SOFTWARE TOOL ARTICLE

BAT: Bisulfite Analysis Toolkit [version 1; referees: 3 approved]

Helene Kretzmer [1,2], Christian Otto[1-3], Steve Hoffmann[1,2]

[1]Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Leipzig, 04109, Germany
[2]Transcriptome Bioinformatics, Research Center for Civilization Diseases (LIFE), University of Leipzig, Leipzig, 04109, Germany
[3]ecSeq GmbH, Leipzig, 04275, Germany

v1 **First published:** 16 Aug 2017, **6**:1490 (doi: 10.12688/f1000research.12302.1)
Latest published: 16 Aug 2017, **6**:1490 (doi: 10.12688/f1000research.12302.1)

Abstract

Here, we present **BAT**, a modular bisulfite analysis toolkit, that facilitates the analysis of bisulfite sequencing data. It covers the essential analysis steps of read alignment, quality control, extraction of methylation information, and calling of differentially methylated regions, as well as biologically relevant downstream analyses, such as data integration with gene expression, histone modification data, or transcription factor binding site annotation.

Open Peer Review

Referee Status: ✓ ✓ ✓

	Invited Referees 1	2	3
version 1 published 16 Aug 2017	✓ report	✓ report	✓ report

1 **Bob Zimmermann**, University of Vienna, Austria

2 **Ishaan Gupta**, Weill Cornell Medicine, USA

3 **Lars Feuerbach**, DKFZ (German Cancer Research Center), Germany

Discuss this article

Comments (0)

Corresponding author: Steve Hoffmann (steve@bioinf.uni-leipzig.de)

Author roles: Kretzmer H: Conceptualization, Software, Validation, Visualization, Writing – Original Draft Preparation; **Otto C**: Conceptualization, Validation, Writing – Original Draft Preparation; **Hoffmann S**: Funding Acquisition, Supervision, Writing – Original Draft Preparation

Competing interests: No competing interests were disclosed.

How to cite this article: Kretzmer H, Otto C and Hoffmann S. **BAT: Bisulfite Analysis Toolkit [version 1; referees: 3 approved]** *F1000Research* 2017, **6**:1490 (doi: 10.12688/f1000research.12302.1)

Grant information: This research was supported by the German BMBF (ICGC MMML-Seq 01KU1002A-J, and ICGC-Data Mining 01KU1505-C and G) the European Union in the framework of the BLUEPRINT Project (HEALTH-F5-2011-282510) and LIFE (Leipzig Research Center for Civilization Diseases), Leipzig University. LIFE is funded by the European Union, by the European Regional Development Fund (ERDF), the European Social Fund (ESF) and by the Free State of Saxony within the excellence initiative. We acknowledge support from the German Research Foundation (DFG) and University of Leipzig within the program of Open Access Publishing.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

First published: 16 Aug 2017, **6**:1490 (doi: 10.12688/f1000research.12302.1)

Introduction

High-throughput DNA methylation sequencing protocols, such as whole-genome bisulfite sequencing (WGBS) and targeted bisulfite sequencing (e. g., RRBS), have made it possible to precisely and accurately measure this major epigenetic modification on a genome wide scale. The impact of DNA methylation on processes, such as cell differentiation, gene expression, chromatin structure, and cancerogenesis, has raised substantial interest in analyzing DNA methylation in many sectors of life sciences. For example, methylomes of a large number of samples have been sequenced in the context of cancer projects and developmental studies[1–5]. Also researchers investigating obesity, neurodegenerative diseases, Alzheimer's, or Parkinson's disease, have begun to focus on DNA methylation[6–9].

A number of time consuming data analysis steps are required in virtually all these projects, i. e., quality control, read alignment, and methylation rate calculation. However, performing each step by hand is highly error prone, takes time, and impacts reproducibility. To ensure a consistent and reproducible processing, we have developed the Bisulfite Analysis Tooklit `BAT`. The workflow enables a fast and easy analysis of bisulfite converted high-throughput sequencing reads. It is specifically designed to facilitate the analysis for biologists and physicians with little bioinformatic knowledge, as well as for bioinformaticians that already work on sequencing data, but are not familiar with the characteristics of bisulfite sequencing data.

Methods

`BAT` is a modular toolkit allowing to easily generate workflows to analyze bisulfite sequencing data. The toolkit includes modules for read alignment (mapping module), methylation level estimation (calling module), grouping of samples (grouping module) and identification of differentially methylated regions (DMR module) (Figure 1). Further modules allow the integration of gene expression, histone modification data, or transcription factor binding site annotation. These modules facilitate the functional analysis of the effects of differential methylation.

Each of the modules can be run on its own, and the minimal system requirements depend on the respective module. The computational most expensive module is the mapping module. Here, the aligner `segemehl`[10] in its bisulfite mode is used, which requires about 55 GB physical RAM for the alignment of reads to the human genome hg19.

The toolkit itself is written in `Perl` and calls software components mainly written in `C` and `R` to ensure swift calculations. All software requirements are listed on our website (www.bioinf.uni-leipzig.de/Software/BAT/install/#requirements). The default parameters for the tools included in the `BAT` pipeline are optimized to process bisulfite sequencing data for most applications. In order to enhance reproducibility and reduce potential errors, the number of parameters that need to be set by the user has been carefully reduced to a minimum. Due to its modularity, however, the toolkit is flexible and can easily be extended or customized to specific needs. To allow for workflow modifications and extensions, standardized formats are used and interfaces to several other tools are provided. Basic steps, e.g., processing from raw reads to a single alignment file from multiple sequencing runs, is split into its pre-, post-, and main processing steps, to allow for the customized extension of the workflow. Error handling is eased by parameter and file checks prior to the analysis, and meaningful error messages allow a quick trouble shooting.

A detailed documentation of all modules, including parameter description, recommended additional tools, analysis reports, and data visualizations produced by the `BAT` workflow are summarized on www.bioinf.uni-leipzig.de/Software/BAT. Moreover, all automatically created visualizations are shown on the webpage. Data and figures displayed there are derived from a small example data set of two groups with four samples each, adopted from Kretzmer *et al*[11]. Our webpage provides raw FASTQ files of one sample as well as the methylation rate files of all eight samples along with expression and annotation data. This example data set and shell scripts covering all modules of `BAT` can be downloaded and adapted together with the toolkit.

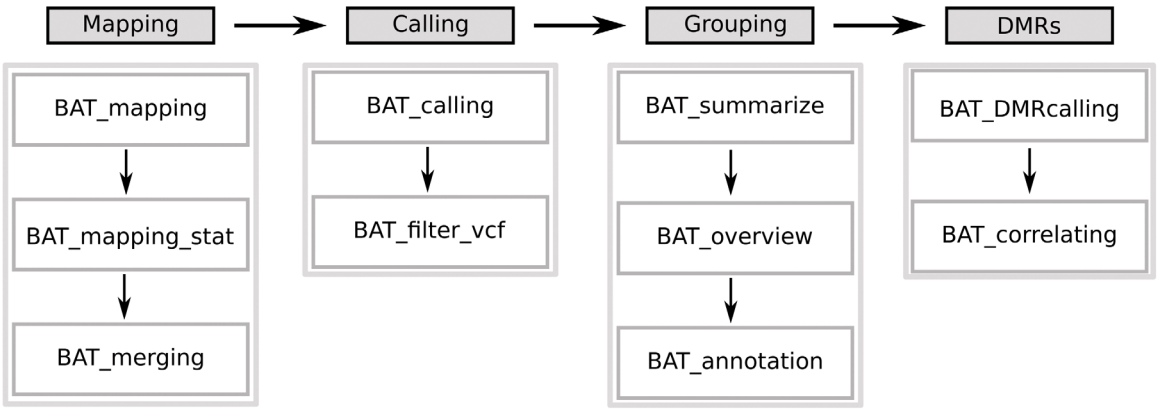

Figure 1. `BAT` workflow. It comprises four modules covering (left to right) read alignment, methylation rate calling, basic group analysis, and DMR calling. The modules consist of a collection of scripts that build up on one another, but easily single steps can be covered by alternative tools.

Furthermore, BAT is provided as Docker[12] image and can be obtained from https://hub.docker.com/r/christianbioinf/bat/. The Docker images ensure platform independent usage of our toolkit. All programs that are used by BAT are already installed in the Docker image and dependencies are resolved. Existing hard drives are mounted to avoid time consuming translocation or upload of the frequently huge data.

Use cases

Resembling a common study design, we assembled a small case-control example dataset, adopted from recently published data[11]. It is a subset of a paired-end human WGBS dataset, comprising 8 samples (control: S1–S4, case: S5–S8). It comprises the raw reads in FASTQ format of one sample and the already called methylation rates of all 8 samples in VCF format. The following modules can now be used to process and analyze bisulfite sequencing data including detection of methylation differences between case and control samples. The use case starts with the alignment of the raw sequencing data using the mapping module. The single components of BAT and their functionality are described in the following:

Mapping

The read alignment step is taken care of by the module BAT_mapping. It includes a bisulfite-sensitive read alignment using segemehl[10], a quality filtering step, and the conversion of the alignments to an indexed and compressed BAM file by samtools[13]. Using BAT_mapping_stat, the quality of the mapping can be assessed by the number and fraction of mapped pairs or reads, the multiplicity of read alignments, and the alignments' error rates. In case of large experiments where a sample is sequenced multiplexed on multiple lanes or flow cells, the read alignments of each sample can easily be merged using BAT_merging, including the addition of read group information to allow for tracebacks of lane effects if necessary.

Calling

Following mapping, the methylation information needs to be extracted from the read alignments. Prior to this methylation calling it is, however, recommended to exclude potential biases by clipping alignment overlaps of paired-end reads (e.g. using bamutil's clipoverlap[14]) or by excluding incompletely converted or artificially introduced cytosines with the M-bias detection method (e.g. using BSeQC[15]). Subsequently, the methylation information can be extracted using the module BAT_calling, which returns a VCF-style file that includes detailed information for each cytosine. This initial set of positions can be filtered by coverage using BAT_filtering to exclude unreliable methylation information from either lowly covered or very highly covered positions (e.g. in repetitive regions). Moreover, it is also possible to filter by genomic context (e.g., to restrict to CG context only). Apart from a VCF file, BAT_filtering reports the methylation level at positions passing the filter in bedGraph format for easy inspection in IGV[16] or upload to the UCSC genome browser[17]. Additionally, the module automatically produces plots showing the distribution of coverages and methylation rates for the complete and the filtered set of positions (Figure 2A), giving the user the opportunity to check and possibly fine tune the filtering parameters.

Groups

The third module now facilitates the transition from single sample analysis to groups of multiple samples. First, methylation information from individual samples is combined to groups and summarized with BAT_summarize. It reports the mean methylation rate per group and position as well as difference of the group's mean methylation rate per position. The summary module can be parameterized to only report positions where each group has a minimum number of samples with sufficient coverage. For convenience, all files are exported in both bedGraph and bigWig format for inspection in UCSC genome browser or IGV. Moreover, a circos plot containing a genome-wide methylation rate heatmap for each sample is automatically produced (Figure 2B). Based on the summary files, a number of overview statistics and plots can be generated using BAT_overview. This includes a hierarchical clustering of the samples based on their methylation profile, a plot of binned mean methylation rates per group (Figure 2C), boxplots of group-wise mean methylation rates (Figure 2D), a smoothed scatterplot showing the correlation between the groups' mean methylation rate per position, and a barplot of the distribution of group methylation differences. Subsequently, BAT_annotation can be used to inspect the methylation of the samples in regions of interest or annotations such as transcription factor binding sites (TFBS), CpG islands, shores, or promoter regions. Therefore, a hierarchically clustered heatmap of all samples (Figure 2E), is produced and the per-group and per-sample mean methylation rate is calculated (Figure 2F).

Differential methylated regions

Finally, the fourth module features the identification and analysis of differentially methylated regions (DMRs) between groups (BAT_DMRcalling). It employs the DMR calling tool metilene[18] which is based on circular binary segmentation of the group methylation difference signal in conjunction with a two-dimensional non-parametric statistical test. Afterwards, the DMRs reported by metilene can be filtered by several criteria, e.g., length (in nt or number of Cs), significance (i.e., q-value), and minimum mean methylation difference, and then converted to BED/bedGraph format. The BED file contains unique identifiers per DMR and reports regions of hyper/hypo methylation. Additionally, the bedGraph file can be used to display the mean group methylation difference of the DMRs. Moreover, BAT_DMRcalling produces overview statistics of the set of filtered DMRs including a histogram of the length and methylation difference of the filtered DMRs, a correlation plot of the mean methylation rate of DMRs in both groups and a plot of the methylation difference vs. the q-value for each DMR. Last but not least, BAT_correlating allows for integration of the DMRs with expression data. Given the methylation information, an expression value of genes, and an association between DMRs and genes, the correlation between both types of data can be examined in order to find correlating DMRs (cDMRs). For each DMR-gene pair, a linear and non-linear correlation coefficient is calculated and a correlation plot (Figure 2G), showing methylation and expression of each sample, is generated.

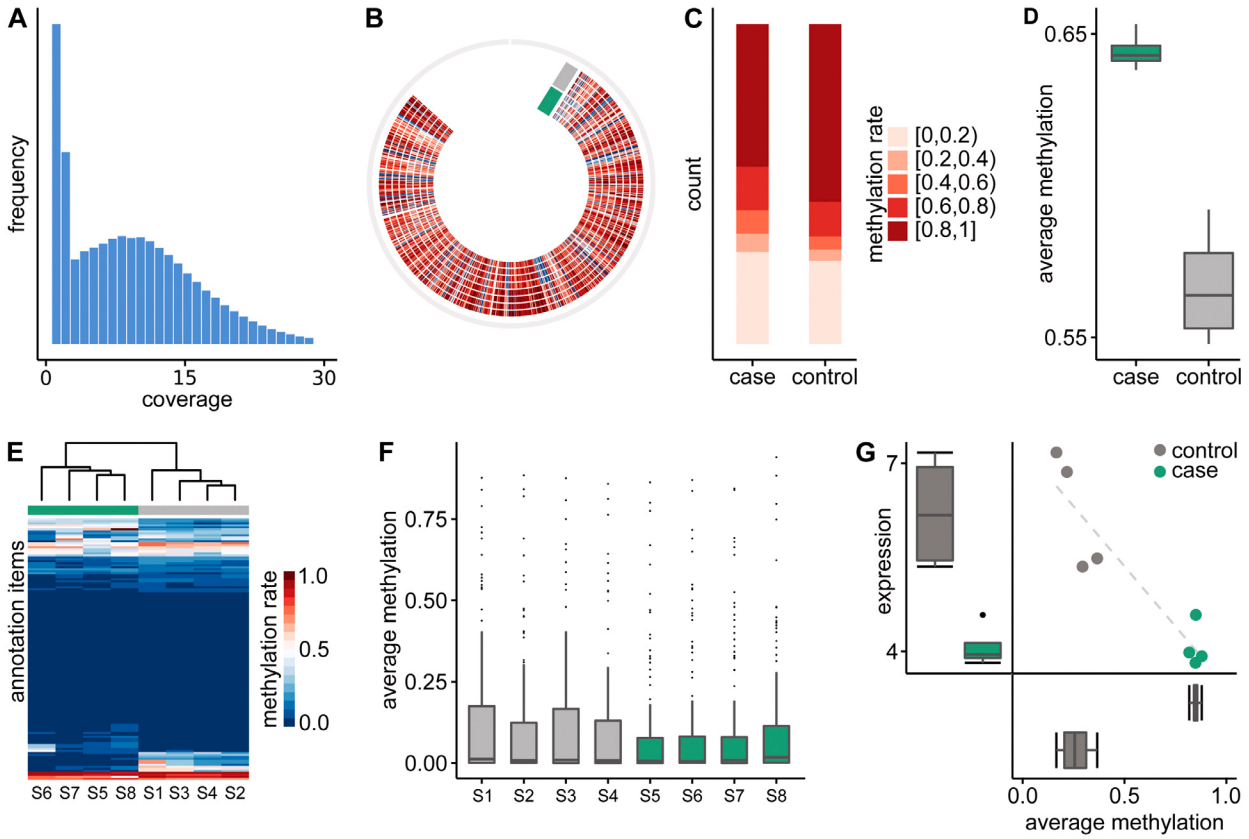

Figure 2. Selection of figures generated on-the-fly by `BAT` during the analysis of the example dataset. Annotation items are ENCODE transcription factor binding sites for GM12878 cell line. **A**) Distribution of coverage. **B**) Circos plot showing the genome-wide methylation level of eight samples as heatmap. **C**) Binned distribution of average methylation rate per CpG for each group. **D**) Boxplots of genome-wide mean methylation rate per group. **E**) Hierarchical clustered heatmap of the methylation rates of all samples over all annotation items. **F**) Boxplots of average methylation rate per annotation item. **G**) Correlating DMR plot shows methylation and expression of a DMR - gene pair. Note that all figures were produced by `BAT` itself, but were minorly post-edited to fit the limited space.

Summary

`BAT` has already successfully been applied in the framework of a large cancer genome study, the ICGC MMML-Seq[11]. The streamlined processing and analysis modules improve and accelerate the analysis by reducing hands on time and user errors. The modularity of `BAT`, as well as its input and output formats, allow to easily extend or customize the default workflows. For instance, it is possible to easily integrate tools such as `BisSNP`[19] or `BS-Snper`[20] or DMR calling tools.

The custom visualizations of the methylation data facilitate data mining and allow to inspect the data quality at each step of the analysis. This is necessary to increase the chance of an early detection of errors, e.g., in library preparation and data handling. Therefore, quality control statistics and graphics are produced continually throughout the entire pipeline.

Taken together, `BAT` is a collection of modular steps for analyzing bisulfite sequencing data that (i) can easily be run on various platforms due to the virtualization via `Docker`, (ii) can be combined with or extended by other tools, (iii) automatically generates publication-ready graphics, and (iv) supports data integration, e.g., annotation or gene expression data.

Software and data availability

Software available from: www.bioinf.uni-leipzig.de/Software/BAT/download

Source code available from: https://github.com/helenebioinf/BAT

Archived source code as at time of publication: http://doi.org/10.5281/zenodo.838200[21].

License: MIT

Example data available from: www.bioinf.uni-leipzig.de/Software/BAT/download/#example_data

Competing interests

No competing interests were disclosed.

Grant information

This research was supported by the German BMBF (ICGC MMML-Seq 01KU1002A-J, and ICGC-Data Mining 01KU1505-C and G) the European Union in the framework of the BLUEPRINT Project (HEALTH-F5-2011-282510) and LIFE (Leipzig Research Center for Civilization Diseases), Leipzig University. LIFE is funded by the European Union, by the European Regional Development Fund (ERDF), the European Social Fund (ESF) and by the Free State of Saxony within the excellence initiative. We acknowledge support from the German Research Foundation (DFG) and University of Leipzig within the program of Open Access Publishing.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

We thank Stephan H. Bernhart for helpful discussion and proof reading. We acknowledge support from the German Research Foundation (DFG) and Universität Leipzig within the program of Open Access Publishing.

References

1. International Cancer Genome Consortium, Hudson TJ, Anderson W, *et al.*: **International network of cancer genome projects.** *Nature.* 2010; **464**(7291): 993–998.
PubMed Abstract | Publisher Full Text | Free Full Text
2. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, *et al.*: **The Cancer Genome Atlas Pan-Cancer analysis project.** *Nat Genet.* 2013; **45**(10): 1113–1120.
PubMed Abstract | Publisher Full Text | Free Full Text
3. ENCODE Project Consortium: **An integrated encyclopedia of DNA elements in the human genome.** *Nature.* 2012; **489**(7414): 57–74.
PubMed Abstract | Publisher Full Text | Free Full Text
4. Bernstein BE, Stamatoyannopoulos JA, Costello JF, *et al.*: **The NIH Roadmap Epigenomics Mapping Consortium.** *Nat Biotechnol.* 2010; **28**(10): 1045–1048.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Martens JH, Stunnenberg HG: **BLUEPRINT: mapping human blood cell epigenomes.** *Haematologica.* 2013; **98**(10): 1487–1489.
PubMed Abstract | Publisher Full Text | Free Full Text
6. van Dijk SJ, Molloy PL, Varinli H, *et al.*: **Epigenetics and human obesity.** *Int J Obes (Lond).* 2015; **39**(1): 85–97.
PubMed Abstract | Publisher Full Text
7. De Jager PL, Srivastava G, Lunnon K, *et al.*: **Alzheimer's disease: early alterations in brain DNA methylation at *ANK1*, *BIN1*, *RHBDF2* and other loci.** *Nat Neurosci.* 2014; **17**(9): 1156–1163.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Schumacher A, Petronis A: **Epigenetics of complex diseases: from general theory to laboratory experiments.** *Curr Top Microbiol Immunol.* 2006; **310**: 81–115.
PubMed Abstract | Publisher Full Text
9. Jowaed A, Schmitt I, Kaut O, *et al.*: **Methylation regulates alpha-synuclein expression and is decreased in Parkinson's disease patients' brains.** *J Neurosci.* 2010; **30**(18): 6355–6359.
PubMed Abstract | Publisher Full Text
10. Otto C, Stadler PF, Hoffmann S: **Fast and sensitive mapping of bisulfite-treated sequencing data.** *Bioinformatics.* 2012; **28**(13): 1698–1704.
PubMed Abstract | Publisher Full Text
11. Kretzmer H, Bernhart SH, Wang W, *et al.*: **DNA methylome analysis in Burkitt and follicular lymphomas identifies differentially methylated regions linked to somatic mutation and transcriptional control.** *Nat Genet.* 2015; **47**(11): 1316–1325.
PubMed Abstract | Publisher Full Text | Free Full Text
12. Merkel D: **Docker: Lightweight linux containers for consistent development and deployment.** *Linux J.* 2014; **2014**(239).
Reference Source
13. Li H, Handsaker B, Wysoker A, *et al.*: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics.* 2009; **25**(16): 2078–2079.
PubMed Abstract | Publisher Full Text | Free Full Text
14. Wang KC, Yang YW, Liu B, *et al.*: **A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression.** *Nature.* 2011; **472**(7341): 120–124.
PubMed Abstract | Publisher Full Text | Free Full Text
15. Lin X, Sun D, Rodriguez B, *et al.*: **BSeQC: quality control of bisulfite sequencing experiments.** *Bioinformatics.* 2013; **29**(24): 3227–3229.
PubMed Abstract | Publisher Full Text | Free Full Text
16. Thorvaldsdóttir H, Robinson JT, Mesirov JP: **Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration.** *Brief Bioinform.* 2013; **14**(2): 178–92.
PubMed Abstract | Publisher Full Text | Free Full Text
17. Karolchik D, Barber GP, Casper J, *et al.*: **The UCSC Genome Browser database: 2014 update.** *Nucleic Acids Res.* 2014; **42**(Database issue): D764–70.
PubMed Abstract | Publisher Full Text | Free Full Text
18. Jühling F, Kretzmer H, Bernhart SH, *et al.*: **metilene: Fast and sensitive calling of differentially methylated regions from bisulfite sequencing data.** *Genome Res.* 2016; **26**(2): 256–62.
PubMed Abstract | Publisher Full Text | Free Full Text
19. Liu Y, Siegmund KD, Laird PW, *et al.*: **Bis-SNP: combined DNA methylation and SNP calling for bisulfite-seq data.** *Genome Biol.* 2012; **13**(7): R61.
PubMed Abstract | Publisher Full Text | Free Full Text
20. Gao S, Zou D, Mao L, *et al.*: **BS-SNPer: SNP calling in bisulfite-seq data.** *Bioinformatics.* 2015; **31**(24): 4006–4008.
PubMed Abstract | Publisher Full Text | Free Full Text
21. helenebioinf: **helenebioinf/BAT: Publication Release.** *Zenodo.* 2017.
Data Source

Open Peer Review

Current Referee Status: ✔ ✔ ✔

Version 1

Referee Report 12 September 2017

doi:10.5256/f1000research.13317.r25075

✔ **Lars Feuerbach**
Comparative Cancer Genomics, Division of Applied Bioinformatics, DKFZ (German Cancer Research Center), Heidelberg, Germany

The manuscript “BAT: Bisulfite Analysis Toolkit” presents a software pipeline for the analysis of sequencing-based analysis of bisulfite treated DNA. It introduces the major modules of this pipeline and familiarize the reader with their basic function, compatibilities and output, but is obviously not intended to provide sufficient detail to allow reimplementation of the described modules. Instead, it refers to external resources such as research articles and documentary webpages, which provide most of this information.

The article excels in providing a researcher who has to choose among several software pipelines for his next methylation project with the necessary information on BAT, without attempting to benchmark it against other approaches.

Especially, the offer of a dockerized pipeline version and a real example datasets ensures the applicability of the software, while simultaneously proving the claim of improved reproducibility.

Another prominent claim, namely the compatibility with other modules for instance alternatives to segemehl, is less well documented. Here the article would profit from an extended example in which some of the modules are exchanged by third party alternatives, e.g. in the alignment step or during the grouping.

Finally, the authors describe the utility of their diagnostic diagrams depicted in figure 2 for the detection of quality problems. To this end a supplementary figure/resource in which a number of examples of how several quality problems manifest in these diagrams is required, not only to proof this statement, but also to educate less experienced users.

Minor comment:
The sentences “However, performing each step by hand is highly error prone, takes time, and impacts reproducibility” in the introduction is formulated unfavorably, as it can be misread as a suggestion that someone would attempt to analyze a WGBS dataset by hand.

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Partly

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Referee Expertise: Epigenetics, Cancer genomics, Software development

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 06 September 2017

doi:10.5256/f1000research.13317.r25078

Ishaan Gupta
Brain and Mind Research Institute, Weill Cornell Medicine, New York City, NY, USA

BAT: Bisulphite analysis toolkit is a timely software which provides an end-to-end solution for performing DNA methylation analysis. The toolkit follows "good software practises" and has a clearly laid out work flows, efficient code , extensive documentation and has limited dependencies. Further, ability to perform the complete analysis from sequencing data to actual interpretation and integration of data as shown in the example data in the toolkit from the manuscript by the same authors "DNA methylome analysis in Burkitt and follicular lymphomas identifies differentially methylated regions linked to somatic mutation and transcriptional control"[1] suggests that the method implemented in the toolkit for calling differentially methylated regions (DMRs) is not only much faster than existing solutions but also extracts biologically relevant information about methylation.

I outline my reasons below :

- Is the rationale for developing the new software tool clearly explained?

The rationale for developing the software well explained as sequencing data especially bisulphite sequencing data are prone to human errors and increasing number of samples being processed for cohorts tackling complex disease phenotypes warrant for a streamlined reproducible workflows. Also the ease of use is a term often loosely used for many bioinformatics tools are under-appreciated by the community but the authors have done well here by providing a docker image that obviates any platform dependencies to provide an out of the box solution.
Suggestion:

As a rationale it would would great if the authors could add a few lines on their method of calculating DMRs in the introduction to contrast with existing tools, I believe this would enhance the manuscript and further convince the readers to use this tool-kit.

- Is the description of the software tool technically sound?

Software documentation is thorough and technically sound.
Moreover, Dr. Hoffmann's lab has been quite consistent in releasing regular updates for their previous tools and is responsive to bug-reports.
Suggestion:
It is accurate that segemehl requires 55GB to align the entire human genome but it would be important to also point out that alignment could be run on individual chromosomes separately and then combined later which significantly reduces this memory intensive step.

- Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

The workflows are well laid out and broken down into the individual modules establishing a replicable software design. Each module can be run individually or together through the perl wrapper and come with appropriate description of flags used in the command line help allowing a look under the hood of the code. Further, each tool is well documented and the code is commented making the tool reproducible.

- Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

The example provided herewith runs well and one can quickly reproduce the plots from Kretzmer *et al* 2015[1].

References

1. Kretzmer H, Bernhart SH, Wang W, Haake A, Weniger MA, Bergmann AK, Betts MJ, Carrillo-de-Santa-Pau E, Doose G, Gutwein J, Richter J, Hovestadt V, Huang B, Rico D, Jühling F, Kolarova J, Lu Q, Otto C, Wagener R, Arnolds J, Burkhardt B, Claviez A, Drexler HG, Eberth S, Eils R, Flicek P, Haas S, Humme M, Karsch D, Kerstens HHD, Klapper W, Kreuz M, Lawerenz C, Lenzek D, Loeffler M, López C, MacLeod RAF, Martens JHA, Kulis M, Martín-Subero JI, Möller P, Nage I, Picelli S, Vater I, Rohde M, Rosenstiel P, Rosolowski M, Russell RB, Schilhabel M, Schlesner M, Stadler PF, Szczepanowski M, Trümper L, Stunnenberg HG, Küppers R, Ammerpohl O, Lichter P, Siebert R, Hoffmann S, Radlwimmer B: DNA methylome analysis in Burkitt and follicular lymphomas identifies differentially methylated regions linked to somatic mutation and transcriptional control.*Nat Genet*. 2015; **47** (11): 1316-1325 PubMed Abstract | Publisher Full Text

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 29 August 2017

doi:10.5256/f1000research.13317.r25076

Bob Zimmermann
Department of Molecular Evolution and Development, University of Vienna, Vienna, Austria

This article presents a tool aggregate which can be a useful one-stop-shop and/or starting off point for analyzing bisulfite data. The authors detail the package and demonstrate its usefulness in an example analysis. Most of the information needed to decide whether to use this package is contained in the article.

A notable exception are the "further modules" mentioned in the first paragraph of the Methods section. While it is clearly useful to integrate gene expression, histone modification data and transcription factor binding site information to your analysis, the reader cannot get an impression of whether the package does this effectively. It would be useful to include either an expansion on this topic or better yet to add example analysis with these modules as well, space permitting. If the authors are space confined, it would be useful to point to where an example of this can be found on the web, as I was unable to locate it on the project page.

A critical omission from the introduction is a short technical background on bisulfite sequencing and its analysis. The reader has no basis to understand why a "VCF-style file that includes detailed information for each cytosine" (in the Calling subsection) would be useful.

Some minor issues were:

- the phrase "grouping of samples" in the second sentence of the Methods section does not really clarify anything about the function of the grouping module. I would suggest to use "sample group analysis"
- "Due to its modularity, however" is awkwardly worded and could be better expressed as "The toolkit's modularity makes it flexible, extensible and customizable for users with specific needs".
- the sentence "Basic steps, e.g. ..." should say "are" instead of "is".
- "Resembling a common study design," in the Use cases section does not express what I believe is the authors' intended meaning, and could be better worded as "In order to illustrate the results of using our toolkit on a common study design,"

Is the rationale for developing the new software tool clearly explained?
Partly

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Partly

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Referee Expertise: Bioinformatics, evolution and development

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.