Published online 11 May 2011 *Nucleic Acids Research, 2011, Vol. 39, Web Server issue* ***W551–W556***
doi:10.1093/nar/gkr312

BiQ Analyzer HT: locus-specific analysis of DNA methylation by high-throughput bisulfite sequencing

Pavlo Lutsik[1,2], Lars Feuerbach[2], Julia Arand[1], Thomas Lengauer[2], Jörn Walter[1,*] and Christoph Bock[2,*]

[1]Department of Genetics/Epigenetics, Saarland University and [2]Max Planck Institute for Informatics, 66123 Saarbrücken, Germany

Received February 27, 2011; Revised April 12, 2011; Accepted April 18, 2011

ABSTRACT

Bisulfite sequencing is a widely used method for measuring DNA methylation in eukaryotic genomes. The assay provides single-base pair resolution and, given sufficient sequencing depth, its quantitative accuracy is excellent. High-throughput sequencing of bisulfite-converted DNA can be applied either genome wide or targeted to a defined set of genomic loci (e.g. using locus-specific PCR primers or DNA capture probes). Here, we describe BiQ Analyzer HT (http://biq-analyzer-ht.bioinf.mpi-inf.mpg.de/), a user-friendly software tool that supports locus-specific analysis and visualization of high-throughput bisulfite sequencing data. The software facilitates the shift from time-consuming clonal bisulfite sequencing to the more quantitative and cost-efficient use of high-throughput sequencing for studying locus-specific DNA methylation patterns. In addition, it is useful for locus-specific visualization of genome-wide bisulfite sequencing data.

INTRODUCTION

DNA methylation is a widely studied epigenetic modification. It is present in all vertebrates and many invertebrate animals as well as in plants (1). In mammals, DNA methylation plays an important role for developmental gene regulation and for germline repression of repetitive elements (2). Aberrant DNA methylation patterns are frequently observed in cancer (3) and may also occur in many other human diseases (4). The link between locus-specific DNA methylation alterations and common diseases has created significant interest in using these epigenetic alterations as biomarkers in drug discovery and clinical diagnostics (5).

To investigate the many roles of DNA methylation in development and disease, researchers depend on experimental methods that accurately measure DNA methylation patterns at high accuracy and affordable cost. Many technologies with different advantages and disadvantages have been developed over the last 20 years, but only bisulfite-based methods provide quantitative DNA methylation data at single-base pair resolution (6). In bisulfite sequencing, the DNA is treated with sodium bisulfite, which selectively converts unmethylated cytosines into uracils but leaves methylated cytosines untouched (7). Hydroxymethylated DNA, which has recently been detected in some mammalian cell types, is also left unconverted and is indistinguishable from methylated DNA using bisulfite-based methods (8).

Bisulfite sequencing has recently been used to obtain the first genome wide, high-resolution maps of DNA methylation in the human genome (9,10). Bisulfite-based methods also performed well in a benchmarking study of DNA methylation mapping technologies (11). Along with technologies for DNA methylation mapping at a genomic scale, locus-specific bisulfite sequencing plays an important role as gold-standard validation method and promises to become a standard technology in clinical diagnostics (12).

Locus-specific bisulfite sequencing has traditionally been performed by Sanger sequencing of a few dozen hand-picked DNA clones, making this method rather time-consuming and costly. To address these limitations, researchers increasingly use high-throughput sequencing instead of Sanger sequencing (13–15), which has three major advantages: (i) due to the increased sequencing throughput, it becomes feasible to obtain highly quantitative DNA methylation patterns for the loci of interest. This is particularly relevant for studying heterogeneous tissue samples and for clinical diagnostics; (ii) due to

*To whom correspondence should be addressed. Tel: +1 765 247 2625; Fax: +49 681 9325 399; Email: cbock@mpi-inf.mpg.de
Correspondence may also be addressed to Jörn Walter. Tel: +49 681 302 2425; Fax: +49 681 302 2703; Email: j.walter@mx.uni-saarland.de

lower per-base costs and the use of multiplexing to sequence many samples and/or loci in a single machine run, the sequencing costs are substantially reduced; and (iii) the cloning step for isolating DNA populations that carry the DNA sequence of a single DNA molecule becomes obsolete because current methods for high-throughput sequencing measure the sequences of individual DNA clones.

A major roadblock for the wider use of high-throughput bisulfite sequencing is the lack of software tools for processing and analyzing the large number of sequencing reads that are generated by this method. Several software tools have been developed for processing small-scale bisulfite sequencing data obtained by conventional Sanger sequencing. The BiQ Analyzer (16) software from our group has recently been updated to version 2.0 and continues to be a useful tool for interactive analysis of small-scale bisulfite sequencing data. Alternative tools include the QUMA web service (17), BISMA (18) and several more specialized programs (19–22). None of these tools can be scaled to the read numbers that are typically obtained by high-throughput sequencing. For this reason, recent studies utilized custom data analysis scripts, none of which are conveniently available (13–15).

Here, we describe BiQ Analyzer HT, a comprehensive software tool for locus-specific analysis of high-throughput bisulfite sequencing data. BiQ Analyzer HT builds on concepts that we originally developed for the popular BiQ Analyzer software (16), but it was redesigned and rewritten to meet the challenges arising for the analysis of high-throughput bisulfite sequencing data. All functionality of BiQ Analyzer HT is available through a web-startable graphical user interface, which guides the user through the data analysis (Figure 1). As an additional option, it is possible to run the computationally intensive parts of the software on a remote high-performance computer while maintaining the user-friendliness of a graphical interface run locally. Finally, BiQ Analyzer HT provides an optional command-line interface to facilitate integration into automatic data analysis pipelines.

PROGRAM OVERVIEW

BiQ Analyzer HT facilitates locus-specific analysis, quality control and visualization of high-throughput bisulfite sequencing data. The tool takes sequencing read data as input, and it produces quality-controlled output tables and diagrams of the inferred DNA methylation information for each sample, locus and DNA methylation site.

BiQ Analyzer HT is a Java-based program which can be run on any computer which has a recent version of the Java Virtual Machine installed. The tool is available as a self-installing Java Web Start distribution, and as a downloadable installation package for computers that are not connected to the Internet. BiQ Analyzer HT's project-based user interface supports the interactive analysis of bisulfite sequencing data for multiple target loci in multiple samples. A typical analysis consists of three phases: (i) data import; (ii) sequence alignment and quality control; and (iii) visualization and export of the inferred DNA methylation information (Figure 1).

To prepare high-throughput sequencing data for analysis with BiQ Analyzer HT, the user first applies vendor-specific software to perform base-calling, to resolve any sample multiplexing and to convert the data into one of two standard formats, FASTA or BAM. When importing FASTA files obtained by locus-specific bisulfite sequencing, BiQ Analyzer HT expects one file per sample and locus. We currently provide a custom script that automatizes data preparation for the Roche 454 sequencing platform (http://biq-analyzer-ht.bioinf.mpi-inf.mpg.de), and we will add similar scripts for other platforms based on user demand. Alternatively, genome-scale bisulfite sequencing data can be imported as BAM files, which are most conveniently generated with BSMAP (23).

When a new BiQ Analyzer HT project is initialized, an output directory is created into which the software writes its analysis results (Table 1). The user specifies the project structure by adding samples and by loading FASTA files that define the genomic reference sequence of each locus. The resulting tree structure is shown in BiQ Analyzer HT's main window. Once the data are loaded, this tree can be ordered either by samples or by loci, depending on the biological question of interest.

Read alignment and inference of DNA methylation information are controlled by parameters that the user selects on the setup screen. While the default values often provide good results, it is recommended that the user runs a first analysis with default parameters, inspects the results and then adjusts the parameters as necessary. Data set-specific choice of quality control parameters can sometimes compensate for quality issues that may be present in the primary data. For example, a decrease in alignment stringency parameters allows for retaining reads with reduced similarity to the reference, which would be removed by the default filtering criteria. This can be essential to process highly polymorphic sequences such as retrotransposable elements and DNA repeats.

Once satisfactory results are obtained, the inferred DNA methylation data can be exported in several formats, including sequence alignments, data tables and DNA methylation plots. Table 1 summarizes all output items. The sequence alignments provide a detailed account of how the DNA methylation levels were inferred. In addition, they can be used to identify allele-specific polymorphisms or evidence of structural variation in the sequence data. The data tables facilitate exploratory data analysis using spreadsheets, in-depth statistics using statistical software such as R/Bioconductor (24) and epigenetic biomarker development using BiQ Analyzer's companion tool MethMarker (25). Finally, the DNA methylation plots visualize the results of BiQ Analyzer HT analyses for use in papers and scientific reports.

The visualization module of BiQ Analyzer HT utilizes the publicly available GSEA library (26) for plotting DNA methylation heatmaps. BAM file handling is implemented using the Picard library (http://picard.sourceforge.net), and parts of the sequence processing code are based on the BioJava framework (27).

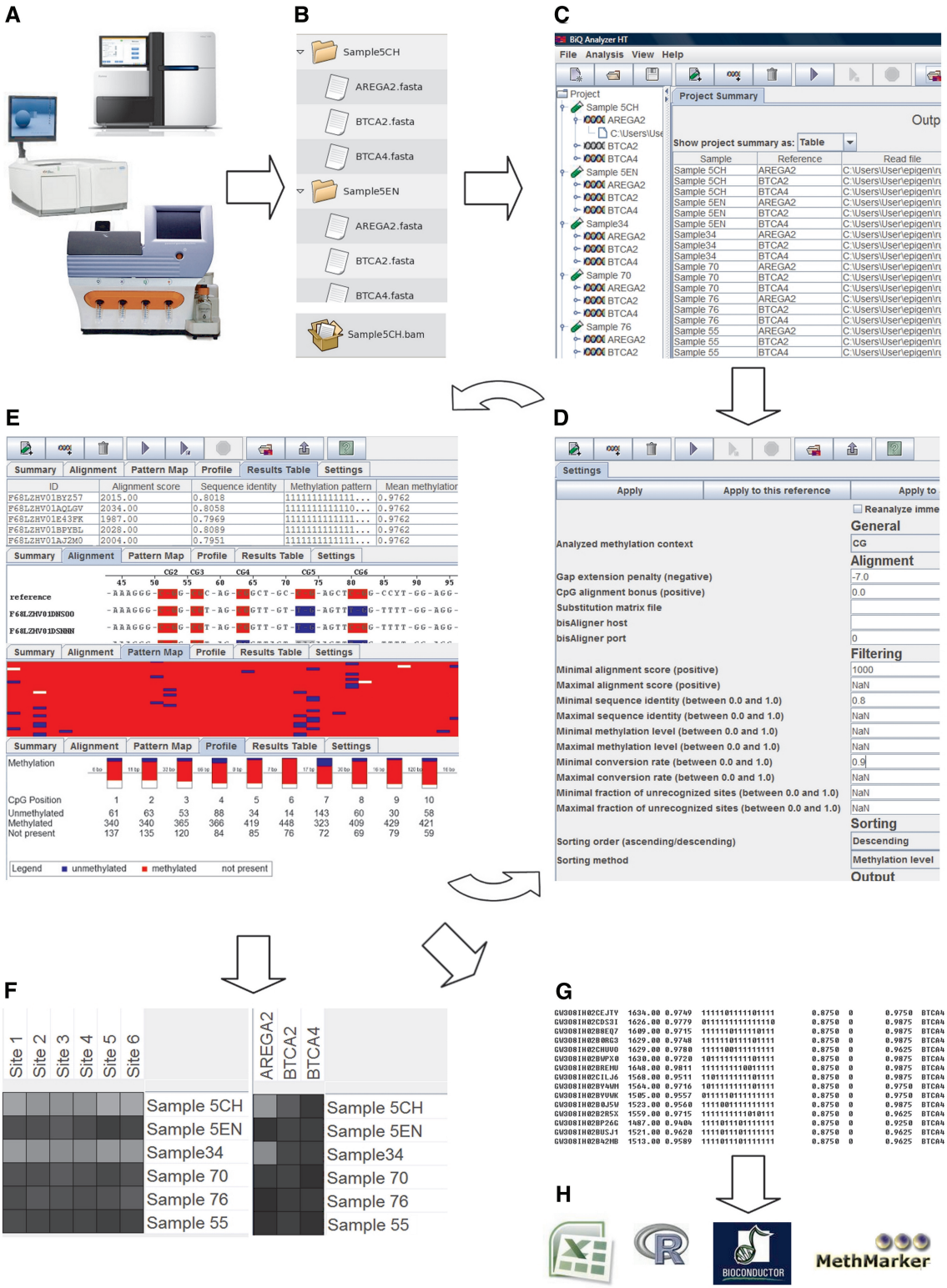

Figure 1. BiQ Analyzer HT workflow. Bisulfite sequencing data are generated either for the entire genome or selectively for a defined set of genomic loci using commercially available high-throughput sequencers (**A**). To reduce sequencing cost, bisulfite-converted DNA from several samples and/or loci is typically barcoded and combined into a single sequencing run. The multiplexed read data are separated and converted into FASTA or BAM files using vendor-provided software and/or custom scripts (**B**), before they are loaded into BiQ Analyzer HT (**C**). Once the data have been loaded, the user sets alignment and quality control parameters (**D**), runs the analysis and inspects the inferred DNA methylation data (**E**) and adjusts the parameters until satisfactory results are obtained. Finally, the DNA methylation measurements can be visualized graphically (**F**) and exported as tab-separated tables (**G**) for in-depth analysis using spreadsheets such as Excel, statistical software such as R/Bioconductor and biomarker development tools such as MethMarker (**H**).

Table 1. Analysis results generated by BiQ Analyzer HT

Category	Title	Access	Format	Description
Tabular	Project summary	GUI	TSV	Basic information summarizing the analysis
	Sample summary	GUI	TSV	DNA methylation summary for each locus in each sample
	Results table	OD	TSV	Alignment quality, estimated bisulfite conversion rate and DNA methylation summary for each sequencing read
	Methylation pattern table	GUI	TSV	DNA methylation patterns for each sequencing read. Columns correspond to DNA methylation sites (typically CpG positions)
	Project results table	GUI	TSV	Combined results table for all samples and loci
Graphical	Methylation pattern map	OD, GUI	PNG	Heatmap-style representation of DNA methylation patterns for each sequencing read. Columns correspond to DNA methylation sites
	Methylation profile	OD, GUI	PNG, SVG	Diagram visualizing the frequency of methylated, unmethylated and missing-value observations for each DNA methylation site
	Project methylation heatmap	GUI[a]	PNG	Heatmap of mean DNA methylation levels for each locus in each sample
	Methylation profile heatmap	GUI[a]	PNG	Heatmap of mean DNA methylation levels for each DNA methylation site at a specific locus
Sequence	Alignment	OD	FASTA	Multiple alignment of sequencing reads for each locus in each sample
	Filtered reads	OD	FASTA	Sequences of all reads that passed quality filtering

[a]The data table from which the heatmap is generated can also be exported for follow-up analysis.
OD ('output directory')–the item is written to the project output directory tree; GUI ('graphical user interface')–the item can be exported via 'Save as...' or 'Copy to clipboard' in the corresponding context menu; FASTA–sequencing reads in multiple-sequence text format; PNG–images in Portable Network Graphics format; SVG–images in Scalable Vector Graphics format; TSV–tab-separated value tables.

DATA PROCESSING

BiQ Analyzer HT implements a data processing pipeline that is run for each combination of locus and sample in the project tree. The pipeline aligns all sequencing reads from the corresponding input file to the locus-specific genomic reference sequence, and based on these alignments it infers which cytosines are methylated or unmethylated by comparing the read sequence with the reference sequence. The key steps of the data processing pipeline are outlined in more detail below. All analyses are conveniently accessible via the graphical interface. They can also be run from the command line, which facilitates integration with automatic data processing pipelines.

Read alignment

The analysis of bisulfite sequencing data crucially depends on accurate alignments. This is an inherently difficult task when complex genomic regions with repetitive elements and structural variation are studied and further complicated by the fact that bisulfite-converted DNA has substantially lower information content than genomic DNA. For this reason, speed-optimized seed-based aligners such as BLAT (28), MAQ (29) and BWA (30)—which are commonly used for aligning high-throughput sequencing data—could undermine the accuracy of BiQ Analyzer HT. After exploring several alternatives, we chose to use the Needleman–Wunsch algorithm (31), which is guaranteed to find the optimal (although not necessarily the correct) alignment between each sequencing read and the reference sequence. Furthermore, we made several modifications to the algorithm that account for recurrent issues with bisulfite-converted DNA (Supplementary Text S1). To partially compensate for the fact that the Needleman–Wunsch algorithm is substantially slower than current short-read aligners, we use a highly optimized implementation of this algorithm. This implementation provides excellent performance for read numbers in the order of 10^4 per locus on a standard laptop computer (Table 2). Furthermore, the read alignment can be outsourced to a remote high-performance computer, which makes it feasible to process in the order of one million reads per locus on a standard laptop computer.

Quality control and read filtering

Based on the pairwise alignment of the sequencing reads with their corresponding genomic reference sequence, the data quality of the bisulfite sequencing experiment is estimated. Basic quality measures include the alignment score and sequence identity with the bisulfite-converted reference sequence, the estimated bisulfite conversion rate (fraction of unconverted cytosines outside of the analyzed methylation context, e.g. 'CG') and the number of DNA methylation sites with missing data. The sequencing read data can be filtered for each of these quality measures in order to quickly discard low quality or otherwise unsuitable reads. The threshold values of each quality measure are set to empirically chosen defaults, but users may need to adjust these parameters interactively to account for the characteristics of their specific data sets.

Inference of DNA methylation patterns. BiQ Analyzer HT's default settings focus on CpG methylation which is the most common modification of eukaryotic DNA. The user can also choose to include other symmetric and asymmetric methylation contexts in the analysis, such as CpHpG and CpHpH. A methylation context is defined

Table 2. Performance comparison of software packages for locus-specific analysis of bisulfite sequencing data

Region	Read count	Performance							
		BiQ Analyzer 2.0		QUMA[a]		BiQ Analyzer HT		BiQ Analyzer HT[b]	
		Memory	ET	Memory	ET	Memory	ET	Memory	ET
RE1	400	350	300	NA[c]	10	95	30	1000	6
RE2	1054	500	911	NA[c]	25	200	50	1000	9
RE3	3150	>1000	3455	NA[c]	70	200	95	1000	16
RE3[d]	10 000	NA[e]		NA[c]	323	300	285	1500	50
RE3[d]	100 000	NA[e]		NA[f]		NA[g]		3500	440
RE3[d]	1 000 000	NA[e]		NA[f]		NA[g]		10 000	1940

All tests, except for the cases noted explicitly, were run on a standard laptop with dual-core processor and 2 GB main memory. The values of peak memory usage are given in MB. ET ('execution time') denotes the total duration of the analysis in seconds.
[a]The QUMA web-server running on a high-performance machine (8 dual-core processors, 16 GB main memory).
[b]BiQ Analyzer HT running on a high-performance machine (8 dual-core processors, 16 GB main memory).
[c]Memory usage of the web-server does not affect performance for the end user.
[d]The data set was obtained by concatenating multiple copies of the initial set of reads obtained for RE3.
[e]The calculation could not be finished due to an error.
[f]The calculation failed because it exceeded the web-server's maximum read threshold.
[g]Tests with BiQ Analyzer HT for the last two read sets were performed only on the high-performance computer.

by a pair of DNA sequence motifs, one of which matching the methylated and the other matching the unmethylated state. The positions of potential methylation sites are detected by scanning the reference sequence for matches of the methylated motif. Next, the methylation state is determined by comparing the read and reference sequences at each potential methylation site, and the site recorded as methylated, unmethylated or missing value ('1′, '0′ and 'x', respectively). The collection of DNA methylation states for all sites in a given sequencing read constitute its methylation pattern, and the number of methylated sites divided by the total number of sites that are not missing values defines the mean methylation of a sequencing read.

Data visualization and export

The inferred DNA methylation data and quality control information can be exported for documentation and follow-up analysis using statistical tools (Table 1). The resulting tables list the quality measures, DNA methylation patterns and mean methylation levels for each sequencing read that has not been filtered out during quality control. Prior to exporting these tables, they can be sorted by one of the quality measures or by the inferred DNA methylation information.

PERFORMANCE EVALUATION

To confirm the practical utility of BiQ Analyzer HT for large data sets and to assess its performance relative to existing low-throughput tools, we benchmarked the tools on data sets with up to one million reads mapping to a single locus (Table 2). These data set were obtained by multiplexed locus-specific bisulfite sequencing on the Roche 454 sequencing platform. Briefly, three classes of repetitive elements (RE1, RE2 and RE3) were amplified from bisulfite-treated mouse DNA, and several thousand reads were sequenced for these repetitive elements. To evaluate BiQ Analyzer HT's performance for higher read numbers, we further constructed artificial test sets from the actual data set of region RE3 by reusing sequencing reads multiple times. The results of this benchmarking shows that all existing tools have severe limitations in the number of reads that can be processed (Table 2). In contrast, with BiQ Analyzer HT, we could successfully analyze a data set with one million reads mapping to a single locus.

CONCLUSIONS

BiQ Analyzer HT provides comprehensive support for locus-specific analysis, quality control and visualization of high-throughput bisulfite sequencing data. It addresses the bioinformatic challenges of using high-throughput sequencing as a fast and cost-efficient alternative to clonal bisulfite sequencing, and it is fully compatible with multiplex analysis of several loci and samples. The alignment algorithm was specifically optimized for bisulfite-converted sequences, and it supports the analysis of both CpG and non-CpG methylation patterns. In summary, the combination of locus-specific high-throughput sequencing and interactive data analysis with BiQ Analyzer HT provides a highly practical approach for measuring the DNA methylation patterns of 10–100′s of loci in 100–1000′s of samples, for example, in the context of biomarker validation and clinical diagnostics.

AVAILABILITY

http://biq-analyzer-ht.bioinf.mpi-inf.mpg.de (This website/software is free and open to all users and there is no login requirement).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Dr Sascha Tierling and Dirk Schuemacher for helpful discussions and the provision of test data, Yassen Assenov for advice with Java programming and Fabian Müller for advice on the BAM format.

FUNDING

CANCERDIP project (HEALTH-F2-2007-200620); ColoNet project (BMBF 0315417-D). Funding for open access charge: Max Planck Institute for Informatics and Saarland University.

Conflict of interest statement. None declared.

REFERENCES

1. Suzuki,M.M. and Bird,A. (2008) DNA methylation landscapes: provocative insights from epigenomics. *Nat. Rev. Genet.*, **9**, 465–476.
2. Bird,A. (2002) DNA methylation patterns and epigenetic memory. *Genes Dev.*, **16**, 6–21.
3. Esteller,M. (2008) Epigenetics in cancer. *N. Engl. J. Med.*, **358**, 1148–1159.
4. Feinberg,A.P. (2007) Phenotypic plasticity and the epigenetics of human disease. *Nature*, **447**, 433–440.
5. Laird,P.W. (2003) The power and the promise of DNA methylation markers. *Nat. Rev. Cancer*, **3**, 253–266.
6. Laird,P.W. (2010) Principles and challenges of genome-wide DNA methylation analysis. *Nat. Rev. Genet.*, **11**, 191–203.
7. Frommer,M., McDonald,L.E., Millar,D.S., Collis,C.M., Watt,F., Grigg,G.W., Molloy,P.L. and Paul,C.L. (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.
8. Huang,Y., Pastor,W.A., Shen,Y., Tahiliani,M., Liu,D.R. and Rao,A. (2010) The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. *PLoS ONE*, **5**, e8888.
9. Lister,R., Pelizzola,M., Dowen,R.H., Hawkins,R.D., Hon,G., Tonti-Filippini,J., Nery,J.R., Lee,L., Ye,Z., Ngo,Q.M. *et al.* (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, **462**, 315–322.
10. Laurent,L., Wong,E., Li,G., Huynh,T., Tsirigos,A., Ong,C.T., Low,H.M., Kin Sung,K.W., Rigoutsos,I., Loring,J. *et al.* (2010) Dynamic changes in the human methylome during differentiation. *Genome Res.*, **20**, 320–331.
11. Bock,C., Tomazou,E.M., Brinkman,A.B., Müller,F., Simmer,F., Gu,H., Jäger,N., Gnirke,A., Stunnenberg,H.G. and Meissner,A. (2010) Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nat. Biotechnol.*, **28**, 1106–1114.
12. Bock,C. (2009) Epigenetic biomarker development. *Epigenomics*, **1**, 99–110.
13. Korshunova,Y., Maloney,R.K., Lakey,N., Citek,R.W., Bacher,B., Budiman,A., Ordway,J.M., McCombie,W.R., Leon,J., Jeddeloh,J.A. *et al.* (2008) Massively parallel bisulphite pyrosequencing reveals the molecular complexity of breast cancer-associated cytosine-methylation patterns obtained from tissue and serum DNA. *Genome Res.*, **18**, 19–29.
14. Taylor,K.H., Kramer,R.S., Davis,J.W., Guo,J., Duff,D.J., Xu,D., Caldwell,C.W. and Shi,H. (2007) Ultradeep bisulfite sequencing analysis of DNA methylation patterns in multiple gene promoters by 454 sequencing. *Cancer Res.*, **67**, 8511–8518.
15. Varley,K.E. and Mitra,R.D. (2010) Bisulfite Patch PCR enables multiplexed sequencing of promoter methylation across cancer samples. *Genome Res.*, **20**, 1279–1287.
16. Bock,C., Reither,S., Mikeska,T., Paulsen,M., Walter,J. and Lengauer,T. (2005) BiQ Analyzer: visualization and quality control for DNA methylation data from bisulfite sequencing. *Bioinformatics*, **21**, 4067–4068.
17. Kumaki,Y., Oda,M. and Okano,M. (2008) QUMA: quantification tool for methylation analysis. *Nucleic Acids Res.*, **36**, W170–W175.
18. Rohde,C., Zhang,Y., Reinhardt,R. and Jeltsch,A. (2010) BISMA–fast and accurate bisulfite sequencing data analysis of individual clones from unique and repetitive sequences. *BMC Bioinformatics*, **11**, 230.
19. Carr,I.M., Valleley,E.M., Cordery,S.F., Markham,A.F. and Bonthron,D.T. (2007) Sequence analysis and editing for bisulphite genomic sequencing projects. *Nucleic Acids Res.*, **35**, e79.
20. Grunau,C., Schattevoy,R., Mache,N. and Rosenthal,A. (2000) MethTools–a toolbox to visualize and analyze DNA methylation data. *Nucleic Acids Res.*, **28**, 1053–1058.
21. Hetzl,J., Foerster,A.M., Raidl,G. and Mittelsten Scheid,O. (2007) CyMATE: a new tool for methylation analysis of plant genomic DNA after bisulphite sequencing. *Plant J.*, **51**, 526–536.
22. Xu,Y.H., Manoharan,H.T. and Pitot,H.C. (2007) CpG PatternFinder: a Windows-based utility program for easy and rapid identification of the CpG methylation status of DNA. *Biotechniques*, **43**, 334336–340, 342.
23. Xi,Y. and Li,W. (2009) BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics*, **10**, 232.
24. Gentleman,R.C., Carey,V.J., Bates,D.M., Bolstad,B., Dettling,M., Dudoit,S., Ellis,B., Gautier,L., Ge,Y., Gentry,J. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.
25. Schüffler,P., Mikeska,T., Waha,A., Lengauer,T. and Bock,C. (2009) MethMarker: user-friendly design and optimization of gene-specific DNA methylation assays. *Genome Biol.*, **10**, R105.
26. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.
27. Holland,R.C., Down,T.A., Pocock,M., Prlić,A., Huen,D., James,K., Foisy,S., Dräger,A., Yates,A., Heuer,M. *et al.* (2008) BioJava: an open-source framework for bioinformatics. *Bioinformatics*, **24**, 2096–2097.
28. Kent,W.J. (2002) BLAT–the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
29. Li,H., Ruan,J. and Durbin,R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, **18**, 1851–1858.
30. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
31. Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.