bioRxiv preprint doi: https://doi.org/10.1101/859900; this version posted November 30, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

wg-blimp: an end-to-end analysis pipeline for whole genome bisulfite sequencing data

Marius Wöste[1,✉], Elsa Leitão[2], Sandra Laurentino[3], Bernhard Horsthemke[2,4], Sven Rahmann[5], and Christopher Schröder[2,5]

[1]Institute of Medical Informatics, University of Münster, Albert-Schweitzer-Campus 1, 48149 Münster, Germany
[2]Institute of Human Genetics, University of Duisburg-Essen, University Hospital Essen, Hufelandstraße 55, 45122 Essen, Germany
[3]Centre of Reproductive Medicine and Andrology, Institute of Reproductive and Regenerative Biology, University Hospital Münster, Albert-Schweitzer-Campus 1, 48149 Münster, Germany
[4]Institute of Human Genetics, University of Münster, Vesaliusweg 12-14, 48149 Münster, Germany
[5]Genome Informatics, Institute of Human Genetics, University of Duisburg-Essen, University Hospital Essen, Hufelandstraße 55, 45122 Essen

Background: **Analysing whole genome bisulfite sequencing datasets is a data-intensive task that requires comprehensive and reproducible workflows to generate valid results. While many algorithms have been developed for tasks such as alignment, comprehensive end-to-end pipelines are still sparse. Furthermore, previous pipelines lack features or show technical deficiencies, thus impeding analyses.**

Results: **We developed wg-blimp (whole genome bisulfite sequencing methylation analysis pipeline) as an end-to-end pipeline to ease whole genome bisulfite sequencing data analysis. It integrates established algorithms for alignment, quality control, methylation calling, detection of differentially methylated regions, and methylome segmentation, requiring only a reference genome and raw sequencing data as input. Comparing wg-blimp to previous end-to-end pipelines reveals similar setups for common sequence processing tasks, but shows differences for post-alignment analyses. We improve on previous pipelines by providing a more comprehensive analysis workflow as well as an interactive user interface. To demonstrate wg-blimp's ability to produce correct results we used it to call differentially methylated regions for two publicly available datasets. We were able to replicate 112 of 114 previously published regions, and found results to be consistent with previous findings. We further applied wg-blimp to a publicly available sample of embryonic stem cells to showcase methylome segmentation. As expected, unmethylated regions were in close proximity of transcription start sites. Segmentation results were consistent with previous analyses, despite different reference genomes and sequencing techniques.**

Conclusions: **wg-blimp provides a comprehensive analysis pipeline for whole genome bisulfite sequencing data as well as a user interface for simplified result inspection. We demonstrated its applicability by analysing multiple publicly available datasets. Thus, wg-blimp is a relevant alternative to previous analysis pipelines and may facilitate future epigenetic research.**

whole-genome bisulfite sequencing | analysis pipeline | epigenetics | methylation | analysis workflow

Correspondence: ***marius.woeste@uni-muenster.de***

Background

Since the development of DNA sequencing, a large number of studies on genetic variation have been conducted, while extensive research on the epigenetic level has only emerged in the recent past. Although most cells within an organism are identical in their genomic sequence, different tissues and cell types vary in their patterns of epigenetic modifications that confer their particular identity. DNA methylation is one of the most important epigenetic marks and occurs mainly at CpG dinucleotides. There are almost 28 million of such sites in the human genome, thus 450k arrays (which cover only 1.6% of all CpGs) are not sufficient to detect small differentially methylated regions (DMRs) (1). As a result, data-intensive whole genome bisulfite sequencing (WGBS) is required to properly identify all CpG methylation levels. While the costs for generating these data sets have been very high, the continuous and sustained reduction of sequencing costs allows more and more WGBS datasets to be generated, creating the need for comprehensive and reproducible analysis tools. Many algorithms have already been established for different aspects of WGBS analyses such as alignment and DMR detection. However, choosing appropriate algorithms and integrating them into an end-to-end analysis workflow is not a trivial task due to combinatorial explosion of possible pipeline setups. Setting up an end-to-end WGBS analysis workflow is further hindered by different requirements of interacting tools, e.g. input and output formats or chromosome naming conventions. Previously developed end-to-end pipelines already consider these problems and only require users to supply their raw data and configuration. However, we find previous approaches to lack features required in common research settings, e.g. methylome segmentation, as well as technical limitations such as installation issues, as described in more detail in the *Results & Discussion* section. As a result, we developed a pipeline approach to address these issues.

Implementation

We present here wg-blimp (whole genome bisulfite sequencing methylation analysis pipeline), a workflow for automated in silico processing of WGBS data. It consists of a comprehensive WGBS data analysis pipeline as well as a user interface for simplified inspection of datasets and potential sharing of results with other researchers. Figure 1 gives an overview of the analysis steps provided.

With FASTQ files and a reference genome as input, wg-blimp performs a complete workflow from alignment to DMR analysis, segmentation and annotation. We choose bwa-meth (2) for alignment as it provides efficient and robust mappings due to its internal usage of BWA-MEM (3). We omit pre-

DRAFT

bioRxiv preprint doi: https://doi.org/10.1101/859900; this version posted November 30, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

alignment trimming of reads because of bwa-meth's internal usage of soft-clipping to mask non-matching read subsequences. Alignments are deduplicated using the Picard toolkit (4). Methylation calling is performed by MethylDackel (5) as it is the recommended tool for use with bwa-meth. Based on the methylation reports created by MethylDackel, wg-blimp computes global methylation statistics. Computing per-chromosome methylation is optional and enables estimation of C > T conversion rates, as unmethylated lambda DNA is commonly added to genomic DNA prior to bisulfite treatment.

For quality control (QC) we use FastQC (6) to evaluate read quality scores. Coverage reports containing information about overall and per-chromosome coverage are generated by Qualimap (7). Qualimap also reports metrics such as GC content, duplication rate, and clipping profiles, thus enabling in-depth quality evaluation of each sample analysed. Quality reports by Picard, Qualimap and FastQC are aggregated into a single interactive HTML report using MultiQC (8)

Multiple algorithms are supported for DMR calling: metilene (9), bsseq (10) and camel (11) are frequently used tools. The application of more than one DMR calling tool is recommended, because these tools identify different, although overlapping sets of DMRs.

We further integrate detection of unmethylated regions (UMRs) and low-methylated regions (LMRs) to identifify active regulatory regions in an unbiased fashion. This segmentation is implemented using MethylSeekR (12) as it provides automatic inference of model parameters using only a user-defined false-discovery rate (FDR) and methylation cutoff. MethylSeekR also implements detection of regions of highly disordered methylation, termed partially methylated domains (PMDs). The presence of PMDs is influencing UMR/LMR detection and is often unknown a priori. As a result, wg-blimp preemptively performs the MethylSeekR workflow with and without PMD computation. Based on the metrics measured by MethylSeekR users may decide wether or not to consider PMDs when analysing UMRs and LMRs.

Resulting DMRs, UMRs, LMRs and PMDs are annotated for overlap with genes, promoters, CpG islands (CGIs) and repetitive elements as reported by Ensembl (13) and UCSC (14) databases. Average coverage per DMR is computed using mosdepth (15) to enable filtering of DMR calls in regions of low coverage.

We base the wg-blimp pipeline on the workflow execution system Snakemake (16) as it enables robust and scalable execution of analysis pipelines and prevents generation of faulty results in case of failure. Snakemake also provides run-time and memory usage logging, thus easing the search for bottlenecks and performance optimization. To minimize errors caused by changing software versions we utilize Bioconda (17) for dependency management and installation.

Once the analysis workflow completes, users may load the results into wg-blimp's user interface. We implemented the interface using the R Shiny framework that enables seamless integration of R features into a reactive web app. The interface aggregates QC reports, pipeline parameters, and allows inspection and filtering of DMRs based on caller output and annotations. UMRs and LMRs computed by MethylSeekR may also be accessed through wg-blimp's Shiny interface, and users may dynamically choose whether or not to include PMDs. Since visualization of genomic data is often employed when inspecting analysis results, access links to alignment data for use with the Integrative Genomics Viewer (IGV) (18) are also provided, as IGV provides a bisulfite mode for use with WGBS data.

Results & Discussion

To evaluate wg-blimp's relevance for WGBS experiments, we compared it to previous end-to-end pipelines and demonstrated its applicability by analysing three exemplary datasets.

Comparison to previous pipelines. Since wg-blimp only integrates published software, and exhaustive evaluation of all conceivable pipeline setups would result in combinatorial explosion, we focus here on a feature-wise comparison of pipelines, similar to (19). We compared wg-blimp to BAT (20), bicycle (21), CpG_Me/DMRichR (10, 22–24), ENCODE-DCC's WGBS pipeline (25), Methy-Pipe (26), Nextflow methylseq (two available workflows) (27), PiGx (28) and snakePipes (19). Pipelines were compared with regards to technical setup (installation, workflow management), WGBS read processing (adapter trimming, alignment, methylation calling, quality control), and post-alignment analyses (DMR detection, segmentation, annotation).

Table 1 gives an overview over each pipeline's setup. Similar to snakePipes, wg-blimp utilizes Bioconda for installation. Using package managers such as Bioconda or workflow environments like Nextflow (29) not only simplifies installation for users but also provides straightforward update processes of both the pipeline itself as well as its dependencies. Thus, we recommend usage of such package managers to ensure stable runtime environments. For workflow management, we prefer using dedicated workflow management systems such as Snakemake or Nextflow over plain shell scripts, as these allow more scalable and robust execution. Users may also consider using cloud computing platforms such as DNAnexus (dnanexus.com). These platforms alleviate setting up own hardware for analysis, with the downside of users providing their data to third-party providers, thus posing potential data privacy risks.

For read processing, wg-blimp employs similar strategies as other pipelines, with popular alignment and methylation calling tools being bwa-meth/MethylDackel and Bismark (24). However, wg-blimp deviates from other pipelines by skipping read trimming, which is handled by BWA-MEM's soft-clipping. For QC we recommend using MultiQC as it produces HTML quality reports in a compact and scalable way. We omitted the details about which metrics are collected by MultiQC for each pipeline, as the pipelines investigated use common tools such as Picard or sambamba (30) (with the exception of BAT, bicycle and Methy-Pipe).

DRAFT

bioRxiv preprint doi: https://doi.org/10.1101/859900; this version posted November 30, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

While most of the pipelines investigated use similar tools for read processing, setups differ for post-alignment analyses. For DMR detection, we pursue a similar setup as snakePipes and BAT by providing multiple DMR callers. wg-blimp and PiGx are the only workflows to perform methylome segmentation. We prefer MethylSeekR over methylKit for segmentation because of its consideration of PMDs.

We further added functionality over other pipelines by implementing an interactive R Shiny GUI. Users may load one or more analysis runs into the Shiny App, thus providing a straightforward way to create a central repository for analysis results to share with fellow researchers. This not only makes distributing individual files unnecessary but also enables a more concise inspection of results. For example, users may switch between segmentation with and without consideration of PMDs using MethylSeekR by toggling a single checkbox instead of having to inspect multiple files. An example of wg-blimp's interface displaying MethylSeekR results is given in Figure 2. More GUI features are discussed in detail in the Supplementary Material.

While we provide additional functionality over previous WGBS pipelines, we would like to emphasize that wg-blimp should not be seen as a replacement for previous approaches, but rather as an extension to the landscape of available workflows. snakePipes, for example, not only provides a WGBS analysis workflow, but is also capable of performing integrative analyses on ChIP-seq, RNA-seq, ATAC-seq, Hi-C and single-cell RNA-seq data. As a result, snakePipes should be preferred over wg-blimp in experiments that aim at integrating different epigenomic assays. In contrast, we prefer wg-blimp over snakePipes for WGBS-only experiments that aim at determining active regulatory regions due to its implementation of segmentation and simplified dataset inspection through its GUI. Thus, when deciding which analysis workflow to choose for a WGBS experiment, we believe there is no "one-fits-all" solution, and we deem wg-blimp one suitable option to consider for future WGBS analyses.

Application to published datasets. We applied wg-blimp to three exemplary publicly available WGBS datasets. Two of these datasets were utilized to demonstrate wg-blimp's DMR calling capabilities and a third to demonstrate methylome segmentation. All analyses were executed on a server equipped with two Intel Xeon E5-2695 v4 CPU's, 528 GB of memory and Debian 9 as operating system (OS). 64 threads were allocated for each analysis.

DMR detection. One of the DMR datasets consists of two pairs of isogenic human monocyte and macrophage samples (31), the other of two pairs of isogenic human blood and sperm samples (each generated from pools of DNA from six men) (32). We chose these two datasets to demonstrate wg-blimp's capability of calling DMRs for cases where few (monocytes vs. macrophages) or many (blood vs. sperm) DMRs are expected due to the degree of relatedness between compared groups.

For the monocyte/macrophage dataset we chose hg38 as reference and used a coverage of at least 5×, at least 4 CpG sites overlapping, and a minimum absolute difference of 0.3 as thresholds for DMR calling. We detected 6,189 DMRs in total, with 4,078 DMRs overlapping genes and 886 DMRs overlapping promoter regions. We were able to recover 112 of the original 114 DMRs reported, even though (31) used hg19 as reference genome and only BSmooth for DMR calling. Most of these DMRs are outside of CpG islands (6,009 DMRs) and lose DNA methylation during differentiation (5,765 DMRs), which is consistent with the original findings (31). Excluding indexing of the reference genome, the whole analysis workflow from FASTQ files to annotated DMRs took 38.87 hours in total. A maximum memory usage of 216.07 GB was reached for bsseq DMR calling (Supplementary Material). bwa-meth alignment was the most time consuming step with a run time of 27.81 hours for a single sample using 16 threads.

For the blood/sperm dataset we used wg-blimp to determine soma-germ cell specific methylation differences. We found 410,247 DMRs ($\geq$ 4 CpGs, $\geq$ 0.3 absolute difference, $\geq 5\times$ coverage), of which 192,953 overlap with genes, 58,183 with promoters and 10,150 with CpG islands. As expected, the number of DMRs is much higher compared to the monocyte/macrophage dataset. Executing the whole workflow required 30.61 hours in total with a maximum memory usage of 208.83 GB.

Methylome segmentation. We applied wg-blimp to a single WGBS sequencing run of H1 embryonic stem cells (ESCs) (33, 34) (SRA accession SRP072141) to demonstrate segmentation using MethylSeekR. We chose H1 embryonic stem cells to compare our integrated segmentation to the results of the original MethylSeekR authors that, among other cell types, also analyzed H1 ESCs (12). FDR cutoff was set to 5% and methylation cutoff to 50% (default values). PMDs were not considered because alpha distribution values did not suggest PMD presence in this methylome (see Supplementary Material). In total, 18,930 UMRs and 31,748 LMRs were detected.

To evaluate segmentation results, we computed each segment center's distance to the nearest transcription start site (TSS) as reported by Ensembl (13). Figure 3 depicts separability of UMRs and LMRs with regards to TSS distances. As expected, most UMRs are in close proximity of a TSS, indicating activity in regulatory regions. Our results are in line with the original findings that also found no PMD presence and UMRs mostly overlapping promoter regions for H1 ESCs (12), despite differences in reference genomes and sequencing strategies.

Excluding reference genome indexing, executing the whole wg-blimp workflow from alignment to segmentation required 11.05 hours to complete. Alignment was the most time consuming step with a run time of 5.72 hours. Maximum memory usage of 168.76 GB was reached by MethylSeekR.

Conclusions

wg-blimp implements a WGBS analysis workflow, improving on previous WGBS pipelines by providing simple instal-

DRAFT

bioRxiv preprint doi: https://doi.org/10.1101/859900; this version posted November 30, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

lation and usage as well as a more extensive set of features. In addition to the analysis workflow wg-blimp includes a reactive R Shiny web interface for simplified inspection and sharing of results. wg-blimp is capable of producing coherent results, as demonstrated by analysing three publicly available datasets. We believe wg-blimp to be an apt alternative to previous WGBS analysis pipelines and hope to ease handling WGBS datasets for fellow researchers, and thus benefit the field of epigenetic research.

Availability and requirements

Project name: wg-blimp.
Project home page: https://github.com/MarWoes/wg-blimp
Operating system(s): UNIX.
Programming language: Python, R.
License: AGPL-3.0

Abbreviations

CGI: CpG island; DMR: differentially methylated region; ESC: embyronic stel cell; FDR: false-discovery rate LMR: low-methylated region; PMD: partially methylated domain; QC: quality control; TSS: transcription start site; UMR: unmethylated region; WGBS: whole genome bisulfite sequencing;

Acknowledgements

We thank Professor Martin Dugas for support.

Author's contributions

MW developed the software and prepared the final version of the manuscript. EL, BH and SL reviewed analysis output and provided feedback for subsequent improvement of the software. SL further provided novel data to test the pipeline with. SR and CS tested the software and provided feedback for subsequent improvement of the software. CS provided suggestions for best-practise WGBS analysis. All authors provided feedback on the manuscript and read and approved the final version of the manuscript.

Funding

This work was supported by the German Federal Ministry of Education and Research under the project Number 01KU1216 (Deutsches Epigenom Programm, DEEP) and the German Research Foundation (Clinical Research Unit CRU326 'Male Germ Cells': DFG grants TU 298/5-1 and HO 949/23-1 as well as DFG grant GR 1547/19-1).

Competing interests

The authors declare that they have no competing interests.

Bibliography

1. Christopher Schröder, Elsa Leitão, Stefan Wallner, Gerd Schmitz, Ludger Klein-Hitpass, Anupam Sinha, Karl-Heinz Jöckel, Stefanie Heilmann-Heimbach, Per Hoffmann, Markus M Nöthen, et al. Regions of common inter-individual dna methylation differences in human monocytes: genetic basis and potential function. *Epigenetics & Chromatin*, 10(1):37, Jul 2017. ISSN 1756-8935. doi: 10.1186/s13072-017-0144-2.
2. Brent S Pedersen, Kenneth Eyring, Subhajyoti De, Ivana V Yang, and David A Schwartz. Fast and accurate alignment of long bisulfite-seq reads. *arXiv preprint arXiv:1401.1129*, 2014.
3. Heng Li. Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv preprint arXiv:1303.3997*, 2013.
4. Broad Institute. Picard toolkit. `http://broadinstitute.github.io/picard/`. *Accessed 13 November 2019.*, 2019.
5. Devon P Ryan. MethylDackel. `https://github.com/dpryan79/methyldackel`. *Accessed 13 November 2019.*, 2019.
6. Simon Andrews. FastQC: A quality control tool for high throughput sequence data. `http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc`. *Accessed 13 November 2019.*, 2019.
7. Konstantin Okonechnikov, Ana Conesa, and Fernando García-Alcalde. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics*, 32(2):292–294, 2015.
8. Philip Ewels, Måns Magnusson, Sverker Lundin, and Max Käller. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19): 3047–3048, 06 2016. ISSN 1367-4803. doi: 10.1093/bioinformatics/btw354.
9. Frank Jühling, Helene Kretzmer, Stephan H Bernhart, Christian Otto, Peter F Stadler, and Steve Hoffmann. metilene: Fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. *Genome research*, 26(2):256–262, 2016.
10. Kasper D Hansen, Benjamin Langmead, and Rafael A Irizarry. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome biology*, 13(10):R83, 2012.
11. Christopher Schröder. *Bioinformatics from genetic variants to methylation.* PhD thesis, Technische Universität Dortmund, 2018.
12. Lukas Burger, Dimos Gaidatzis, Dirk Schübeler, and Michael B. Stadler. Identification of active regulatory regions from DNA methylation data. *Nucleic Acids Research*, 41(16): e155–e155, 07 2013. ISSN 0305-1048. doi: 10.1093/nar/gkt599.
13. Fiona Cunningham, Premanand Achuthan, Wasiu Akanni, James Allen, M Ridwan Amode, Irina M Armean, Ruth Bennett, Jyothish Bhai, Konstantinos Billis, Sanjay Boddu, et al. Ensembl 2019. *Nucleic acids research*, 47(D1):D745–D751, 2018.
14. Maximilian Haeussler, Ann S Zweig, Cath Tyner, Matthew L Speir, Kate R Rosenbloom, Brian J Raney, Christopher M Lee, Brian T Lee, Angie S Hinrichs, Jairo Navarro Gonzalez, et al. The UCSC Genome Browser database: 2019 update. *Nucleic Acids Research*, 47 (D1):D853–D858, 11 2018. ISSN 0305-1048. doi: 10.1093/nar/gky1095.
15. Brent S Pedersen and Aaron R Quinlan. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics*, 34(5):867–868, 10 2017. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx699.
16. Johannes Köster and Sven Rahmann. Snakemake–a scalable bioinformatics workflow engine. *Bioinformatics (Oxford, England)*, 28(19):2520–2522, October 2012. ISSN 1367-4811. doi: 10.1093/bioinformatics/bts480.
17. Björn Grüning, Ryan Dale, Andreas Sjödin, Brad A Chapman, Jillian Rowe, Christopher H Tomkins-Tinch, Renan Valieris, and Johannes Köster. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature Methods*, 15:475–476, 2018.
18. James T Robinson, Helga Thorvaldsdóttir, Aaron M Wenger, Ahmet Zehir, and Jill P Mesirov. Variant review with the integrative genomics viewer. *Cancer Research*, 77(21): e31–e34, 2017. ISSN 0008-5472. doi: 10.1158/0008-5472.CAN-17-0337.
19. Vivek Bhardwaj, Steffen Heyne, Katarzyna Sikora, Leily Rabbani, Michael Rauer, Fabian Kilpert, Andreas S Richter, Devon P Ryan, and Thomas Manke. snakePipes: facilitating flexible, scalable and integrative epigenomic analysis. *Bioinformatics*, 05 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz436. btz436.
20. H Kretzmer, C Otto, and S Hoffmann. BAT: Bisulfite analysis toolkit [version 1; peer review: 3 approved]. *F1000Research*, 6(1490), 2017. doi: 10.12688/f1000research.12302.1.
21. Osvaldo Graña, Hugo López-Fernández, Florentino Fdez-Riverola, David González Pisano, and Daniel Glez-Peña. Bicycle: a bioinformatics pipeline to analyze bisulfite sequencing data. *Bioinformatics*, 34(8):1414–1415, 12 2017. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx778.
22. Benjamin I. Laufer, Hyeyeon Hwang, Annie Vogel Ciernia, Charles E. Mordaunt, and Janine M. LaSalle. Whole genome bisulfite sequencing of down syndrome brain reveals regional dna hypermethylation and novel disorder insights. *Epigenetics*, 14(7):672–684, 2019. doi: 10.1080/15592294.2019.1609867. PMID: 31010359.
23. Keegan Korthauer, Sutirtha Chakraborty, Yuval Benjamini, and Rafael A Irizarry. Detection and accurate false discovery rate control of differentially methylated regions from whole genome bisulfite sequencing. *Biostatistics*, 20(3):367–383, 02 2018. ISSN 1465-4644. doi: 10.1093/biostatistics/kxy007.
24. Felix Krueger and Simon R. Andrews. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, 27(11):1571–1572, 04 2011. ISSN 1367-4803. doi: 10.1093/bioinformatics/btr167.
25. Carrie A Davis, Benjamin C Hitz, Cricket A Sloan, Esther T Chan, Jean M Davidson, Idan Gabdank, Jason A Hilton, Kriti Jain, Ulugbek K Baymuradov, Aditi K Narayanan, Kathrina C Onate, Keenan Graham, Stuart R Miyasato, Timothy R Dreszer, J Seth Strattan, Otto Jolanki, Forrest Y Tanaka, and J Michael Cherry. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Research*, 46(D1):D794–D801, 11 2017. ISSN 0305-1048. doi: 10.1093/nar/gkx1081.
26. Peiyong Jiang, Kun Sun, Fiona M. F. Lun, Andy M. Guo, Huating Wang, K. C. Allen Chan, Rossa W. K. Chiu, Y. M. Dennis Lo, and Hao Sun. Methy-pipe: An integrated bioinformatics pipeline for whole genome bisulfite sequencing data analysis. *PLOS ONE*, 9(6):1–11, 06 2014. doi: 10.1371/journal.pone.0100360.

DRAFT

bioRxiv preprint doi: https://doi.org/10.1101/859900; this version posted November 30, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

27. Phil Ewels, Rickard Hammarén, Alexander Peltzer, Patrick Hüther, Sven F., Paolo Di Tommaso, Maxime Garcia, Johannes Alneberg, Andreas Wilm, and Alessia. nf-core/methylseq: nf-core/methylseq version 1.3, February 2019.
28. Alexander Gosdschan, Katarzyna Wreczycka, Bren Osberg, and Ricardo Wurmus. PiGx. *https://github.com/BIMSBbioinfo/pigx_bsseq*. *Accessed 13 November 2019.*, 2019.
29. Paolo Di Tommaso, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. Nextflow enables reproducible computational workflows. *Nature Biotechnology*, 35(4):316, 2017.
30. Artem Tarasov, Albert J. Vilella, Edwin Cuppen, Isaac J. Nijman, and Pjotr Prins. Sambamba: fast processing of NGS alignment formats. *Bioinformatics*, 31(12):2032–2034, 02 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv098.
31. Stefan Wallner, Christopher Schröder, Elsa Leitão, Tea Berulava, Claudia Haak, Daniela Beißer, Sven Rahmann, Andreas S Richter, Thomas Manke, Ulrike Bönisch, et al. Epigenetic dynamics of monocyte-to-macrophage differentiation. *Epigenetics & Chromatin*, 9(1): 33, Jul 2016. ISSN 1756-8935. doi: 10.1186/s13072-016-0079-z.
32. Sandra Laurentino, Jann-Frederik Cremers, Bernhard Horsthemke, Frank Tuettelmann, Karin Czeloth, Michael Zitzmann, Eva Pohl, Sven Rahmann, Christopher Schroeder, Sven Berres, Klaus Redmann, Claudia Krallmann, Stefan Schlatt, Sabine Kliesch, and Joerg Gromoll. Healthy ageing men have normal reproductive function but display germline-specific molecular changes. *medRxiv*, 2019. doi: 10.1101/19006221.
33. Garrett Jenkinson, Elisabet Pujadas, John Goutsias, and Andrew P Feinberg. Potential energy landscapes identify the information-theoretic nature of the epigenome. *Nature genetics*, 49(5):719, 2017.
34. Thorsten M Schlaeger, Laurence Daheron, Thomas R Brickler, Samuel Entwisle, Karrie Chan, Amelia Cianci, Alexander DeVine, Andrew Ettenger, Kelly Fitzgerald, Michelle Godfrey, et al. A comparison of non-integrating reprogramming methods. *Nature biotechnology*, 33(1):58, 2015.

DRAFT

bioRxiv preprint doi: https://doi.org/10.1101/859900; this version posted November 30, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figures

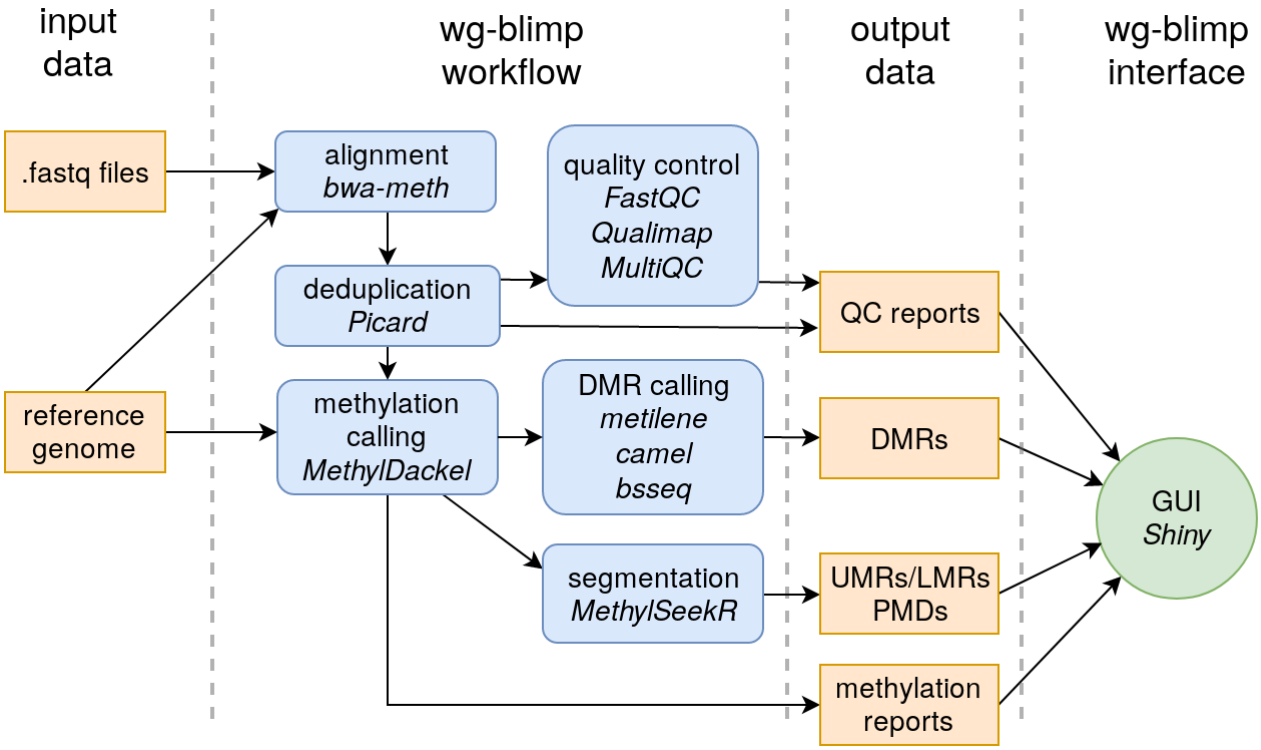

Fig. 1. *wg-blimp workflow overview.* Users only need to provide FASTQ files and a reference genome, and wg-blimp will perform alignment, deduplication, QC checks, DMR calling, segmentation and annotation. Once the pipeline results are available, users can inspect results using a web interface.

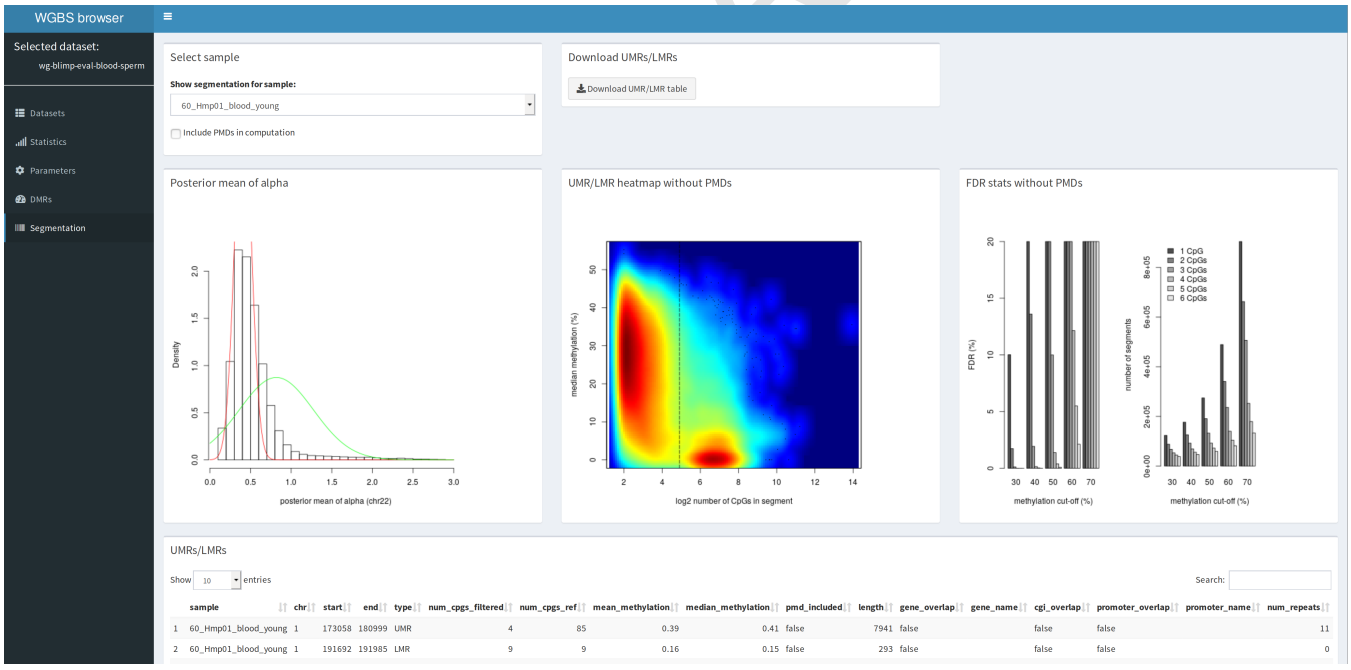

Fig. 2. *Segmentation tab of wg-blimp R Shiny GUI.* Once the analysis pipeline completes, users may load results into wg-blimp's R Shiny App. The tab depicted here displays MethylSeekR results and allows users to include or exclude PMD computation by toggling a single checkbox.

Tables

bioRxiv preprint doi: https://doi.org/10.1101/859900; this version posted November 30, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

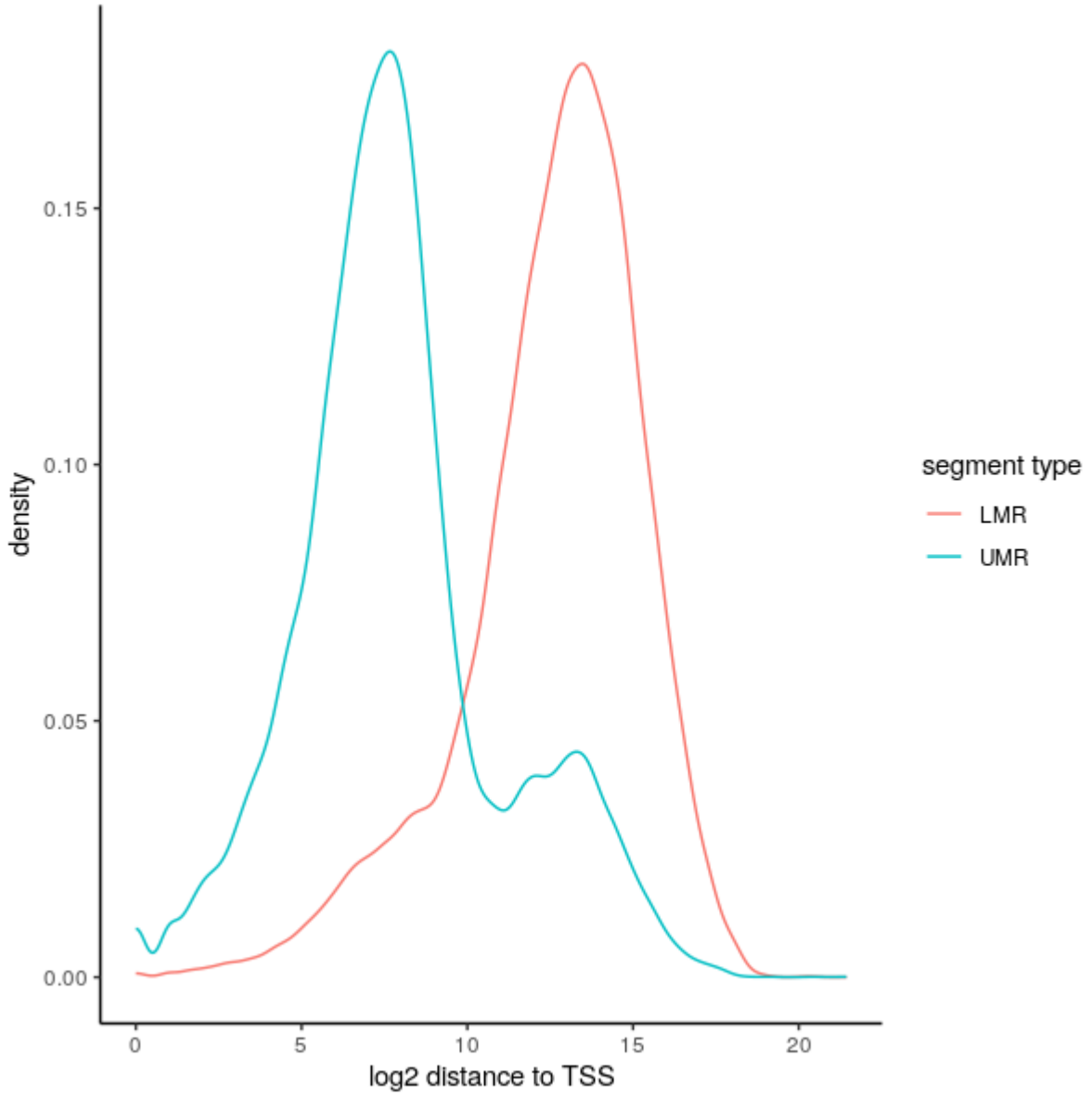

Fig. 3. *Distance from UMR/LMR centers to closest TSS for H1 ESCs.* UMRs/LMRs were automatically inferred using wg-blimp's MethylSeekR integration. UMRs and LMRs show a clear separation, with most UMRs being located in close proximity of TSSs.

bioRxiv preprint doi: https://doi.org/10.1101/859900; this version posted November 30, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 1. *Comparison of WGBS end-to-end pipelines.* Most pipelines use similar software for "standard" WGBS analysis tasks such as alignment or QC. wg-blimp improves on existing pipelines by providing a more comprehensive workflow as well as an interactive user interface.

pipeline	installation	workflow management	adapter trimming	alignment	methylation calling	quality control	DMR detection	segmentation	annotation
wg-blimp	Bioconda	Snakemake	/	bwa-meth	MethylDackel	MultiQC	bsseq camel metilene	MethylSeekR	CGI's genes repetitive regions
BAT	manual Docker	Perl/R/shell scripts	/	segemehl	haarz	BAT	metilene	/	arbitrary BED file
bicycle	manual Docker Live CD	Java Application			bicycle			/	arbitrary BED file
CpG_Me/DMRichR	manual	shell/R scripts	Trim Galore!	Bismark	Bismark	MultiQC	bsseq dmrseq	/	CGIs genes
ENCODE pipeline	DNAnexus		Trim Galore!	Bismark	Bismark	SAMtools Bismark	/	/	genes
Methy-Pipe	manual	Makefile	Methy-Pipe					/	genes
Nextflow methylseq (Bismark)	Nextflow		Trim Galore!	Bismark	Bismark	MultiQC	/	/	/
Nextflow methylseq (bwa-meth)	Nextflow		Trim Galore!	bwa-meth	MethylDackel	MultiQC	/	/	/
PiGx	GNU guix	Snakemake	Trim Galore!	Bismark	methylKit	MultiQC	methylKit		CGIs genes
snakePipes	Bioconda	Snakemake	Cutadapt Trim Galore! Fastp	bwa-meth	MethylDackel	MultiQC	dmrseq DSS metilene	/	genes

DRAFT