Published online 22 February 2008

Nucleic Acids Research, 2008, Vol. 36, No. 5 ***e34***
doi:10.1093/nar/gkn083

Bisulfite sequencing Data Presentation and Compilation (BDPC) web server—a useful tool for DNA methylation analysis

Christian Rohde[1], Yingying Zhang[1], Tomasz P. Jurkowski[1], Heinrich Stamerjohanns[1], Richard Reinhardt[2] and Albert Jeltsch[1,*]

[1]School of Engineering and Science, Jacobs University Bremen, Campus Ring 1, 28725 Bremen and [2]Max Planck Institute for Molecular Genetics, Ihnestrasse 63-73, D-14195 Berlin-Dahlem, Germany

Received December 17, 2007; Revised February 7, 2008; Accepted February 8, 2008

ABSTRACT

During bisulfite genomic sequencing projects large amount of data are generated. The Bisulfite sequencing Data Presentation and Compilation (BDPC) web interface (http://biochem.jacobs-university.de/BDPC/) automatically analyzes bisulfite datasets prepared using the BiQ Analyzer. BDPC provides the following output: (i) MS-Excel compatible files compiling for each PCR product (a) the average methylation level, the number of clones analyzed and the percentage of CG sites analyzed (which is an indicator of data quality), (b) the methylation level observed at each CG site and (c) the methylation level of each clone. (ii) A methylation overview table compiling the methylation of all amplicons in all tissues. (iii) Publication grade figures in PNG format showing the methylation pattern for each PCR product embedded in an HMTL file summarizing the methylation data, the DNA sequence and some basic statistics. (iv) A summary file compiling the methylation pattern of different tissues, which is linked to the individual HTML result files, and can be directly used for presentation of the data in the internet. (v) A condensed file, containing all primary data in simplified format for further downstream data analysis and (vi) a custom track file for display of the results in the UCSC genome browser.

INTRODUCTION

DNA methylation encodes additional information on the DNA in the form of methyl groups covalently attached to the C5 position of cytosine. In mammals, the methylation of cytosines occurs at cytosine–guanine dinucleotides (CG-sites) in a cell type and tissue-specific pattern (1,2). Methylation of the DNA works in concert with other epigenetic marks like histone modification (3) in the regulation of gene expression and development. In general, DNA methylation of gene promoters silences gene activity (2). It is involved in X-chromosome inactivation, genomic integrity, cell development and differentiation (4). DNA methylation is essential for mammalian development, because in mice deletion of any of the known active DNA methyltransferase leads to embryonic lethality or developmental abnormalities and early death (5,6). Abnormal DNA methylation is linked to cancer development and other diseases (7,8).

After complete sequencing of the human genome, the decoding of the epigenome, which contains the blueprint for the activity of genetic elements has come into the focus of research attention. Recently, human (9) and the *Arabidopsis* (3) DNA methylation was studied genome wide using immunoprecipitation of DNA by antibodies directed against methylcytosine in combination with array technology. In addition, DNA methylation in human chromosomes 6, 20 and 22 gene promoter regions was studied in detail by bisulfite genomic sequencing (10), which is the standard method for analysis of DNA methylation, because it provides a reliable and detailed picture of the methylation state of the DNA. Specific techniques are needed to study DNA methylation, because the information about the methylation status of DNA is lost during *in vitro* DNA amplification and cloning. After sodium bisulfite treatment, methylated and unmethylated cytosines can be discriminated, because unmethylated cytosines are converted into uracil whereas methylated cytosines remain as cytosine (11–13). The workflow of a bisulfite sequencing experiment consists of the design of primers specific for converted DNA that usually do not contain CG-sites, and

*To whom correspondence should be addressed. Tel: +49 421 200 3247; Fax: +49 421 200 3249; Email: a.jeltsch@jacobs-university.de

bisulfite conversion of the DNA followed by PCR. The analysis of individual DNA molecules by sequencing of sub-cloned PCR products provides the most reliable and detailed information about the methylation state of each CG-site. Therefore, in most cases the next steps are subcloning of the PCR products and sequencing of several individual clones to generate a statistically significant data set (the workflow is described in detail in 14).

For analysis of sequencing results from bisulfite converted DNA, the experimental sequences are aligned to the *in silico* converted genomic target sequence. This step is facilitated by the BiQ Analyzer software (15), which is a very popular program for initial analysis of bisulfite sequencing results. BiQ analyzer uploads the target sequence and an arbitrary number of sequences of subcloned PCR products, creates the alignment, guides the user through each step of the analysis and stores the alignment and the results in an HTML file. However, the BiQ Analyzer only works for individual PCR products and creates one separate output file for each PCR product such that it cannot assist further data analysis and compilation.

Medium or large-scale projects accumulate large amounts of data, because DNA methylation is usually studied at many genomic sites in several biological samples. In this article, we assume that DNA methylation is analyzed in different tissues or biological samples, which we abbreviate here as 'tissue'. Different sets of primers are used for analysis of the methylation of different genomic regions called 'amplicons'. The same amplicon can be studied in different tissues. Thereby, different PCR products of the same amplicon type are generated. Consequently, methylation data of a large number of PCR products have to be studied and the results integrated, analyzed and presented. Often, the methylation data for PCR products of the same amplicon need to be compared to detect methylation differences between tissues. Furthermore, general statistics comparing results of all PCR products are required and, for data presentation, publication grade representations of the methylation pattern together with linked and publication ready HTML files for data presentation in the internet are needed.

RESULTS AND DISCUSSION

The Bisulfite sequencing Data Presentation and Compilation (BDPC) web interface supports the compilation, analysis and presentation of bisulfite DNA methylation data of any degree of complexity ranging from small-scale exploratory data sets to large-scale methylome projects (http://biochem.jacobs-university.de/BDPC/). The program supports data presentation by preparing publication grade figures showing the methylation pattern of each PCR product and preparing a set of linked HTML files for internet data presentation. It provides a data summary for each PCR product, different result compilation files and supports further in depth analysis by preparing a condensed output file, which contains all primary data.

The BDPC software is written in PHP. The source code of the program is available on request. It is designed to analyze and compile BiQ Analyzer result files, which provides the results in the form of separate HTML output files, one for each PCR product. Data files prepared manually or with other software can be compiled using BDPC as well, if the data are provided in the BDPC compatible format as explained in Figure 1. For uploading, the data files have to be named with the amplicon name and stored in one folder for each tissue, respectively. The folder name is used by BDPC and interpreted as the tissue designation. Optionally, the user may provide the chromosomal locations and coordinates of the amplicons as additional

```
1  <pre>
2  Genomic Sequence with numbered CpG dinucleotides:
3  Use these line for numbering of the CG-sites or keep this line empty
4  Use these line for numbering of the CG-sites or keep this line empty
5  Put Sequence analyzed here here or keep this line empty
6  Keep this line empty
7  Methylation data for these CpG dinucleotides (1=methylated, 0=unmethylated, x=unknown):
8  [2]clone_1    x    0    0    0    0    0    0    0    0    0    0
9  []clone_02    0    1    1    1    1    1    0    1    1    1    0
10 []clone_03    0    0    0    1    0    0    0    1    0    0    0
11 []clone_04    0    0    0    1    0    0    0    1    0    0    0
12 []clone_05    0    0    0    1    0    0    0    1    0    0    0
13 []clone_06    0    0    0    1    0    0    0    1    1    1    0
14 []clone_07    0    0    0    1    0    0    0    1    0    0    0
15 []clone_08    0    0    0    1    0    0    0    1    0    0    0
16 []clone_09    0    0    0    1    0    0    0    1    0    0    0
17 []clone_10    0    0    0    1    0    0    0    1    0    0    0
18 []clone_11    0    0    0    1    0    0    0    1    0    0    0
19 []clone_12    0    0    0    1    0    0    0    1    1    1    0
20 []clone_13    x    0    0    0    0    0    0    0    0    0    0
21 </pre>
```

Figure 1. BDPC compatible data format. For uploading to BDPC, a data file needs the information shown in bold. The example file can be downloaded from the BDPC website in the 'Example files' area. For presentation purpose here line numbering is used, which must not be done in the data files. The phrases in lines 2 and 7 are mandatory and must be written exactly as shown here. The lines 3 and 4 can be used for numbering the CG-sites, line 5 to give the sequence analyzed. This information will be carried over into the BDPC output files. In line 8 the '[2]' is mandatory, whereas for the following lines two squared brackets are sufficient. From line 8 on, the results are organized in such a way that each column represents a CG-site and each line the sequencing result of an individual clone. In the results, the '1' represents a methylated CG-site, the '0' an unmethylated CG-site and the 'x' a CG-site which is not present. The data is separated with tabs. The HTML tags in line 1 and 21 are not required.

information, which then will be used by BDPC to generate a UCSC custom track file. The format of this file is described in the BDPC online manual. The folders (together with the additional information file) need to be compressed into one single file using the ZIP standard format (http://www.pkware.com/) and uploaded.

Figure 2 gives an overview of the role of BDPC in the workflow of a bisulfite DNA methylation study and the result files provided after analysis. BDPC can compile datasets of large sizes. Internally, we have processed data sets comprising more than 1000 result files without difficulties. Here, as an example, we illustrate the application of BDPC on a relatively small dataset comprising five independent amplicons, which cover the transcriptional start site of the human FAM3B gene as shown in Figure 3. These amplicons were analyzed in the DNA of four

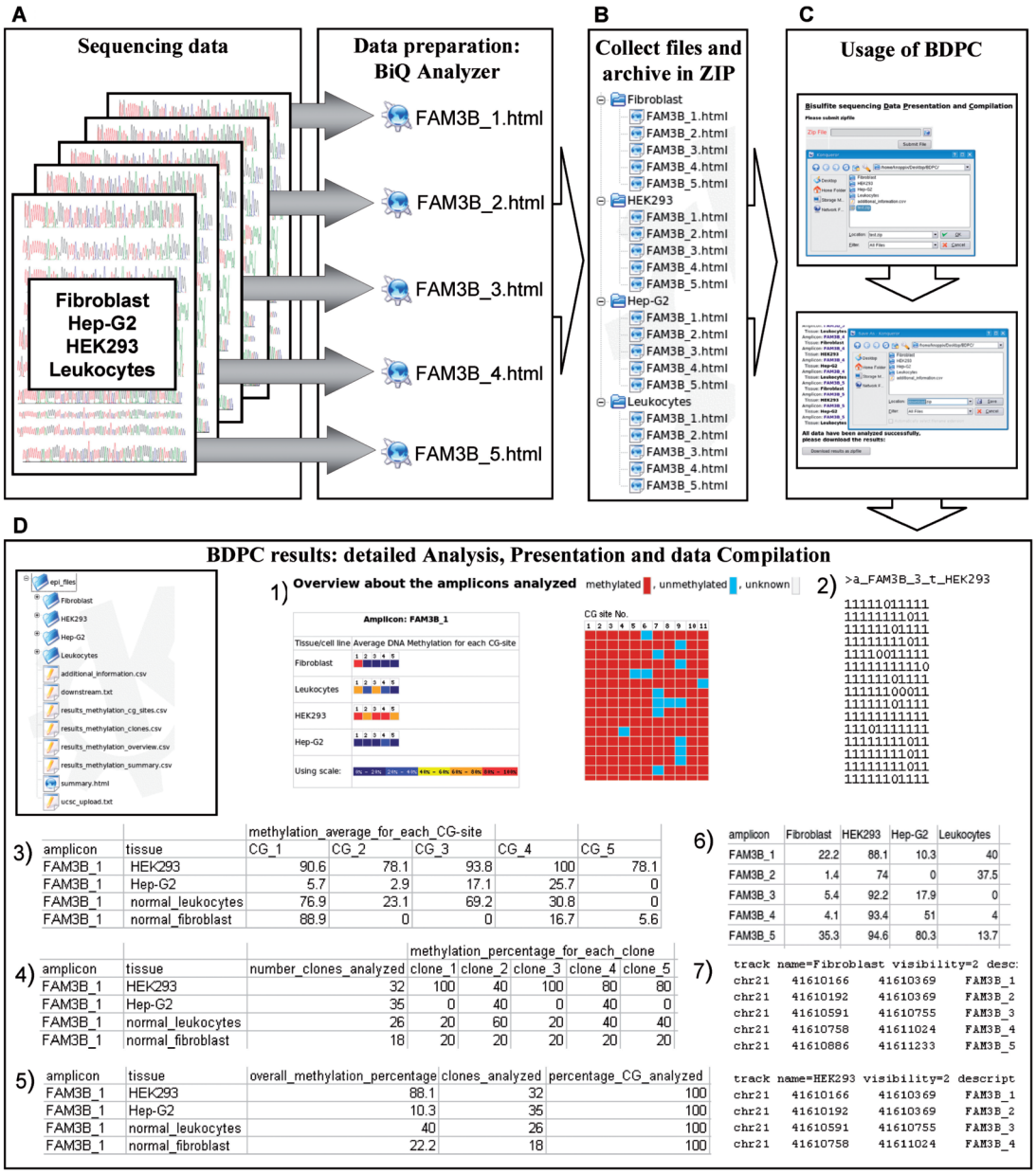

Figure 2. Workflow of bisulfite genomic sequencing analysis using BDPC. (**A–C**) Initial analysis of sequencing data is done using the BiQ Analyzer (A). The data are organized in folders (B) and uploaded to BDPC for analysis and data compilation (C). Afterwards the results can be downloaded in one ZIP file and extracted locally. (**D**) BDPC generates the following files: 1) The amplicon overview 'summary.html' file linked to the primary result HTML files with embedded pictures, 2) the 'downstream.txt' file compiling all primary data in one file, 3) the 'results_methylation_cg_sites.csv' file, 4) the 'results_methylation_clones.csv' file, 5) the 'results_methylation_summary.csv' file, 6) the 'results_methylation_overview.csv' table comparing the methylation results of all amplicons in all tissues and 7) the 'ucsc_upload.txt' file. (**E**) For each individual PCR product, a presentation ready HTML file is generated, that contains: 1) The sequence analyzed with numbered CG-dinucleotides. 2) The DNA methylation status of each CG-dinucleotide visualized graphically. Here each column corresponds to one CG-site analyzed in the PCR product. Each row represents one subcloned PCR product. Methylated CG-dinucleotides are presented as a red square, unmethylated as a blue square and CG-dinucleotides, which are not present are indicated in white. 3) The DNA methylation summary over all clones and statistics of the presence of CG-dinucleotides. 4) The average methylation level for each CG-site presented in a color-coded picture. 5) The average methylation for each subcloned DNA molecule presented in a table.

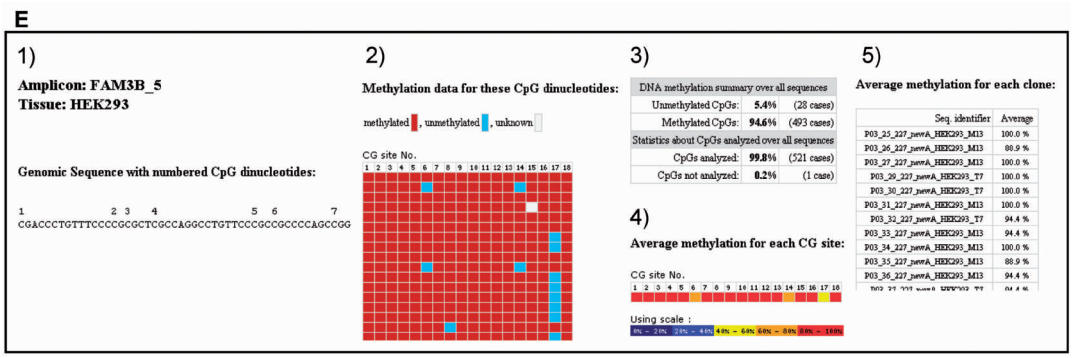

Figure 2. Continued.

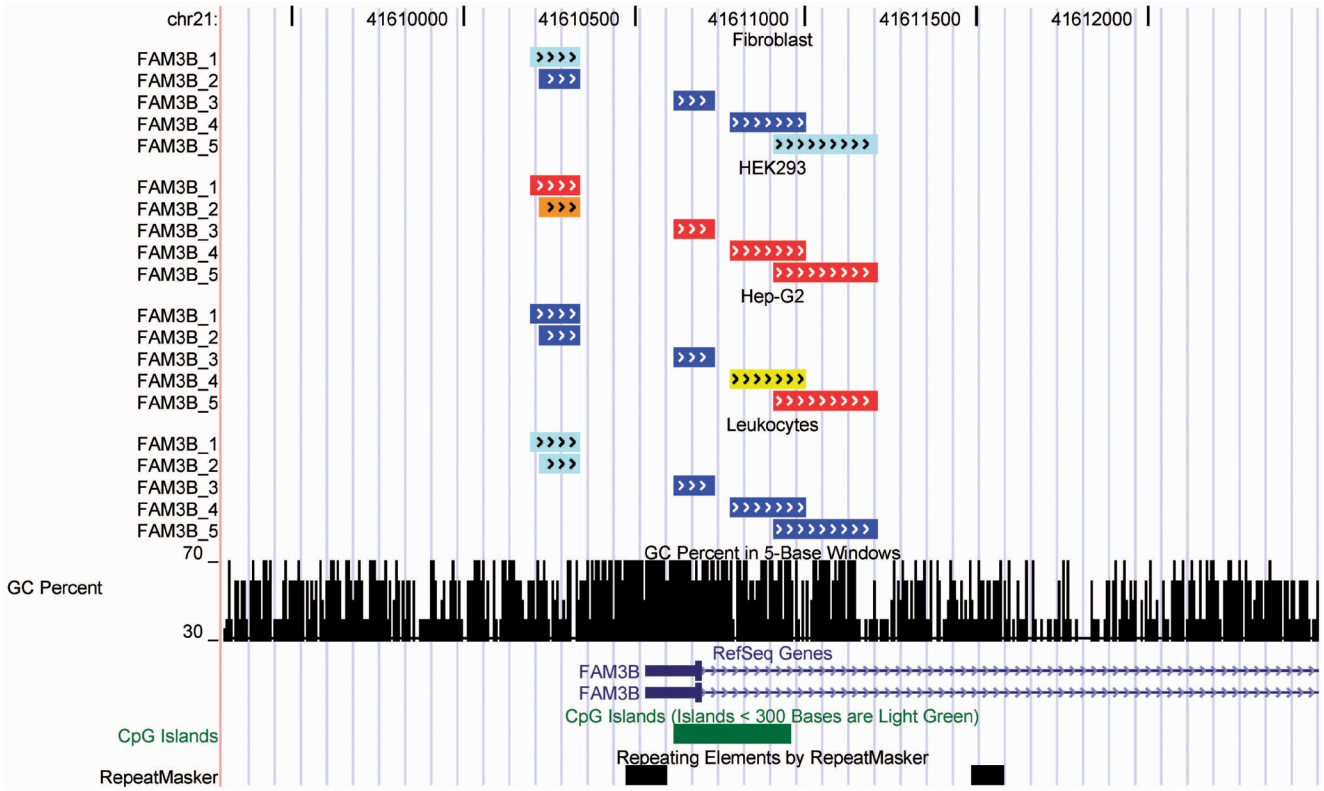

Figure 3. Display of BDPC results in the UCSC genome browser. Here, the position 41 609 300–41 612 500 of the NCBI36 assembly of human chromosome 21 is shown. The picture was generated by uploading the 'ucsc_upload.txt' file as a custom track at http://genome.ucsc.edu/cgi-bin/hgGateway. From top to bottom the figure shows methylation levels for different amplicons for HEK293, Leukocytes, Hep-G2, and Fibroblast. The arrows in the annotated PCR product indicate the DNA strand targeted for DNA methylation analysis. The overall DNA methylation level of the products is indicated by color: 0–20% is colored blue, 20.01–40% cyan, 40.01–60% yellow, 60.01–80% orange and 80.01–100% red. In addition, the GC-percentage, the RefSeq gene annotations, the annotated CpG-Islands and the repetitive sequence elements are displayed.

different cell lines or tissues, namely HEK293, Hep-G2, a Fibroblast cell line and Leukocytes. The sequencing results were analyzed with BiQ Analyzer and the result files were edited manually if necessary. Next, the result files were arranged as described earlier and shown in Figure 2, compressed in one ZIP file and analyzed using BDPC software.

After analysis, the results can be downloaded in ZIP format. BDPC provides the following files:

(i) One summary file that gives for each PCR product the amplicon name, the tissue analyzed, the overall DNA methylation percentage, the number of clones analyzed and the percentage of CG positions analyzed ('results_methylation_summary.csv').

(ii) A formatted overview table, which allows direct comparison of the average methylation levels of each amplicon in each tissue ('results_methylation_overview.csv').

(iii) One file giving the average methylation of each CG site for each PCR product ('results_methylation_cg_sites.csv'). Here, a threshold is implemented, such that only averages are calculated, if at least five results are available for the respective CG site.

Otherwise the methylation state of the CG site will be annotated as 'not determined'.

(iv) One file which contains the average methylation of each individual clone for each PCR product ('results_methylation_clones.csv').

(v) One primary data HTML file for each PCR product that contains the amplicon name, the tissue analyzed, the overall DNA methylation percentage and the methylation observed at each CG site, which is presented in a condensed picture in PNG format that can be directly used for data presentation.

(vi) A summary HTML file comparing the methylation pattern obtained with each amplicon in the different tissues ('summary.html'). The figures in this file are directly linked to the HTML files showing the individual results as described in (v). This file system can be used for immediate data presentation in the internet.

(vii) A custom track file for direct uploading of the results in the UCSC genome browser ('ucsc_upload.txt') (Figure 3).

(viii) One condensed data file collecting all primary data in simplified format to make the data accessible for later downstream analysis ('downstream.txt').

In summary, BDPC provides a useful resource for analysis of bisulfite DNA methylation data. BDPC does not only simplify the presentation and compilation procedure, but it also improves the analysis, because calculation of the overall methylation percentage considers whether a CG was found in the original sequencing run. In addition, the methylation state of a CG site is identified as 'not determined' if <5 clones contained data for this site. These features are important in case of poor data quality and presence of genetic polymorphisms. Evaluation of data quality is also assisted by the compilation of the overall coverage of CG sites for all PCR products provided in the 'results_methylation_summary.csv' output file.

Examples of the application of BDPC output files are listed below:

- The average data can be directly used for comparison of the methylation states of individual CG sites in the different tissues and to compute *P*-values for the significance of differences observed between two tissues by applying the statistics of a binomial distribution. Also, these data can be used to display methylation profiles for individual amplicons (Figure 4A).
- The methylation levels of individual clones can be used directly to compare the methylation of different tissues (Figure 4B). Also, they can be used for pair wise comparison of the methylation pattern of different tissues by *t*-test as shown in Figure 5. This procedure is accurate if both tissues show a unimodal distribution of methylation levels. In case of a bimodal distribution of at least one tissue, a simple *t*-test might not detect methylation differences although they are statistically significant.

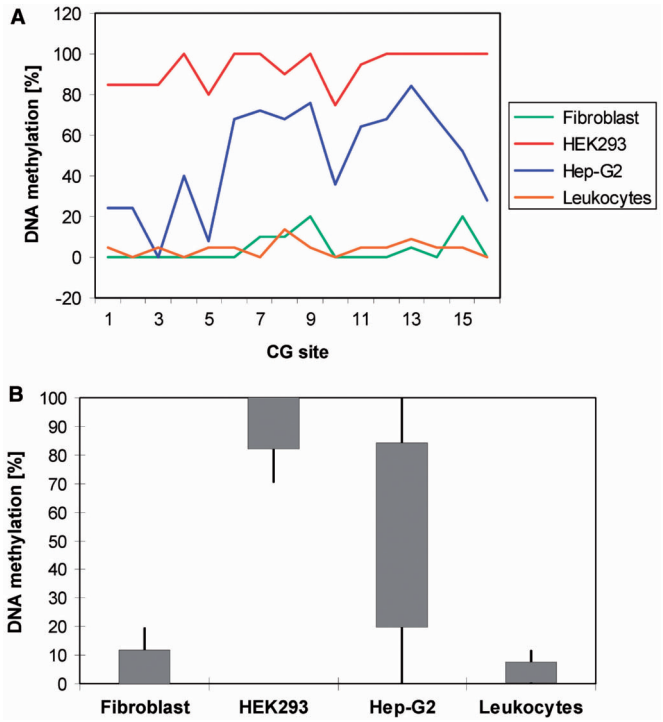

Figure 4. Examples of the application of BDPC output files for data presentation. (**A**) Distribution of methylation levels of individual CG sites on the FAM3B_4 amplicon in different tissues. This diagram was generated using the data compiled in the 'results_methylation_cg_sites.csv' file. (**B**) Overall methylation of clones of the FAM3B_4 amlicon in different tissues. This figure was prepared using the results compiled in the 'results_methylation_clones.csv' file by calculating the average methylation in each tissue together with the SE. The figure shows average ± 1 SE as grey box, average ± 2 SE as lines. The broad distribution observed in Hep-G2 is due to a biphasic distribution of methylation levels among the clones (see Figure 5).

- As mentioned above, the UCSC upload file can be used to display the methylation data in the UCSC genome browser (Figure 3).

ACKNOWLEDGEMENTS

This work was supported by a grant of the German minister of education and research (NGFN-2 program). We thank Ms S. Becker for technical assistance. Funding to pay the Open Access publication charges for this article was provided by Max Planck Institute, Berlin.

Conflict of interest statement. None declared.

REFERENCES

1. Hermann,A., Gowher,H. and Jeltsch,A. (2004) Biochemistry and biology of mammalian DNA methyltransferases. *Cell. Mol. Life Sci.*, **61**, 2571–2587.
2. Klose,R.J. and Bird,A.P. (2006) Genomic DNA methylation: the mark and its mediators. *Trends Biochem. Sci.*, **31**, 89–97.
3. Martin,C. and Zhang,Y. (2007) Mechanisms of epigenetic inheritance. *Curr. Opin. Cell Biol.*, **19**, 266–272.
4. Bird,A. (2002) DNA methylation patterns and epigenetic memory. *Genes Dev.*, **16**, 6–21.

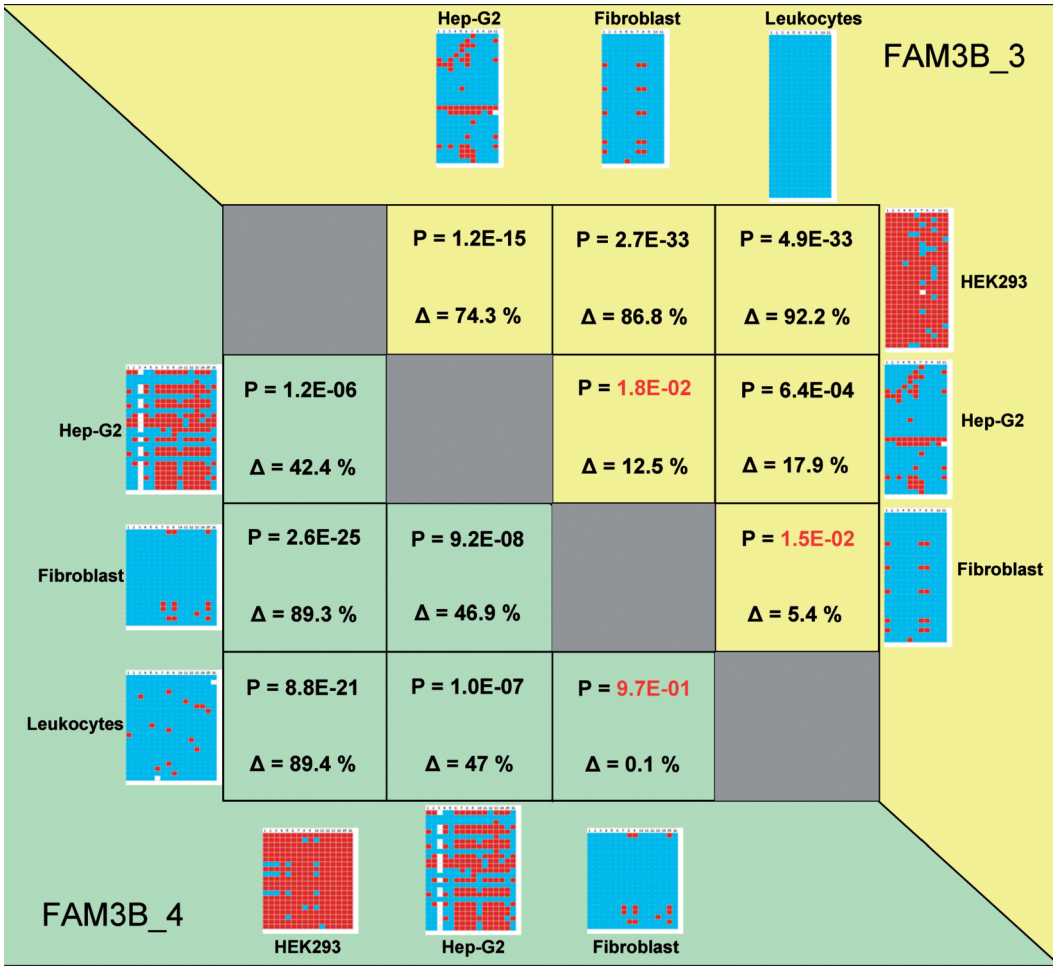

Figure 5. Pair wise comparison of methylation data obtained for the FAM3B_3 and FAM3B_4 amplicons in different cell lines and tissues. The figure shows the methylation patterns observed in different tissues for two amplicons: FAM3B_3 (in the yellow shaded part) and FAM3B_4 (in the green shaded part). In the table, the pairwise differences of the methylation levels (Δ in percentage) in different tissues and the *P*-values of the statistical significance of the differences are listed. *P*-values indicating no significant difference are colored red. The differences in the methylation levels were calculated using the methylation data given in 'results_methylation_summay.csv'. The *P*-values are calculated using the methylation levels of individual clones (provided in 'results_methylation_clones.csv') using a two-flanked *t*-test for samples with differing variance.

5. Li,E., Bestor,T.H. and Jaenisch,R. (1992) Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell*, **69**, 915–926.
6. Okano,M., Bell,D.W., Haber,D.A. and Li,E. (1999) DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, **99**, 247–257.
7. Feinberg,A.P. (2004) The epigenetics of cancer etiology. *Semin. Cancer Biol.*, **14**, 427–432.
8. Egger,G., Liang,G., Aparicio,A. and Jones,P.A. (2004) Epigenetics in human disease and prospects for epigenetic therapy. *Nature*, **429**, 457–463.
9. Weber,M., Hellmann,I., Stadler,M.B., Ramos,L., Paabo,S., Rebhan,M. and Schubeler,D. (2007) Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat. Genet.*, **39**, 457–466.
10. Eckhardt,F., Lewin,J., Cortese,R., Rakyan,V.K., Attwood,J., Burger,M., Burton,J., Cox,T.V., Davies,R., Down,T.A. *et al.* (2006) DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat. Genet.*, **38**, 1378–1385.
11. Frommer,M., McDonald,L., Millar,D., Collis,C., Watt,F., Grigg,G., Molloy,P. and Paul,C. (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.
12. Clark,S.J., Harrison,J., Paul,C.L. and Frommer,M. (1994) High sensitivity mapping of methylated cytosines. *Nucleic Acids Res.*, **22**, 2990–2997.
13. Grunau,C., Clark,S.J. and Rosenthal,A. (2001) Bisulfite genomic sequencing: systematic investigation of critical experimental parameters. *Nucleic Acids Res.*, **29**, E65–E65.
14. Zhang,Y., Rohde,C., Tierling,S., Stamerjohanns,H., Reinhardt,R., Walter,J. and Jeltsch,A. (2008) In Tost,J. (ed.), *Methods for DNA methylation analysis*, 2nd edn. Humana Press, Totowa, NJ.
15. 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.