BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 13 2012, pages 1698–1704
doi:10.1093/bioinformatics/bts254

Sequence analysis

Advance Access publication May 10, 2012

Fast and sensitive mapping of bisulfite-treated sequencing data

Christian Otto[1,2], Peter F. Stadler[1,2,3,4,5,6] and Steve Hoffmann[1,2,*]

[1]Interdisciplinary Center for Bioinformatics and Bioinformatics Group, Department of Computer Science, University Leipzig, 04107 Leipzig, Germany, [2]Transcriptome Bioinformatics Group, LIFE — Leipzig Research Center for Civilization Diseases, University Leipzig, 04107 Leipzig, Germany, [3]RNomics Group, Fraunhofer Institute for Cell Therapy and Immunology, 04103 Leipzig, Germany, [4]Santa Fe Institute, Santa Fe, NM 87501 USA, [5]Department of Theoretical Chemistry, University of Vienna, A-1090 Vienna, Austria and [6]Max-Planck-Institute for Mathematics in Sciences, 04103 Leipzig, Germany

Associate Editor: Michael Brudno

ABSTRACT

Motivation: Cytosine DNA methylation is one of the major epigenetic modifications and influences gene expression, developmental processes, X-chromosome inactivation, and genomic imprinting. Aberrant methylation is furthermore known to be associated with several diseases including cancer. The gold standard to determine DNA methylation on genome-wide scales is 'bisulfite sequencing': DNA fragments are treated with sodium bisulfite resulting in the conversion of unmethylated cytosines into uracils, whereas methylated cytosines remain unchanged. The resulting sequencing reads thus exhibit asymmetric bisulfite-related mismatches and suffer from an effective reduction of the alphabet size in the unmethylated regions, rendering the mapping of bisulfite sequencing reads computationally much more demanding. As a consequence, currently available read mapping software often fails to achieve high sensitivity and in many cases requires unrealistic computational resources to cope with large real-life datasets.

Results: In this study, we present a seed-based approach based on enhanced suffix arrays in conjunction with Myers bit-vector algorithm to efficiently extend seeds to optimal semi-global alignments while allowing for bisulfite-related substitutions. It outperforms most current approaches in terms of sensitivity and performs time-competitive in mapping hundreds of millions of sequencing reads to vertebrate genomes.

Availability: The software `segemehl` is freely available at `http://www.bioinf.uni-leipzig.de/Software/segemehl`.

Contact: E-mail: steve@bioinf.uni-leipzig.de

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

Received on November 7, 2011; revised on April 24, 2012; accepted on April 25, 2012

1 INTRODUCTION

Cytosine DNA methylation is one of the major epigenetic modifications in eukaryotes (Esteller, 2005). The epigenetic modification pathways governing DNA methylation and histone modifications are strongly coupled with each other (Cedar and Bergman, 2009). Hypermethylations in promotors of genes are associated with stable repression of its activity (as in case of X-chromosome inactivation) that can be maintained throughout cell divisions (Weber and Schübeler, 2007). During mammalian development, methylation patterns are largely rearranged. In very early stages, methylation marks are erased to allow flexible short-term regulation by histone modifications, while wide-spread *de novo* methylations in later stages enable long-term silencing of pluripotency-related or imprinted genes (Reik, 2007). In mammalian genomes, DNA methylation also ensures genomic integrity by inactivating and immobilizing transposable elements and hence preventing chromosomal instability, translocation, or gene disruption (Weber and Schübeler, 2007). In cancer cells, this overall stable landscape of DNA methylations is heavily distorted by wide-spread and massive hypomethylations, e.g. in repetitive sequences, and by silencing of tumor-suppressor genes by hypermethylating their promotors (Esteller, 2007).

Capturing DNA methylations on genome-wide scales in high resolution has become technically and economically feasible only with the advent of high-throughput sequencing (HTS) technologies. DNA methylations are commonly captured either by sequencing methylated DNA that was isolated by antibodies or proteins, as in *methylated DNA immunoprecipitation* (Weber *et al.*, 2005) and methyl-CpG binding domain-based (MBD) isolated genome sequencing (Serre *et al.*, 2010), or by sequencing DNA reads treated with sodium bisulfite to selectively convert unmethylated cytosines to uracils (Frommer *et al.*, 1992). Since the first approach merely enriches sequencing reads with methylation marks by pull down with antibodies or proteins, it is not possible to accurately pinpoint frequency, exact location, and sequence context of the modifications. The isolation procedure is further biased towards enrichment of highly methylated regions (Lister and Ecker, 2009). Sequencing techniques based on bisulfite treatment, on the other hand, facilitate single-base resolution, so that the methylation state of each single cytosine can be analyzed. Thus, they are capable of detecting intermediate methylation levels in heterogeneous samples or imprinted genes. One drawback of this method is the fact that hydroxymethylated cytosines, present in some mammalian cell types, cannot be distinguished from methylation marks after conversion with sodium bisulfite. The role of the hydroxymethylation is not yet known but it might be involved in demethylation or alterations of the chromatin structure (Huang *et al.*, 2010).

Due to its high resolution and the possibility of unbiased genome coverage, bisulfite sequencing has been established as the 'gold

[*] To whom correspondence should be addressed.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

© The Author 2012. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

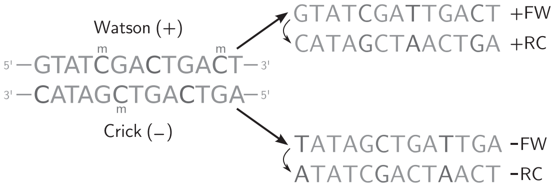

Fig. 1. Possible read types (+FW, +RC, −FW and −RC) in bisulfite sequencing protocols. Methylated and unmethylated cytosines in the genomic sequence (left) are coloured in red and blue, respectively, and positions in the read sequences (right) derived from genomic cytosines are coloured correspondingly. Note that the intermediate conversion of unmethylated cytosines into uracils after bisulfite treatment is omitted

standard' method to capture DNA methylation. In the earliest approach of this type, reduced representation bisulfite sequencing by Meissner *et al.* (2005), genomic regions with CpG dinucleotides are enriched by prior digestion with MspI. More recent protocols avoid this bias. Both methylC-seq (Lister *et al.*, 2009) and BS-seq (Cokus *et al.*, 2008) are protocols for the construction of the bisulfite-treated libraries for HTS. They mainly differ in their amplification procedure: while methylC-seq involves only a single amplification, BS-seq uses two amplification steps to ensure only fully bisulfite-converted sequences to be amplified and hence sequenced. In BS-seq, first, adapters containing unmethylated cytosines are ligated to the DNA fragment. After treatment with sodium bisulfite, the first amplification is performed using primers complementary to fully bisulfite-converted adapters, then digested with DnpI and again amplified using common Solexa adapters. This results in four different types of bisulfite reads: +FW and +RC from the plus strand and −FW and −RC from the minus strand (Fig. 1). In case of methylC-seq, only two of these read types (+FW and −FW) may occur and are expected to be sequenced at similar rate. Beyond the extensive study of Lister *et al.* (2009) as part of the UCSD Human Reference Epigenome Mapping Project, the methylomes of silkworm (Xiang *et al.*, 2010), honey bee (Lyko *et al.*, 2010); and Human peripheral blood mononuclear cells (Li *et al.*, 2010) have been analyzed by means of bisulfite sequencing. Moreover, this technology has been applied to identify methylation variations in epigenetic domains across cancer types (Hansen *et al.*, 2011).

Standard DNAseq mapping algorithms may run into problems when dealing with the potentially high number of converted cytosines in bisulfite sequencing reads: the bisulfite conversion causes a large number of mismatches between read and reference genome that should not be penalized. The asymmetry of the resulting matching rule, i.e. a genomic cytosine should match a thymine in the read but not *vice versa*, complicates the issue. Early bisulfite mapping methods used very time-consuming strategies. `BSMAP` (Xi and Li, 2009), for instance, iterates over all possible C/T conversions, `CokusAlignment` (Cokus *et al.*, 2008) uses an exhausting tree search with base probability vectors. More recent methods either allow for asymmetric bisulfite-related mismatches, typically implemented by means of hash-tables as in `MAQ` (Li *et al.*, 2008) and `RMAP` (Smith *et al.*, 2009), or use a collapsed alphabet so that the asymmetry is disregarded altogether. In the latter type of methods, each cytosine is converted to a thymine (or guanine to adenine to match the minus strand) in the reads and in the genomic sequence. Both `BS Seeker` (Chen *et al.*, 2010) and `Bismark` (Krueger and Andrews, 2011) use `Bowtie` (Langmead *et al.*, 2009) to map the converted strings using different alignment policies. The resulting alignments are then post-processed to recover the methylated positions. None of the available tools can account for insertions and deletions (indels) in the read alignment. This is a major drawback since indels are known to be the predominant error type in 454 sequencing data and small indels contribute significantly to the genetic variation in human (Mills *et al.*, 2011). Overall, currently available bisulfite mappers may not be able to cope with higher error rates potentially caused by erroneous PCR clones, low-quality reference genomes, extensive allelic variations, or mapping to the genome of a closely related organism. For example, the amphioxus genome exhibits substantial allelic variation with 3.7% SNPs and 6.8% polymorphic indels (Putnam *et al.*, 2008). In the *Ciona intestinalis*, another important model organism, the average SNP rate is 1.2% but the variations are not uniformly distributed and locally increase to 10–15% within windows of 100 nt (Dehal *et al.*, 2002).

`segemehl` is an efficient read mapping tool based on suffix arrays that readily accommodates indels using an extended version of the matching statistics (Hoffmann *et al.*, 2009). In this study, we demonstrate that the mapping of bisulfite sequencing data can be incorporated into the framework using a hybrid approach that combines seed searches in the suffix array on a collapsed alphabet with optimal semi-global alignments around seed matches using a specialized extension of Myers bit-vector algorithm.

2 RESULTS

2.1 Seed search on collapsed alphabet

Our matching strategy uses seeds that serve as anchors for subsequent semi-global alignments (see Supplementary Fig. S1). The seed search is efficiently facilitated in `segemehl` by an enhanced suffix array (ESA; Abouelhoda *et al.*, 2004). The ESA data structure supports exact query searches with an effort proportional to the length *m* of the query sequence and largely independent of the size of the reference genome. Our method aims to identify seed matches starting at each position. To improve the sensitivity of the seed search in the presence of sequencing errors, a limited number of mismatches as well as indels (insertions and deletions) is evalutated. For the technical details, we refer to Hoffmann *et al.* (2009).

To facilitate efficient mapping of bisulfite-treated sequencing data, it is necessary to cope with a high number of bisulfite-related mismatches. To overcome this issue, the nucleotide alphabet is collapsed to three characters. To enable the seed search on both strands, two reference genomes and their corresponding ESAs, one with C-to-T and one with G-to-A conversions, are created. Since forward (C-to-T) and backward (G-to-A) ESA can be used consecutively, only disk storage but not core memory is affected. The reduced alphabet requires somewhat longer seeds to ensure unambiguous matches, leading to an increase in runtime by a small constant factor compared to ordinary read matching.

2.2 Myers bit-vector algorithm

After seed matching, `segemehl` calculates semi-global alignments of the query with the reference genome loci indicated by the seeds using the fast bit-vector algorithm of Myers (1999). To prevent spurious hits, `segemehl` uses a user-defined accuracy threshold (option `-A`) specifying the minimal required percentage of matches

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

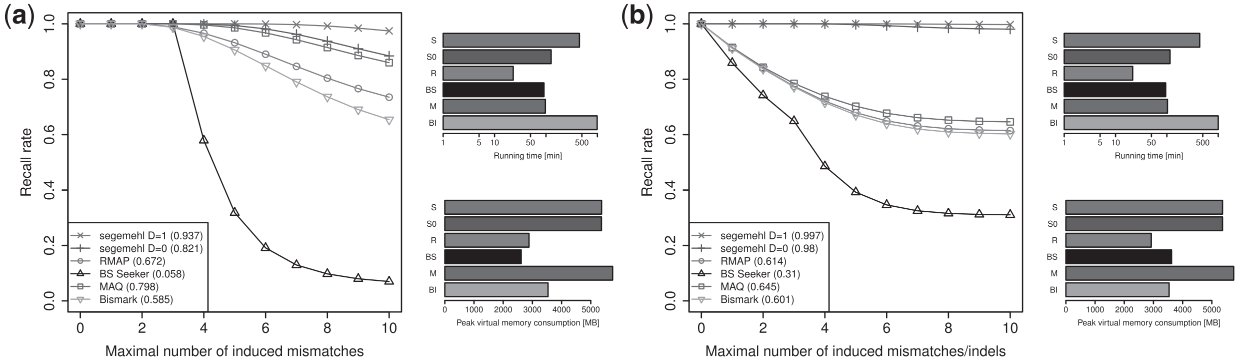

Fig. 2. Performance evaluation on artificial datasets. The benchmarks assessed the performance of `segemehl` with D = 1 (in red) and D = 0 (in dark red), `RMAP` (in green), `BS Seeker` (in black), `MAQ` (in blue) and `Bismark` (in orange) in terms of recall rate, running time (in user mode) and peak virtual memory consumption by mapping 10 million artificial bisulfite reads to a 200 MB large random reference. Furthermore, **(a)** mismatches at a rate of 10% or **(b)** mismatches + indels at a rate of 5% were randomly introduced into the bisulfite reads. The recall rate is the relative number of mapped reads where the score of the best alignment is found to be unique and the original position on the artificial reference was recovered correctly. The recall rate was estimated on subsets of the artificial reads with limited number of introduced mismatches or mismatches + indels. The overall recall rate of each program with the entire query dataset is given in the legend. Note that the preprocessing time is not included in the measurement

within the calculated read alignment. By default, `segemehl` reports all read matches where the minimal accuracy criteria is met. In case of best-only (option `-H 1`), only those read matches are reported whose alignment contains the minimal number of errors (mismatches + insertions + deletions) among all matches.

To efficiently cope with bisulfite conversions, we further extended the bit-vector algorithm of Myers to fully support the nucleic acid code of the International Union of Pure and Applied Chemistry (IUPAC) which encodes nucleotide ambiguity, e.g. the IUPAC nucleotide symbol 'Y' denotes either cytosine or thymine. By means of the IUPAC nucleotide code, bisulfite-sensitive alignments can be computed where asymmetric bisulfite-related conversions are implicitly treated as matches. The overhead of this extension of Myers' algorithm is only nominal, see Methods for further details. Overall, the major advantage of this mapping strategy is that, in contrast to other bisulfite mappers, no post-processing of the mapping is required.

2.3 Evaluation on artificial data

To evaluate the performance of the bisulfite version of `segemehl`, we compared it with existing methods on artificial and real-life datasets. The artificial query datasets were composed of 10 million reads of length 80 nt randomly selected from a 200 MB large reference sequence. The reference itself was generated with a uniform nucleotide distribution and randomly methylated cytosines on both strands at a rate of 50%. To mimick the methylC-seq protocol, only +FW and −FW reads were generated from each strand of the reference. The sodium bisulfite treatment was simulated by converting each unmethylated cytosine on the reference into a thymine. We remark that `segemehl` can also map bisulfite sequencing data generated with the BS-seq library preparation protocol of Cokus *et al.* (2008). In this case, the mapping is extended to both strands with each of the alphabet conversions rather than only C-to-T on the plus strand and G-to-A on the minus strand of the genome sequence. The sensitivity of `segemehl` on artificial datasets mimicking the BS-seq library preparations are very similar to the results on methylC-seq datasets (data not shown).

To consider genomic aberrations such as mutations and polymorphisms as well as sequencing errors, we further introduced random sequence errors into the bisulfite reads at different error rates (5 and 10%) and for error types (mismatches or mismatches and indels at the ratio 4:1). In our benchmark, we compare `segemehl` v0.1 with `RMAP` v2.05, `BS Seeker`, `MAQ` v0.7.1 and `Bismark` v0.5.1. We executed all programs with default parameters but adjusted some options, e.g. error limits and filtering constraints, to obtain more sensitive mappings and hence to assess the capability of each approach to cope with more difficult settings, see Section 4 for further details on parameters and the evaluation procedure. Overall, `segemehl` obtained recall rates >92% (with D = 1 difference allowed in the seed) and 81% (for exact seed matches) in every setting and hence outperforms all other programs in this respect. In the least challenging scenario with low mismatch rate (5%), all programs except for `BS Seeker` are able to recover the original position of >89% of the reads correctly (Supplementary Fig. S2a). By increasing the mismatch rate (10%), the recall rates of other programs drop considerably down to 60 and 70%. `segemehl` still achieves a recall rate >93% (D = 1) and 82% (D = 0). Among the other programs, `MAQ` performs best and is only slightly inferior to `segemehl` in its less sensitive setting (Fig. 2a). As for the introduction of indels into the read data, `segemehl` largely retains its good performance, while a substantial loss is observed with the other tools (Fig. 2b and Supplemantary Fig. S2b). This is also true in case of the artificial dataset with the low error rate (5%) including only few indels (Fig. 2b). In more challenging scenarios, the recall rates of these programs even drops <40% (Supplemantary Fig. S2b).

`segemehl` obtains the higher recall rates at the cost of a reduced time performance. On average, `segemehl` with D = 1 is ∼5-fold slower than with D=0. Hence, the choice of this parameter is a tradeoff between speed and recall and is dependent on the user's requirements. The running times of `MAQ` and `BS Seeker` are comparable to `segemehl` with lower sensitivity whereas `RMAP` is about four times faster. In terms of memory, the programs consume between 2.6 GB (in case of `BS Seeker`) and 5.6 GB of virtual memory (in case of `MAQ`). Note that the actual amount of used

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

physical memory is lower than the virtual memory consumption. For example, segemehl requires ∼5.2 GB of virtual but only 3.2 GB of physical memory.

To verify that the superior sensitivity of segemehl does not lead to a substantial loss of mapping specificity, we counted the number of false positive mappings in each artificial benchmark. Among the uniquely mapped reads with <13 mismatches+indels, segemehl does not report a single false positive mapping. Thus, by limiting the number of permitted errors and restricting to uniquely mapped reads, segemehl does not lose specificity while achieving very high sensitivities.

To assess whether higher mapping sensitivities may also assist in calling the methylation state more accurately, we assembled additional methylation calling benchmarks. The datasets with an expected 10-fold coverage were composed of 2.5 million bisulfite reads, mimicking the methylC-seq protocol, which were randomly selected from a 10 MB large reference sequence where 50% of the cytosines on each strand have been artificially methylated. Furthermore, errors (mismatches or mismatches + indels) were introduced in the read sequences at different rates (5, 10 and 15%). To assess the performance of re-calling, these artificial methylation states using the mapping output, we mapped each dataset with each of the bisulfite mapping tools, filtered out ambiguously mapped reads and determined the methylation states using simple majority voting under a minimal coverage of 5, see Section 4 for further details. For each dataset and mapping tools, recall rates and false discovery rates at different score cutoffs were estimated (see Fig. 3 and Supplementary Fig. S3).

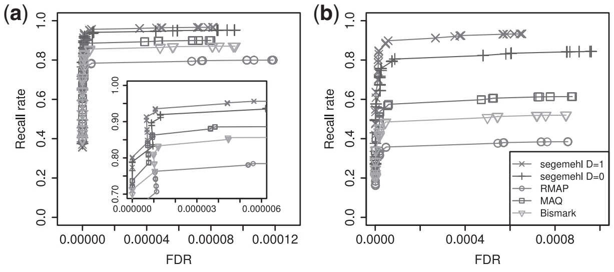

Fig. 3. Performance in methylation calling benchmarks. Recall rate as function of FDR after evaluating the performance in methylation calling using the mapping output of segemehl with D=1 (in red) and D=0 (in dark red), RMAP (in green), MAQ (in blue) and Bismark (in orange). We therefore mapped 2.5 million artificial bisulfite reads, containing mismatches at a rate of **(a)** 10% or **(b)** 15%, with each program to a 10 MB large reference sequence, see Section 4 for details on generation and evaluation of the datasets. The inlay in the left panel magnifies the area where the FDR is close to zero (same units on axes). Note that the same colours and symbols are used in both panels. The peak recall rates with 10% mismatches (left panel) are 0.966 and 0.952 for segemehl with D=1 and D=0, respectively, 0.8 for RMAP, 0.9 for MAQ and 0.871 for Bismark. In case of 15% mismatches (right panel), the peak recall rates are 0.933 and 0.843 for segemehl with D=1 and D=0, respectively, 0.385 for RMAP, 0.614 for MAQ and 0.52 for Bismark

Overall, the methylation calls using segemehl's mapping output obtained higher recall rates at a lower false discovery rate in every setting compared with the other mappers. For example, with low and medium error rates (5% and 10%), it is possible to recover $>95\%$ of the methylation marks correctly with segemehl whereas the recall rates of methylation calls using the output of RMAP, MAQ and Bismark vary between 80% and 90%. In the most challenging scenarios with high error rates (15%), the mapping output of segemehl can still be used to infer the methylation state of $>84\%$ and 93% of the cytosines with D=0 and D=1, respectively, while retaining FDRs $<0.1\%$. In addition, we simulated bisulfite reads from an artificial genome containing sites with four different methylation rates (20, 40, 60 or 80%). We estimated the methylation rates and calculated the differences from the simulated levels (error) for the alignments of each program. Errors of the segemehl based estimator were compared with estimators based on other alignment methods. Overall, with segemehl alignments, the accuracy of the estimated methylation rates is superior to the other tools tested—in particular in benchmarks with higher error rates (see Supplementary Fig. S4 and S5).

We emphasize that methylation calling is primarily a statistical problem inherently distinct from read mapping. Hence, we used here simple benchmark settings with uniform methylation patterns and sequencing errors. Partial chemical conversion, for instance, may reduce the sensitivity of a simple methylation calling procedure such as majority voting and call for a more sophisticated statistical model. It does not affect, however, the mappability of individiual reads.

2.4 Mapping of real-life data

Next, we compared the bisulfite mapping tools on two real-life datasets. Both SRR019048 (15 331 851 reads of length 87) and SRR019597 (5 943 586 reads of length 76) are part of the whole genome shotgun bisulfite sequencing dataset of the human H1 cell line by Lister *et al.* (2009). segemehl clearly outperforms the other programs in both datasets by reporting more mapped reads with a lower number of errors. The results including running time, memory usage, fraction of unique best mapped reads and fraction of unique best mapped reads at a given maximum error cutoff are given in Table 1. The latter measure makes it possible to determine whether a higher overall number of mapped reads is merely reached by allowing more errors in the read alignment or whether it is obtained by also mapping more reads with few errors indicating better mapping capabilities of the method. segemehl is able to map an additional number of ∼234000 and ∼88000 reads from the datasets SRR019048 and SRR019597, respectively, with only up to two mismatches, insertions or deletions. Only a small difference in the number of mapped reads is observed between segemehl's D=0 and D=1 options. Similar to the artificial benchmarks, the number of mapped reads with BS Seeker is significantly lower compared with RMAP, MAQ and Bismark, which show similar results in both real-life datasets. By allowing non-unique mappings, segemehl is able to obtain mappings for $>98\%$ of the reads in each of the real-life datasets. In addition to these rather challenging datasets due to their poor base calling qualities, we analyzed a recent bisulfite dataset with good base calling qualities by Lister *et al.* (2011) and obtained concordant results (see Supplementary Table S1).

Strikingly, the running time of segemehl is lower compared with RMAP, MAQ and Bismark even for the sensitive D=1 parameter setting. The increase varies from 13% to 189% for SRR019048 and SRR019597. The less sensitive setting comes with a 26-fold and 18-fold decrease in the running time compared with MAQ. In addition, RMAP and MAQ, in contrast to the other programs including segemehl, do not support multi-threading and hence cannot benefit from commonly available multi-core machines. This

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

Table 1. Performance evaluation on real-life datasets

	Running user	Memory^a	Mismatches + indels							
	time (min)	(MB)	=0	≤1	≤2	≤3	≤4	≤5	≤10	Max
SRR019048										
segemehl (D=1)	844	74 995	**22.8**	**33.4**	**38.8**	**43.2**	**47.2**	**51.1**	**68.7**	**87.2**
segemehl (D=0)	**224**	74 999	**22.8**	33.4	38.8	43.1	47.2	51.1	68.5	87.0
BS Seeker	247	9280	22.7	32.5	37.2	40.7	40.7	40.7	40.7	40.7
RMAP^b	1003	**7687**	22.8	32.6	37.3	40.8	43.8	46.7	60.0	78.3
MAQ^b	22 635	8798	22.7	32.1	36.5	39.8	42.7	45.5	58.1	79.5
Bismark	1909	14 649	22.7	32.3	36.9	40.3	43.2	46.0	58.5	79.1
SRR019597										
segemehl (D=1)	256	74 995	39.9	**53.3**	**59.7**	**64.2**	**67.9**	**71.2**	**82.9**	90.0
segemehl (D=0)	**72**	74 999	39.9	53.3	59.7	64.2	67.9	71.2	82.8	**90.2**
BS Seeker	86	6081	39.8	52.3	58.1	62.0	62.0	62.0	62.0	62.0
RMAP^b	487	5846	**39.9**	52.5	58.2	62.1	65.0	67.6	77.4	87.5
MAQ^b	4782	**3425**	39.1	51.3	56.8	60.5	63.3	65.8	75.1	86.5
Bismark	741	14 649	39.7	52.1	57.8	61.5	64.4	67.0	76.3	87.6

The tests assessed the performance of segemehl with D=0 or D=1 (-F 1, -H 1, -A 70), BS Seeker (-t N, -e 80, -m 3), RMAP (-B, -m 20), MAQ (-M c, -n 3, -e 500) and Bismark (-directional, -n 3, -e 500) by mapping two lanes of a whole genome shotgun bisulfite sequencing dataset of the human H1 cell line (published by Lister *et al.* 2009) against the human genome in terms of running time (in user mode), peak virtual memory consumption and fraction of unique best mapped reads (overall or subdivided by the maximal number of mismatches + indels in the alignment). Note that last measure only considers read mappings where the score of the best alignment is found to be unique. The best value in each measure, e.g., lowest running time, lowest memory consumption or highest number of unique best mapped reads, is printed in boldface. The real-life datasets consists of 15 331 851 reads of length 87 nt and 5 943 586 reads of length 76 nt in case of SRR019048 and SRR019597, respectively. Note that the preprocessing time is not included in the time measurement. Details on the selected parameters of each program are given in Section 4.

[a] Virtual memory consumption shown whereas the required physical memory considerably less. For example, segemehl uses only ~52 GB of physical memory. [b] RMAP and MAQ do not provide multi-threading.

is a major technical shortcoming in the light of the size of datasets to be mapped. The high mapping accuracy and speed is paid for by the rather high memory consumption, which exceeds that of the other tools by a factor of five to ten: the ESA of the human genome used by segemehl consumes ~73 GB of virtual memory but only 53 GB of the physical memory. The software thus requires equipment at the top end of what at present can be considered standard hardware. At the cost of higher running time, it is also possible to run segemehl on each chromosome separately with a peak memory consumption of 6 GB of RAM. Detailed information on merging the mapping output of each chromosome, updating sequence alignment map (SAM) tags, and enforcing (if desired) the best-only matching strategy is given on our website together with the necessary tools.

3 DISCUSSION

The analysis of bisulfite sequencing data has remained a challenging problem. Existing tools either do not provide an all-in-one solution but are based on post-processing output of common mapping tools (e.g. Bowtie) leading to losses in sensitivity, or show undesirable runtime performance—in particular for vertebrate datasets. In addition, none of the existing tools is able to consider insertions or deletions and even very few indels, e.g., originating from sequencing errors or genomic variations, effectively obstruct the mapping of sequencing reads. We have presented here a novel approach to this problem based on segemehl to efficiently perform bisulfite mapping with high sensitivity. Our method is insensitive to contaminations and handles insertions and deletions already during the initial seed search. Compared with competing methods, our approach provides significantly higher recall rates as measured on artificial datasets. Although the recall rate of most other tools is drastically reduced by a larger number of mismatches or a few indels in the read data, these effects only slightly affect the sensitivity of segemehl. This increase in sensitivity does not come at the cost of specificity and may finally result in better performance in calling methylation state or methylation level as well.

The algorithm is specifically designed to map also ambiguous reads. In some application scenarios, these reads are of interest and convey useful biological information. For example, repetitive elements were reported to be hypermethylated (Weber and Schübeler, 2007) but may be extensively demethylated during development (Gehring *et al.*, 2009) or tumorgenesis (Esteller, 2007; Watanabe and Maekawa, 2010).

Due to its highly time-efficient index structure, segemehl has strong advantages over the existing methods in mapping real-life datasets of human both in terms of sensitivity and running time, at the expense of a higher memory requirements. By supporting multi-threading, the software can furthermore take full advantage of multi-processor architectures and completes mapping of >540 million sequencing reads (SRX006240 dataset by Lister *et al.*, 2009) on the human genome in only around three and a half days using a two Quad-core machine with 64 GB of core memory. It further supports mapping of bisulfite sequencing data from both currently existing library protocol, methylC-seq and BS-seq, and provides output in standardized SAM format for which various post-processing utilities are available such as samtools (Li *et al.*, 2009).

In addition to mapping bisulfite sequencing data, our approach might also assist in mapping ancient DNA (Briggs *et al.*, 2007;

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

Prüfer *et al.*, 2010), where read ends are heavily exposed to deamination, i.e., cytosines are converted to thymine, over the large time-scales. Due to the short read length of ancient sequencing data, trimming of $5'$ and $3'$ end of reads may not be adequate and impede their mappability. By simply adjusting the conversion functions, this version of `segemehl` can also be applied to datasets generated with the PAR-CLIP protocol where protein binding sites can be identified genome-wide at high resolution by use of UV cross-linking and photoactivatable nucleosides such as 4SU or 6SG. These are specifically converted near cross-linking and hence binding sites and might assist in post-processing to reduce the number of false positives. By regarding these specific conversions as matches, `segemehl` becomes insensitive to the number of these conversions under any parameter setting. We have not investigated the performance of `segemehl` on these types of sequencing data so far. They are, however, a natural objective of future research.

4 METHODS

4.1 Seed search

In `segemehl`, the seed search is facilitated by use of the ESA as described in Hoffmann *et al.* (2009). In brief, the concept of suffix array is based on lexicographically sorting all suffixes of the genomic sequence. By additionally using lcp-table and child table, the ESA is equivalent to a suffix tree (Abouelhoda *et al.*, 2004). The suffix tree is a directed rooted tree in which edges are labeled with a non-empty string such that each suffix is formed by the concatenation of edge labels of exactly one path from the root to a leaf. Hence, a simple seed search in the ESA can be imagined as top-down traversal of the corresponding suffix tree with the query sequence. To facilitate imperfect seed searches and hence allow for mismatches, insertions and deletions, it is possible to enumerate alternative paths along the perfect matching path. However, the number of alternative matching paths increases exponentially with higher numbers of permitted errors. Hence, for the sake of time efficiency, the number of errors during seed search is limited. To perform seed searches at each query position, a greedy substring search was implemented. This approach uses suffix link information on previously computed matching paths and hence avoids recomputations. To construct an ESA, first, the suffix array table is generated by sorting all suffixes of the genomic sequence using the algorithm introduced by Ko and Aluru (2003). Second, the additional tables, namely lcp table, child table, and suffix link table, can be efficiently constructed according to Abouelhoda *et al.* (2004). Overall, the construction of an ESA index requires $\mathcal{O}(n)$ in time where n denotes the length of the genomic sequence.

For DNAseq and RNAseq reads, the seed search is performed on a four-letter nucleotide alphabet, $\Sigma_{\text{DNA}}=\{A,C,G,T\}$, in both read and genome sequence. In the case of mapping bisulfite sequencing reads, the substitutions of genomic cytosines to thymines in the read sequence need to be taken care of during the search and should not be penalized as errors. Considering these bisulfite conversions explicitly would imply a potentially exponential enumeration and hence hamper the mapping performance considerably. We therefore introduce two conversion functions $f_{\text{C}\to\text{T}}$ and $f_{\text{G}\to\text{A}}$ such that

$$f_{\text{C}\to\text{T}}(x)=\begin{cases}\text{T} & x=\text{C}\\ x & \text{otherwise}\end{cases}$$

$$f_{\text{G}\to\text{A}}(x)=\begin{cases}\text{A} & x=\text{G}\\ x & \text{otherwise}\end{cases}$$

where $x\in\Sigma_{\text{DNA}}$. In the first stage of the algorithm, the $f_{\text{C}\to\text{T}}$ converted reads are mapped to a reference that has been converted with $f_{\text{C}\to\text{T}}$. To consider bisulfite conversion on the minus strand, it is necessary to additionally map the read converted with $f_{\text{G}\to\text{A}}$ to the $f_{\text{G}\to\text{A}}$ converted reference in the second stage because DNA methylations are strand-specific. Note that by mapping the four-letter alphabet to a three-letter alphabet bisulfite-related conversions appear as matches but at the same time the asymmetry of the substitution leads to an implicit underestimation of the edit distance.

4.2 Bisulfite-sensitive semi-global alignment

Following the seed search in the ESA, `segemehl` extends the seeds to semi-global alignments. In contrast to the seed search on the converted references, the alignment should use the asymmetric bisulfite matching rule where a genomic cytosine and thymine in the read produces a match but not *vice versa*. For this purpose, we extended the highly efficient bit-vector algorithm of Myers (1999). Instead of computing only one entry in the dynamic programming matrix at a time, this algorithm computes w entries simultaneously where w as the word size of the machine. It takes advantage of the high efficiency of low-level bit operations due to bit-level parallelism in common processors. Thus, the core of Myers' bit-vector algorithm is entirely based on bit operations including differentiating between matches and mismatches which are initially precomputed and stored in bit-vectors. For each character x of the alphabet Σ_{DNA}, a bit-vector B_x of length m is constructed where m denotes the length of the read sequence r. Subsequently, the bit-vectors are initialized by a function in such a way that i-th bit in B_x is set iff x and the i-th character of r produce an alignment match. This algorithm has runtime of $\mathcal{O}((m/w)\cdot l)$ where l is the length of the read and reference substring. In our implementation, the reference substring is bounded by $m+2\cdot k$ where k is defined by the maximal permitted errors in the read alignment. Hence, the algorithm has a runtime complexity of $\mathcal{O}((m/w)\cdot(m+2\cdot k))$ in time. Since $w=64$ in our implementation, the algorithm runs in $\mathcal{O}(m+2\cdot k)$ for reads of size up to 64.

Commonly, the function to differentiate matches and mismatches simply tests for character equality. Here, it was extended to fully support the IUPAC nucleotide code. For example, the IUPAC symbol Y, denoting a pyrimidine, produces a match with both C and T. By converting Ts into Ys within the read sequence, the asymmetric bisulfite matching rule is implicitly integrated. Again, due to the strand specificity of DNA methylations, read sequences matching to the minus strand are translated differently, i.e., every adenine in the read is converted into an R, the IUPAC symbol for a purine. Overall, the necessary modifications only concern the initialization procedure and hence result merely in nominal overhead.

4.3 Benchmarking procedure

With a few exceptions, all programs were executed with default parameters for artificial and real-life datasets. Some options, such as error limits and filtering constraints, were adjusted to obtain a higher sensitivities. We executed `segemehl` in default mode, where at most one mismatch or indel is permitted in the seed (option `-D`) and where the maximum expectation value (option `-E`) is set to five. In addition, we also executed `segemehl` in a less sensitive but more time-performant configuration with `-D 0`. Seeds with >500 matches in the genome are dismissed by default (option `-M`). Due to the high number of read errors, the minimal required alignment accuracy (option `-A`) was adjusted to 80% and 70% in artificial and real-life benchmarks, respectively, and `segemehl` was set to report best-scoring hits only. For `BS Seeker`, `MAQ`, and `MAQ`, we permitted the maximum of three mismatches to obtain optimal sensitivity and adjusted the option `-e` in `BS Seeker` to the largest read length occurring in the benchmarks. In the same vein, the maximum number of mismatches was used for `RMAP` (`-m 20`). To avoid hits to be discarded due to the sum-of-base-qualities-policy, `MAQ` was executed with the option `-e 500`. The parameters were set analogously for `Bismark`. In contrast to `BS Seeker`, `RMAP` or `Bismark`, where non-unique best mapped reads (regarding their alignment score) are dismissed by default, `MAQ` reports a best hit in any case but assigns a mapping quality of zero in case multiple hits with equal score (sum of base qualities at mismatch positions) were found. Such ambiguously best mapped reads were rejected before any of the evaluations.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

For each program, artificial, and real-life datasets, we assessed running time (in user mode), peak virtual memory consumption and recall rate in mapping the different datasets on the same machine with two 2.27 GHz 64-Bit Quad-Core CPUs and 126 GB of RAM. The time and the memory measurements were performed using unix `ps`. Note that the preprocessing times for generating index structures like the ESA or Burrows–Wheeler transform are not included in the measures. We estimated the recall rate as the relative number of reads where the score of the best alignment is found to be unique (i.e. unique best mapped reads) and the original position on the artificial reference was recovered correctly. However, optimal read alignments under the unit cost model may become ambiguous with insertions and deletions. Therefore, any reported position with a deviation of less than 11 nt from its original position was deemed as correct. In addition to the overall recall rate of each program in the datasets, we calculated the recall rates at a given maximum number of read errors (mismatches or mismatches+indels). Note that the number of errors in the optimal read alignment may be smaller than the number of introduced errors. For example, an unmethylated cytosine that is converted to thymine during the bisulfite treatment but subsequently called as cytosine due to a base calling error will not affect the alignment score. For each program, we illustrated the overall running time and memory consumption as well as the recall rate of these subdatasets as function of their maximal number of introduced errors (Fig. 2 and Supplementary Fig. S2).

In the artificial methylation benchmarks, all programs were executed as described earlier and the mapping output of each mapper was used to call the methylation states. To ensure a fair comparison, we implemented a simple methylation caller based on majority voting. Given the position of a cytosine, the list of bases within read sequences, which uniquely map to this position on the same strand, commonly denoted as cross-section, is extracted and the most frequent base in the cross-section is determined. If there is a tie or the most frequent base is neither C nor T, the cytosine is not called and hence counted as false negative (FN). If the most abundant character is C or T, the cytosine is called methylated or unmethylated, respectively. The methylation call is counted as true positive (TP) if the call matches the artificial methylation state and counted as false positive (FP) otherwise. To limit the number of false positives, only cytosine sites were called where the coverage on the strand exceeded a given minimal value. We further calculate the methylation rate, that is the fraction of non-converted over the sum of non-converted and converted bases. The entire methylation caller is by construction unaware of any particular features of the mapping tools.

We provide the methylation caller as additional file. It uses the output generated by `mpileup` from the `samtools` package which can easily be created from any SAM- or BAM-formatted mapping output. By estimating the recall rate, that is TP/(TP+FN), and false discovery rate, that is FP/(TP+FP), we can compare the performance in methylation calling using the mapping output of each bisulfite mapper.

Funding: This publication was supported by LIFE and BMBF through ICGC MMML-Seq (01KU1002J). LIFE—Leipzig Research Center for Civilization Diseases, Universität Leipzig, is funded by means of the European Union, by the European Regional Development Fund (ERDF) and by means of the Free State of Saxony within the framework of the excellence initiative. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Conflict of Interest: none declared.

REFERENCES

Abouelhoda,M.I. *et al.* (2004) Replacing suffix trees with enhanced suffix arrays. *J. Discrete Algor.*, **2**, 53–86.

Briggs,A.W. *et al.* (2007) Patterns of damage in genomic DNA sequences from a Neandertal. *Proc. Natl Acad. Sci. USA*, **104**, 14616–14621.

Cedar,H. and Bergman,Y. (2009) Linking DNA methylation and histone modification: patterns and paradigms. *Nat. Rev. Genet.*, **10**, 295–304.

Chen,P.Y. *et al.* (2010) BS Seeker: precise mapping for bisulfite sequencing. *BMC Bioinformatics*, **11**, 203.

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

Dehal,P. *et al.* (2002) The draft genome of ciona intestinalis: insights into chordate and vertebrate origins. *Science*, **298**, 2157–2167.

Esteller,M. (ed.) (2005) *DNA Methylation: Approaches, Methods, and Applications*. CRC Press, Boca Raton, FL, USA.

Esteller,M. (2007) Cancer epigenomics: DNA methylomes and histone-modification maps. *Nat. Rev. Genet.*, **8**, 286–298.

Frommer,M. *et al.* (1992) A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl Acad. Sci. USA*, **89**, 1827–1831.

Gehring,M. *et al.* (2009) Extensive demethylation of repetitive elements during seed development underlies gene imprinting. *Science*, **324**, 1447–1451.

Hansen,K.D. *et al.* (2011) Increased methylation variation in epigenetic domains across cancer types. *Nat. Genet.*, **43**, 768–775.

Hoffmann,S. *et al.* (2009) Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Comput. Biol.*, **5**, e1000502.

Huang,Y. *et al.* (2010) The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. *PLoS One*, **5**, e8888.

Ko,P. and Aluru,S. (2003) Space efficient linear time construction of suffix arrays. In *Combinatorial Pattern Matching (CPM 03)*, pp. 200–210. Springer.

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

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

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.* (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, **462**, 315–322.

Lister,R. *et al.* (2011) Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature*, **471**, 68–73.

Li,H. *et al.* (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, **18**, 1851–1858.

Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAM tools. *Bioinformatics*, **25**, 2078–2079.

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

Lyko,F. *et al.* (2010) The honey bee epigenomes: differential methylation of brain DNA in queens and workers. *PLoS Biol.*, **8**, e1000506.

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

Mills,R.E. *et al.* (2011) Natural genetic variation caused by small insertions and deletions in the human genome. *Genome Res.*, **21**, 830–839.

Myers,G. (1999) A fast bit-vector algorithm for approximate string matching based on dynamic programming. *J. ACM*, **46**, 395–415.

Prüfer,K. *et al.* (2010) Computational challenges in the analysis of ancient DNA. *Genome Biol.*, **11**, R47.

Putnam,N.H. *et al.* (2008) The amphioxus genome and the evolution of the chordate karyotype. *Nature*, **453**, 1064–1071.

Reik,W. (2007) Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature*, **447**, 425–432.

Serre,D. *et al.* (2010) MBD-isolated genome sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome. *Nucleic Acids Res.*, **38**, 391–399.

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

Watanabe,Y. and Maekawa,M. (2010) Methylation of DNA in cancer. *Adv. Clin. Chem.*, **52**, 145–167.

Weber,M. *et al.* (2005) Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nature Genet.*, **37**, 853–862.

Weber,M. and Schübeler,D. (2007) Genomic patterns of DNA methylation: targets and function of an epigenetic mark. *Curr. Opin. Cell. Biol.*, **19**, 273–280.

Xiang,H. *et al.* (2010) Single base-resolution methylome of the silkworm reveals a sparse epigenomic map. *Nat. Biotechnol*, **28**, 516–520.

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

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016