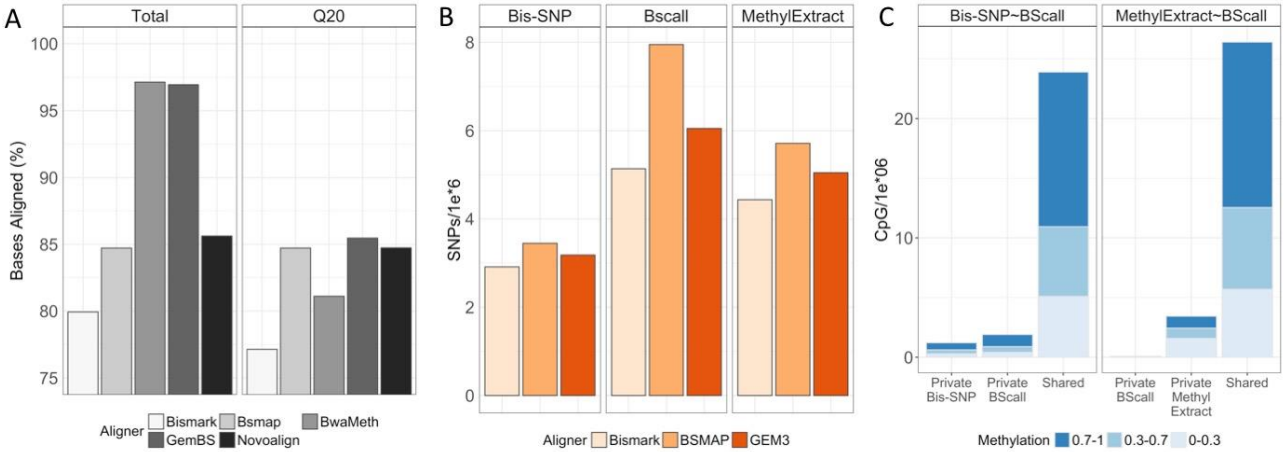


Fig 2. gemBS performance compared with other BS-Seq analysis tools (58X coverage). A) Mapping sensitivity across BS aligners (Q20 = alignments with mapping quality > 20). B) SNPs calls by caller and mapper C) Shared CpG calls and methylation estimates across callers (GEM3 alignment)

For example, all of the conversion steps before and after mapping are performed on the fly on a read-pair-by-read-pair basis in the mapper itself. GEM3 only performs one alignment against a single composite reference since its internal design allows the handling large indices. Because of its large index, GEM3 requires comparatively large memory resources (e.g. 48GB RAM to process the human genome), which nowadays however is readily available on most midrange workstations. As with other mappers, the index needs to be generated only once prior to mapping and is reused for multiple mappings when mapping to the same reference (see Figure 1).

When we compare BScall against Bis-SNP (Liu *et al.*, 2012) and MethylExtract (Barturen *et al.*, 2013), two other bisulfite variant callers (see Table 1), we find that BScall is more than 50 times faster than any of them (Bis-SNP and MethylExtract take approximately the same time for processing). In contrast to non-bisulfite variant calls that focus on the identification of variable sites, the objective of BScall is to confidently identify homozygous cytosines that can be used for downstream methylation analyses. Despite this difference, the accuracy and sensitivity of BScall for SNP variants is comparable to non-bisulfite variant callers such as FreeBayes (Garrison and Marth 2012). Amongst the 3 callers, MethylExtract reports most SNPs, followed by BScall and Bis-SNP see Figure 2B). Between BScall and Bis-SNP SNP calls are more similar (71% of SNP shared) than MethylExtract (17% and 20% of SNP shared) likely because both tools implement a similar model for genotype calling (Liu *et al.*, 2012). In fact, most SNPs identified by Bis-SNP are also identified by BScall but do not pass the filtering thresholds. Out of the SNP identified by Bis-SNP, 99.9% are also found by BScall when relaxing all filters.

At the CpG level, about 85% of position are shared across amongst the callers and show a similar distribution of methylation levels (see Figure 2). Those CpG unique to either Bis-SNP or BScall show similar methylation levels as expected, but those unique to MethylExtract have a large proportion of unmethylated CpG. Together, with the observation that MethylExtract calls around 30 million CpG (gemBS ~26 million and Bis-SNP ~25 million CpGs, see Table S4) while there are only 28 million present in the human genome, a significant amount of these are likely false genotype calls.

We also noticed that the alignments (i.e. the mapper used to produce them) influence the calling results. All callers produced more cytosine calls based on alignments from BSMAP which were however are less concordant than calls produced from GEM3 or Bismark alignments (see Table S3).

3.2 Validation

In order to further evaluate the robustness of gemBS to cope with different BS-Seq protocols, namely WGBS, RRBS and PBAT, we tested gemBS on a variety of publicly available datasets (see Figure 3, Table S7). Different BS-sequencing protocols exhibit different biases due to BS-induced DNA degradation and PCR amplification that can have an impact on genomic coverage distribution (Laird, 2010; Olova *et al.*, 2017) and read alignment itself can be affected by read length, single/paired-end sequencing and dimerization in paired-end. Rates of unique mappings were highest for WGBS data from BLUEPRINT (~84%), ENCODE (average=~69%) and Lister et al. (2013)(~69%),

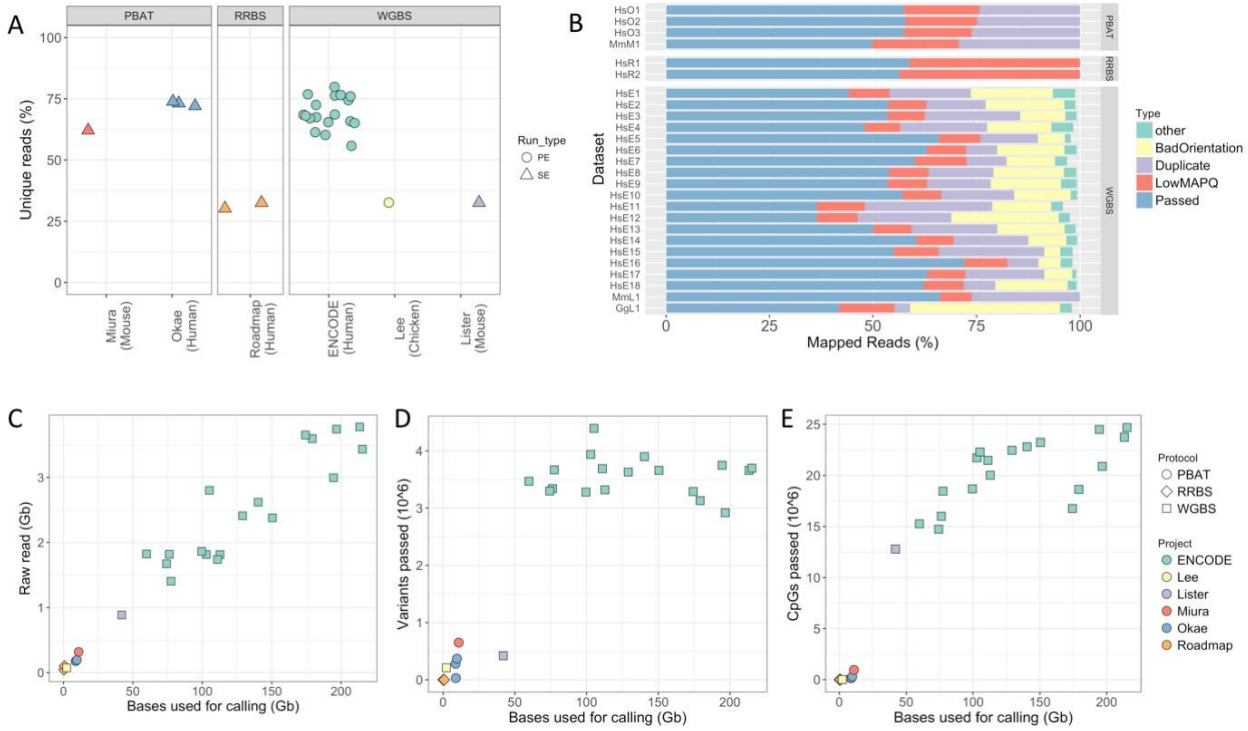


Fig 3. gemBS performance for different BS-Seq protocols (PBAT, RRBS, WGBS). A) Unique reads mapped, B) Types of reads mapping to conventional chromosomes, C-E) Relationship between bases used for calling and sequence input, variants called after filters and CpGs passing filters. (Datasets from publically available sources for 3 different species (Project: ENCODE (HsE), Roadmap (HsR); Studies: Okae et al. (2017) (HsO), Lister et al. (2012) (MmL), Miura et al. (2012)(Mml), Lee et al. (2015) (GgL))).

followed by PBAT data from Miura et al. (2012) and Okae and Arima (2016) with ~62% and an average of 73% unique mappings, respectively. RRBS data from the Roadmap project yielded particular low rates (32%), due to mapping issues for very short read lengths (29bp) (see Figure 3A). Filtering the mapped reads additionally for wrong orientation of read pair, insert size and duplicates (except RRBS) reduces the number of reads used for genotype calling depending on the dataset to 36-72% (Figure 3B). Although, there are clear differences in the types of artefacts for each data type and their proportion may vary substantially, the number of bases left after filtering is still highly correlated with the initial sequencing depth. Variant calling is most efficient at high coverages (>50Gb post-filtered bases) as is detecting CpGs (Figure 3D-E).

SNP calls from BS-Seq data are mainly used to determine cytosine context correctly (CpG and non-CpG). Nevertheless, they can also be used to identify allele specific methylation or as a quality check to identify sample mix-ups. To assess the accuracy of SNP calls from gemBS (BScall), we used a sample from the 1000 genomes dataset (NA12878) that has been extensively genotyped by multiple approaches, and for which high quality public sequencing datasets are available. We first analyzed the original dataset using both gemBS and FreeBayes, and then performed *in silico* bisulfite conversion on the same data set (supplement for details). The >12 million SNPs that form the phase 3 genotype calls
served as gold standard for comparing the output. To assess the impact of sequencing depth, we also subsampled the original data taking at random 50%, 25% and 12.5% of the read pairs. An *in silico* bisulfite-converted dataset was generated from each of the sub-sampled datasets, and both the genomic and bisulfite-converted datasets were analyzed using gemBS.
Using the original dataset, gemBS produced confident genotype calls (passing standard filtering) on 96.92% of the SNP panel, with a discordancy rate when compared to the 1000 genomes gold standard of 0.30% (see Table 3). FreeBayes had similar figures after filtering out sites with genotype quality (GQ) < 20 (matching the filtering used for gemBS), with a call rate of 95.55% and a discordancy rate of 0.38%. With the *in silico* bisulfite converted dataset the call rate dropped to 94.74%, while the discordancy rate stayed low at 0.32%. Comparing the gemBS genotype calls from the original and bisulfite converted datasets gave a discordancy percentage of 0.03%, while comparing the calls from FreeBayes (using the original data) and gemBS gave discordancy rates of 0.03% and 0.04% respectively for the original and bisulfite converted datasets. The discrepancy between SNPs called from genomic data and simulated bisulfite data increased with decreasing sequencing depth as in practice only half of the reads from bisulfite data can be used for genotype calling. The discordancy rate with the phase3 gold standard SNP calls, however, remained at similar levels.
We conclude that gemBS can accurately call SNPs from genomic and BS data, and as a key feature of gemBS proves the utility of gemBS above other currently available analysis tools.

Table 3. Number of SNP calls by Freebayes and BScall compared to SNP calls from 1000 Genome Project

Sample1	Sample2	Sample_1 typed	Sample_2 typed	Both typed	Concordant	% typed 1	% typed 2	% discordant
1000g	Freebayes	12513267	11955900	11955900	11910953	99,989%	95,535%	0,376%
1000g	BScall 100	12513267	12129210	12128302	12092106	99,989%	96,920%	0,298%
1000g	BScall 100 BS	12513267	11856511	11855613	11818244	99,989%	94,741%	0,315%
1000g	BScall 50	12513267	12037064	12036093	11997425	99,989%	96,184%	0,321%
1000g	BScall 50 BS	12513267	11418888	11417945	11381715	99,989%	91,244%	0,317%
1000g	BScall 25	12513267	11499437	11498405	11461191	99,989%	91,888%	0,324%
1000g	BScall 25 BS	12513267	8707530	8706711	8677565	99,989%	69,579%	0,335%
1000g	BScall 12.5	12513267	7852793	7851963	7823090	99,989%	62,749%	0,368%
1000g	BScall12.5 BS	12513267	3186102	3185739	3168250	99,989%	25,459%	0,549%

BS= simulated bisulfite data, concordant = SNPs called in both data sets identically, discordant= SNP called with different genotype

4 Conclusion

Here we present gemBS, a bioinformatics pipeline for BS mapping and genotype and methylation calling. gemBS is fast, efficient and easy scalable for large data sets. Due to its flexible architecture it allows for modular processing and the incorporation of third party applications. It can be run on a single workstation, a computer cluster with shared file system or a distributed system without shared file systems. Contemporary features include the distribution/installation through Singularity or as a Docker image.

gemBS produces accurate results and robustly handles different data types while exceeding current popular tools in performance and functionality. gemBS has been the BS analysis pipeline for the BLUEPRINT project and has been recently adopted as the standard processing pipeline for bisulfite sequencing within the IHEC consortium. We therefore believe gemBS has the potential to become the standard analysis tool for BS-Seq analysis.

Funding

This work was funded by the European Commission (FP7-HEALTH: READNA, BLUEPRINT; FP7-INFRASTRUCTURES: ESGI) and the Spanish MINECO (BIO2015-71969-REDI).

Conflict of Interest: none declared.

References

Barturen,G. *et al.* (2013) MethylExtract: High-Quality methylation maps and SNV calling from whole genome bisulfite sequencing data. *F1000Research*, **2**, 1–23.

Bock,C. (2012) Analysing and interpreting DNA methylation data. *Nat. Rev. Genet.*, **13**, 705–19.

Krueger,F. *et al.* (2012) DNA methylome analysis using short bisulfite sequencing data. *Nat. Methods*, **9**, 145–51.

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

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

Lister,R. *et al.* (2013) Global epigenomic reconfiguration during mammalian brain development. *Science*, **341**, 1237905.

Liu,Y. *et al.* (2012) Bis-SNP: combined DNA methylation and SNP calling for Bisulfite-seq data. *Genome Biol.*, **13**, R61.

Marco-Sola,S. (2017) Efficient Approximate String Matching Techniques for Sequence Alignment Advisor Paolo Ribeca.

Marco-Sola,S. *et al.* (2012) The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat. Methods*, **9**, 1185–1188.

Miura,F. *et al.* (2012) Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. *Nucleic Acids Res.*, **40**.

Okae,H. and Arima,T. (2016) DNA Methylation Dynamics During Early Human Development. *J. Mamm. Ova Res.*, **33**, 101–107.

Olova,N. *et al.* (2017) Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. *bioRxiv*, 165449.

Pedersen,B.S. *et al.* (2014) Fast and accurate alignment of long bisulfite-seq reads. *arXiv Prepr. arXiv …*, **00**, 1–2.

Xi,Y. and Li,W. (2009) BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics*, **10**, 1–9.

DETAILS OF GENOTYPE AND METHYLATION MODEL

1. Outline of approach

At a given genomic position, the bases covering that position are recorded. Considering only single base polymorphisms there are 10 possible genotype at site; we calculate the probability of each genotype given the observed bases, taking into account the sequencing error rate, the methylation rate and the estimated bisulfite conversion rates. Given the calculated genotype probabilities, we can select the most likely genotype and calculate the support for the call. For the 7 genotypes that contain at least one C or G allele (AC, AG, CC, CG, CT, GG, GT), we calculate the maximum likelihood estimate of the methylation conditional on each genotype in turn, and this methylation estimate is then used to calculate the genotype probabilities.

2. Model

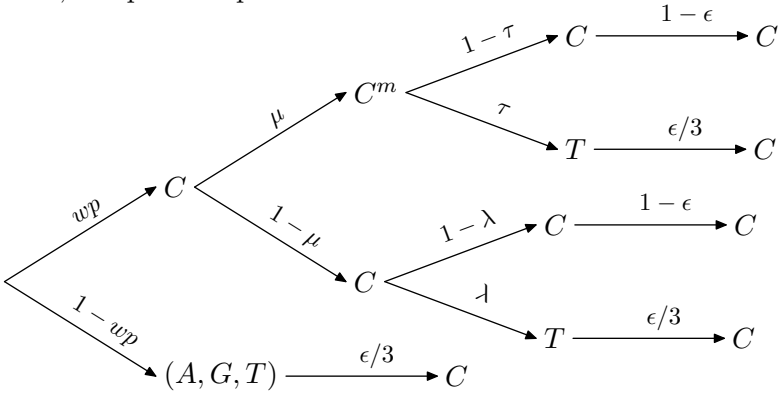
The observed bases can be split into two groups, those that have been subject to the bisulfite conversions process and those that have not. The first group provide information about the methylation rate whereas the second group provide the majority of the information about the genotype call. If we consider reads that map to the C2T converted reference, and after converting the reads to the forward direction on the top strand, the bases C and T have been subject to bisulfite conversion while the bases A and G have not. Similarly, if we consider reads that map to the G2A converted reference, the bases A and G have been subjected to bisulfite conversion while C and T bases have not. For the following development of the model the potentially converted bases will be written in lower case and the other bases will be written in upper case. Reads mapping to the C2T strand will therefore provide the bases A, c, G, t while reads mapping to the G2A strand will provide the bases a, C, g, T (again, after placing all reads on the top strand in the forward direction).

Let the frequencies of the 4 possible bases be given by $F = (f_A, f_C, f_G, f_T)$. Rather than using F directly, we parameterize F in terms of (w, p, q) where:

$$
\begin{aligned}
w &= f_C + f_T \\
p &= \begin{cases} f_C/(f_C + f_T) & \text{if } w > 0, \\ 0 & \text{if } w = 0. \end{cases} \\
q &= \begin{cases} f_G/(f_G + f_A) & \text{if } w < 1, \\ 0 & \text{if } w = 1. \end{cases}
\end{aligned}
$$

Each genotype can now be described in terms of (w, p, q), e.g., $F(\mathrm{AC}) = (1/2, 1, 0), F(\mathrm{GG}) = (0, 0, 1)$. Let ϵ be the sequencing error rate, μ the methylation rate, λ the probability that a non-methylated C is converted to T, and τ the probability that a methylated C is converted to T. If we assume the simple model that methylation, conversion and errors

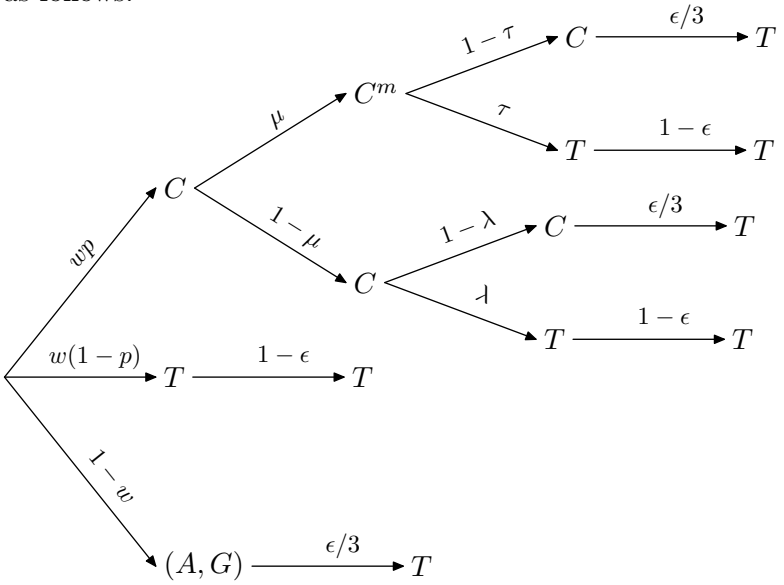
operate independently for each observed base, then taking as example an observed c base, the possible paths that could lead to the observation are shown below:



The probability of the observation can then be calculated as the sum of the probabilities of the paths therefore:

$$\begin{aligned} p(c) &= wp(\mu((1-\tau)(1-\epsilon)+\tau\epsilon/3)+(1-\mu)((1-\lambda)(1-\epsilon)+\lambda\epsilon/3))+(1-wp)\epsilon/3 \\ &= \frac{1}{3}\big((3-4\epsilon)wp(\mu(\lambda-\tau)+1-\lambda)+\epsilon\big) \end{aligned}$$

In a similar way, for an observed t base the probability graph and function would be as follows:



$$\begin{aligned} p(t) &= wp(\mu((1-\tau)\epsilon/3 + \tau(1-\epsilon)) + (1-\mu)((1-\lambda)\epsilon/3 + \lambda(1-\epsilon))) + w(1-p)(1-\epsilon) + (1-w)\epsilon/3 \\ &= \frac{1}{3}\big((3-4\epsilon)(wp(\lambda - \mu(\lambda-\tau)) + w(1-p)) + \epsilon\big) \end{aligned}$$

Using the substitutions $z = \mu(\lambda - \tau) + 1 - \lambda$ and $\kappa = \epsilon/(3-4\epsilon)$, the two probabilities can be written more concisely as:

$$\begin{aligned} p(c) &= \frac{3-4\epsilon}{3}(wzp + \kappa) \\ p(t) &= \frac{3-4\epsilon}{3}((w(1-zp) + \kappa) \end{aligned}$$

We note that methylation can be detected on both strands, c and t bases give information about methylation on the positive strand and g and a bases give information about methylation on the negative strand. If the genotype at a given site is CG then methylation could be observed on both strands. In this case it is reasonable to consider that the methylation process is independent on the two strands, so the values of μ and z are strand specific; in the following equations we use + and - subscripts to distinguish the strand specific variables (i.e., m_+ is the methylation on the + strand).

The sequencing error probabilities, ϵ (and therefore κ) vary between observations as they derive from the individual base quality scores assigned during sequence generation. Observation specific errors can be accounted for in the model, however it is more convenient to consider a common error rate for all bases of a particular type (A, C, G, T, a, c, g, t) that are estimated from the geometric average of the individual error rates (or, equivalently, the arithmetic average of the phred scaled sequencing quality scores). For a base type x, therefore , we have common values for ϵ_x and κ_x). The log likelihood of the set of base counts $(n_A, n_C, n_G, n_T, n_a, n_c, n_g, n_t)$ can then be written as follows (where K is a constant that doesn't depend on any of the unknown variables):

$$\begin{aligned} L =& K + \\ & n_A \log((1-w)(1-q) + \kappa_A) + n_C \log(wp + \kappa_C) + \\ & n_G \log((1-w)q + \kappa_G) + n_T \log(w(1-p) + \kappa_T) + \\ & n_c \log(wpz_+ \kappa_c) + n_t \log(w(1-pz_+) + \kappa_t) + \\ & n_g \log((1-w)qz_- + \kappa_g) + n_a \log((1-w)(1-qz_-) + \kappa_a) \end{aligned}$$

3. Methylation and genotype estimation

Genotype estimation is performed by calculating the log likelihood for each of the 10 possible genotypes and selecting the that with the highest likelihood. For genotypes where $f_C > 0$ or $f_G > 0$ then it an estimate of the relevant methylation rate is required. Conditional on genotype, expressions for the maximum likelihood estimates

of the methylation rates can be obtained by setting the partial derivatives of the log likelihood to zero:

$$\begin{aligned}
\hat{z_+} &= \frac{n_c(w-\kappa_t) - n_t\kappa_c}{(n_c+n_t)wp}, \\
\hat{z_-} &= \frac{n_g(1-w-\kappa_a) - n_a\kappa_g}{(n_g+n_a)(1-w)q}, \\
\hat{m_+} &= \frac{\hat{z_+} - 1 + \lambda}{\lambda - \tau}, \\
\hat{m_-} &= \frac{\hat{z_-} - 1 + \lambda}{\lambda - \tau}.
\end{aligned}$$

If $\hat{m_+}$ or $\hat{m_-}$ is less than zero or more than 1 then the estimate is truncated to zero or one respectively. We note that for homozygous C or G in the absence of sequencing errors and with perfect conversion (so $\lambda = 1, \tau = 0$) then the estimator above reduces to the commonly used formulae for methylation:

$$\hat{m_+} = \frac{n_c}{(n_c+n_t)}, \hat{m_-} = \frac{n_g}{(n_g+n_a)}$$

4. Goodness of fit test

We implement a goodness of fit test where we first maximize the likelihood allowing w, p, q, m_+, m_- to vary independently between 0 and 1; this we call the free model as there are no constraints imposed on (w, p, q). The goodness of fit statistic is then the difference between the $\log_{10}$ likelihoods of the free model and the best genotype model. This log likelihood ratio statistic is large if the data do not fit well a diploid model (for example, if there is evidence of >2 alleles or if the allelic ratio in heterozygotes is skewed from 1/2).

SUPPLEMENT

1. Benchmark of GEM3 and BScall against other analysis tools

We used a publically available dataset from the BLUEPRINT consortium (dataset: EGAD00001002322, sample: EGAN00001170522), namely a plasma cell sample extracted from the bone marrow (2x101bp, 58X genome coverage, additionally down sampled to 27X). All software was run with default parameters according to the authors specification on nodes of 2 x Xeon E5-2680v3 (12cores each) with 2.5 GHz and 256 GB of main memory using a Linux operative system (Red Hat 6.7). Mapping was performed against human genome assembly GRCh38.

Table S1. Software and parameters used in this study

Task	Software (Version)	Components (Version)	Parameter settings	Reference
Read alignment	Bismark (0.16.1)	bowtie2 (2.2.9)	-p 4	Krueger et al. (2011), Langmead et al (2012)
	BSMAP (2.9.0)	BSMAP (2.9.0)	-w 2 -q 20 -z 33 -p 8 -r 0	Xi et al. (2009)
	Novoalign (3.5.1)	Novoalign (3.5.1)	-t 20,2.5 --hlimit 7 -b2 -H 20	http://www.novocraft.com/
	Bwa-meth (0.10)	BWA (0.7.7-r441*)	-t 24	Pedersen (2014), Li (2013)
	gemBS (3.0)	GEM3 (3.1.0)	-p -r -m 1 -M 4	Marco-Sola et al. (2012), Marco-Sola (2017)
SNP and methylation calling	gemBS (3.0)	BScall (2.1)	-L5 -p	
	Bis-SNP (0.82.2)	Bis-SNP (0.82.2)	(1)-maxQ 40 (2)-stand_call_conf 20 -mbq 0 -stand_emit_conf 0 -mmq 30	Liu et al. (2012)
	MethylExtract (1.8.2)	MethylExtract (1.8.2)	flagW=99,147 flagC=83,163 p=24 FirstIgnor=5	Barturen et al (2013)

(1) Table Recalibration (Mills and 1000 Genome Gold standard indels, dbsnp v.138)

(2) SNP genotyper

Table S2. Read alignments

Cov	Mapper	Bases aligned	%	Quality > 20	%	Read Pairs Mapped	%	Read Pairs Uniquely Mapped	%
27X	Bismark	72.260.168.008	80%	69.732.899.548	77%	776.609.162	87%	715.449.606	80%
	BSMAP	76.588.875.760	85%	76.588.875.760	85%	737.709.554	82%	737.709.554	82%
	Bwa-meth	87.788.193.351	97%	73.369.620.896	81%	869.922.674	97%	755.131.135	84%
	GEM3	87.631.005.316	97%	77.255.445.602	85%	831.555.476	93%	748.962.640	84%
	Novalign	77.397.811.626	86%	76.602.297.580	85%	856.003.322	96%	786.633.096	88%
58X	Bismark	150.550.722.614	80%	145.284.706.844	77%	1.618.021.532	87%	1.490.606.206	80%
	BSMAP	159.567.965.437	85%	159.567.965.437	85%	1.536.961.596	82%	1.536.961.596	82%
	Bwa-meth	182.936.003.960	97%	152.757.232.094	81%	1.812.656.124	97%	1.572.235.855	84%
	GEM3	182.573.758.778	97%	160.959.781.503	85%	1.732.454.186	93%	1.560.450.802	84%
	Novalign	161.253.221.507	86%	159.595.406.636	85%	1.783.428.846	96%	1.638.898.342	88%

Table S3. SNPs called by different callers based on alignments from multiple mappers

Coverage	Caller	Mapper	SNPs
27X	Bis-SNP	Bismark	2.392.513
		BSMAP	2.855.296
		GEM3	2.716.501
	BScall	Bismark	5.157.821
		BSMAP	7.284.116
		GEM3	5.872.541
	MethylExtract	Bismark	4.093.287
		BSMAP	4.887.502
		GEM3	4.780.534
58X	Bis-SNP	Bismark	2.909.705
		BSMAP	3.448.045
		GEM3	3.182.798
	BScall	Bismark	5.138.678
		BSMAP	7.948.014
		GEM3	6.051.209
	MethylExtract	Bismark	4.438.764
		BSMAP	5.714.662
		GEM3	5.049.129

Table S4. Number and methylation level of CpGs detected (58X coverage)

CpG	BScall	Bis-SNP	MethylExtract
Unmethylated (0-30%)	5.666.889	5.360.912	7.277.673
Intermediate methylated (30-70%)	6.670.084	6.498.758	7.835.127
High methylated (70-100%)	14.304.843	13.469.834	14.876.963
Total	26.641.816	25.329.504	29.989.763

Table S5. Bis-SNP/BScall and MethylExtract/BScall pairwise comparison. Number and methylation levels of shared CpGs (58X coverage)

CpG	Shared Equal Methylated	Shared Different Methylated	Private Bis-SNP	Private BScall
Unmethylated	5.093.653 (21,32%)		286.807 (23,64%)	415.006 (21,85%)
Intermediate methylated	5.841.360 (24,45%)		326.378 (26,9%)	471.822 (24,84%)
Methylated	12.953.795 (54,22%)		599.670 (49,44%)	1.012.272 (53.30%)
Total	23.888.808 (85,76%)	853.908 (3,06%)	1.212.855 (4,35%)	1.899.100 (6,81%)

CpG	Shared Equal Methylated	Shared Different Methylated	Private MethylExtract	Private BScall
Unmethylated	5.684.621 (21,53 %)		1.573.343 (45,83 %)	66.618 (78,40 %)
Intermediate methylated	6.853.342 (25,96 %)		860.520 (25,06 %)	10.650 (12,53 %)
Methylated	13.856.699 (52,49 %)		999.052 (29,02 %)	7.700 (9,06 %)
Total	26.394.662 (87,76 %)	162.186 (0,53 %)	3.432.915 (11,41 %)	84.968 (0,28 %)

2. gemBS across multiple BS-Seq protocol

Publically available data for different bisulfite sequencing protocols was downloaded through the SRA (see Table S7) and processed with standard gemBS configurations.

Table S6. Standard gemBS configurations for different BS-Seq protocols

gemBS task	Parameter	WGBS	RRBS	PBAT
mapping	non_stranded:[1]	FALSE	FALSE	TRUE
	remove_individual_bams:	TRUE	TRUE	TRUE
	mapq_threshold:	10	10	10
calling	qual_threshold:	13	13	13
	reference_bias: [2]	2	2	2
	left_trim: [3]	5	5	5
	right_trim:[3]	0	0	0
	keep_improper_pairs:	FALSE	FALSE	TRUE
	keep_duplicates:	FALSE	TRUE	FALSE
	haploid:	FALSE	FALSE	FALSE
	conversion:	auto	auto	auto
	remove_individual_bcfs:	TRUE	TRUE	TRUE
	contig_pool_limit:	25000000	25000000	25000000
	mode:	strand_specifc	strand_specifc	strand_specifc
extract	phred_threshold:[5]	10	10	10

[1] selects the proper C->T and G->A read conversions for directional and non-directional libraries
[2] weight given to the reference genotype
[3] bases clipped from the respective end of a read (pair) to remove artificial introduced cytosines
[4] conversion rate unmethylated cytosine = 95%, methylated cytosine = 0.01%
[5] phred scaled genotype quality score

3. gemBS across multiple BS-Seq protocol

Validation for gemBS SNP calls was performed using data from the 1000 Genomes Project (sample: NA12878, run: SRR622457 (2 x 101bp, 88X genome coverage)). Simulated WGBS data was generated by *in silico* bisulfite conversion using the methylation profile from the plasma cell sample as a template and with values for the bisulfite conversion rates similar to those seen in the same sample (99% conversion rate for non-methylated cytosines; 5% conversion for methylated cytosines). Sub-sampling was performed by randomly selecting read-pairs prior to the alignment stage. gemBS was run with the standard parameters and filters. FreeBayes was run with the standard parameters except that the –genotype-qualities flag was set (to allow generation of the GQ fields for filtering) and the -@ option was used to force FreeBayes to produce output.

for all SNP sites in the reference panel, whether the sites were variant or not in the test samples.