BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 21 2005, pages 4067–4068
doi:10.1093/bioinformatics/bti652

Sequence analysis

BiQ Analyzer: visualization and quality control for DNA methylation data from bisulfite sequencing

Christoph Bock[1,*], Sabine Reither[2], Thomas Mikeska[2], Martina Paulsen[2], Jörn Walter[2] and Thomas Lengauer[1]

[1]Max-Planck-Institut für Informatik, Saarbrücken, Germany and [2]Universität des Saarlandes, FR 8.3 Biowissenschaften, Genetik/Epigenetik, Saarbrücken, Germany

Received on June 9, 2005; revised on August 17, 2005; accepted on August 26, 2005
Advance Access publication September 1, 2005

ABSTRACT

Summary: Manual processing of DNA methylation data from bisulfite sequencing is a tedious and error-prone task. Here we present an interactive software tool that provides start-to-end support for this process. In an easy-to-use manner, the tool helps the user to import the sequence files from the sequencer, to align them, to exclude or correct critical sequences, to document the experiment, to perform basic statistics and to produce publication-quality diagrams.

Emphasis is put on quality control: The program automatically assesses data quality and provides warnings and suggestions for dealing with critical sequences. The *BiQ Analyzer* program is implemented in the Java programming language and runs on any platform for which a recent Java virtual machine is available.

Availability: The program is available without charge for noncommercial users and can be downloaded from http://biq-analyzer.bioinf.mpi-inf.mpg.de/

Contact: cbock@mpi-inf.mpg.de

1 INTRODUCTION

DNA methylation is a frequent biochemical modification of eukaryotic DNA. In vertebrates, it almost exclusively affects the C5 position of cytosines that belong to CpG dinucleotides (i.e. a cytosine is directly followed by a guanine). Although this phenomenon has been known for several decades, it has recently witnessed a boost of attention. DNA methylation is assumed to play an important role in cancer (Feinberg and Tycko, 2004) and ageing (Issa, 2003). It is the cause for several developmental diseases (Walter and Paulsen, 2003). It has been brought into connection with chromatin remodeling (Reik *et al*., 2003), low success rates in mammalian cloning (Reik *et al*., 2003) and RNA interference (Kawasaki and Taira, 2004).

The most accurate and probably the most widely used experimental protocol for analyzing DNA methylation makes use of a selective conversion of unmethylated cytosines to uracils by bisulfite treatment (Frommer *et al*., 1992; Hajkova *et al*., 2002). Subsequent amplification, cloning, sequencing and comparison to the genomic sequence allows for identifying the unmethylated cytosines, which then appear as thymines in a multiple sequence alignment. Although this protocol is generally reliable, it gives rise to some potential error sources, which we address with our program.

Currently, few software tools exist that are tailored to support DNA methylation research. On the one hand, several primer design websites (Li and Dahiya, 2002; Tusnady *et al*., 2005) help the experimenter to prepare DNA methylation experiments, a problem that is upstream of the data processing task that we consider here. On the other hand, there is a basic Microsoft Excel template (Anbazhagan *et al*., 2001), which can assist with the calculation of average methylation and similar statistics when methylation data have already been generated and cleaned up (downstream of our task). The only software that partially overlaps in scope with *BiQ Analyzer* is MethTools (Grunau *et al*., 2000), a set of Perl scripts that generate publication-quality diagrams (lollipops and logos) from methylation data. *BiQ Analyzer* differs from MethTools in several respects. First, *BiQ Analyzer* imports sequence files directly from the sequencer without the need for any manual intervention and assists the user with all steps of alignment and quality control. Second, *BiQ Analyzer* does not only calculate summary statistics but can export the methylation data in full detail and in a format that makes it easy to import them into any statistics package or spreadsheet program. Third, *BiQ Analyzer* supports standardized experiment documentation. Finally, *BiQ Analyzer* provides an interactive graphical interface that guides the user through quality control and gives continuous feedback on problematic sequences.

2 QUALITY CONTROL METHODS

Potential error sources in bisulfite sequencing arise from three phases of the experimental protocol: bisulfite conversion, PCR and sequencing. Each of these steps can give rise to characteristic errors in the sequences, which the experimenter must address before deriving methylation profiles.

Here we describe these error types, their impact on methylation data and the quality control methods that *BiQ Analyzer* applies to identify the critical sequences.

Incomplete conversion. In bisulfite sequencing we assume that all unconverted Cs were originally methylated. Therefore, when the bisulfite treatment fails to convert unmethylated Cs, methylation will be overestimated. Fortunately, for vertebrates it is possible to identify those sequences with a low conversion rate. Assuming that Cs outside a CpG context are always unmethylated (Reik *et al*.,

[*]To whom correspondence should be addressed.

2003), *BiQ Analyzer* calculates the conversion rate of a sequence as the ratio between the number of correctly converted Cs outside a CpG context divided by the sum of converted and unconverted Cs outside a CpG context. By default, *BiQ Analyzer* highlights all sequences with a conversion rate lower than 90% as critical.

Clone sequences. PCR amplification can produce a vast over-representation of sequences from one or few individual chromosomes. Usage of such identical sequences results in biased estimation of DNA methylation. *BiQ Analyzer* implements a heuristic clone detection method. It highlights those sequences as critical that are identical in all correctly aligned C positions. The advantage of this method over simple sequence comparison is that it is insensitive to sequence truncations and sequencing errors at non-C positions.

Sequencing errors. Sequencing errors changing C to T and vice versa can lead to errors in the methylation data derived from the sequences. Therefore, *BiQ Analyzer* suggests excluding all sequences that fall below a local sequence identity level of 80% against the genome sequence (conversions and truncations are ignored). Furthermore, in our experiments we regularly observe ambiguous base insertions within a CpG context (i.e. CG → CTG or CG → TCG). In these cases, *BiQ Analyzer* reports the methylation state of the CpG dinucleotide as unknown.

The threshold levels for minimum conversion rate and minimum sequence identity are based on our experience with bisulfite sequencing and the user can change them in the configuration file of *BiQ Analyzer*.

3 PROGRAM OVERVIEW

BiQ Analyzer is a software tool designed to mimic the manual process of DNA methylation analysis. In several steps, the user is guided from the import of sequences, across several phases of quality control and multiple sequence alignment, to a questionnaire documenting the experiment. In each of the quality control steps, the program makes suggestions how to handle critical sequences, but the ultimate decision to include or exclude a sequence always stays with the user. Based on the user decisions during that process, the program finally generates a one-file HTML documentation (including publication-quality methylation diagrams in the widely-used 'lollipop' style) and saves the derived methylation data to the system clipboard, ready for subsequent analysis with a spreadsheet or a statistics program.

As a Java application, *BiQ Analyzer* runs on almost any platform, requiring only a recent version of the Java virtual machine (which can be downloaded from www.javasoft.com) and a screen resolution of at least 1024*768 pixels. For the multiple sequence alignment, a local version of ClustalW (Thompson *et al*., 1994) is used, which we include in the standard download package. The alignment step is computationally expensive and can be slow on older computers. Therefore, the program also provides an option to calculate the alignment over the internet on a high-performance computer at Max-Planck-Institut für Informatik.

4 CONCLUSION

BiQ Analyzer provides start-to-end support for the visualization and quality control of DNA methylation data from bisulfite sequencing. For the frequent user of bisulfite sequencing it will lead to significant speed up of the data analysis process. The occasional user will benefit from the extensive hints that help to perform a rigorous quality control. Beyond that, *BiQ Analyzer* promises to be a first step towards standardization in quality control and documentation. This is a necessary prerequisite for the second generation of DNA methylation databases that will validate data quality and that will accept direct submissions from the public. Non-commercial users can download *BiQ Analyzer* free of charge from http://biq-analyzer.bioinf.mpi-inf.mpg.de/.

ACKNOWLEDGEMENT

We thank Tarang Khare for helpful discussions and Joachim Büch for technical support. This work was conducted within the context of the EU Network of Excellence 'Biosapiens' (LSHG-CT-2003-503265) and the EU Network of Excellence 'The Epigenome' (LSHG-CT-2004-503433).

Conflict of Interest: none declared.

REFERENCES

Anbazhagan,R. *et al*. (2001) Spreadsheet-based program for the analysis of DNA methylation. *Biotechniques*, **30**, 110–114.

Feinberg,A.P. and Tycko,B. (2004) The history of cancer epigenetics. *Nat. Rev. Cancer*, **4**, 143–153.

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.

Grunau,C. *et al*. (2000) MethTools—a toolbox to visualize and analyze DNA methylation data. *Nucleic Acids Res*., **28**, 1053–1058.

Hajkova,P. *et al*. (2002) DNA-methylation analysis by the bisulfite-assisted genomic sequencing method. *Methods Mol. Biol*., **200**, 143–154.

Issa,J.P. (2003) Age-related epigenetic changes and the immune system. *Clin. Immunol*., **109**, 103–108.

Kawasaki,H. and Taira,K. (2004) Induction of DNA methylation and gene silencing by short interfering RNAs in human cells. *Nature*, **431**, 211–217.

Li,L.C. and Dahiya,R. (2002) MethPrimer: designing primers for methylation PCRs. *Bioinformatics*, **18**, 1427–1431.

Reik,W. *et al*. (2003) Mammalian epigenomics: reprogramming the genome for development and therapy. *Theriogenology*, **59**, 21–32.

Thompson,J.D. *et al*. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*., **22**, 4673–4680.

Tusnady,G.E. *et al*. (2005) BiSearch: primer-design and search tool for PCR on bisulfite-treated genomes. *Nucleic Acids Res*., **33**, e9.

Walter,J. and Paulsen,M. (2003) Imprinting and disease. *Semin. Cell Dev. Biol*., **14**, 101–110.