Genome analysis

P3BSseq: parallel processing pipeline software for automatic analysis of bisulfite sequencing data

Phuc-Loi Luu¹, Daniela Gerovska², Mikel Arrospide-Elgarresta², Sugoi Retegi-Carrión², Hans R. Schöler³,⁴ and Marcos J. Araúzo-Bravo¹,²,⁵,*

¹Computational Biology and Bioinformatics, Max Planck Institute for Molecular Biomedicine, 48149 Münster, Germany, ²Computational Biology and Systems Biomedicine, Biodonostia Health Research Institute, 20014 San Sebastián, Spain, ³Department of Cell and Developmental Biology, Max Planck Institute for Molecular Biomedicine, 48149 Münster, Germany, ⁴Medical Faculty, University of Münster, 48149 Münster, Germany and ⁵IKERBASQUE, Basque Foundation for Science, 48011 Bilbao, Spain

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on January 25, 2016; revised on September 14, 2016; accepted on September 29, 2016

Abstract

Motivation: Bisulfite sequencing (BSseq) processing is among the most cumbersome next generation sequencing (NGS) applications. Though some BSseq processing tools are available, they are scattered, require puzzling parameters and are running-time and memory-usage demanding.

Results: We developed P3BSseq, a parallel processing pipeline for fast, accurate and automatic analysis of BSseq reads that trims, aligns, annotates, records the intermediate results, performs bisulfite conversion quality assessment, generates BED methylome and report files following the NIH standards. P3BSseq outperforms the known BSseq mappers regarding running time, computer hardware requirements (processing power and memory use) and is optimized to process the upcoming, extended BSseq reads. We optimized the P3BSseq parameters for directional and non-directional libraries, and for single-end and paired-end reads of Whole Genome and Reduced Representation BSseq. P3BSseq is a user-friendly streamlined solution for BSseq upstream analysis, requiring only basic computer and NGS knowledge.

Availability and Implementation: P3BSseq binaries and documentation are available at: http://sourceforge.net/p/p3bsseq/wiki/Home/

Contact: mararabra@yahoo.co.uk

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

DNA methylation is a methyl-group addition to some DNA 5′ cytosines and is one of the most important epigenetic modifications, inducing gene silencing, cell differentiation and cancer. Next generation sequencing (NGS) of bisulfite sequences (BSseqs) generates BSseq reads (BSRs) that allow to study DNA methylation at nucleotide resolution. There are two approaches for BSseq data generation: Whole Genome BSseq (WGBS) and Reduced Representation BSseq (RRBS) (Meissner et al., 2005), widely used for diagnosis of cancer cells (Dehan et al., 2009) and analysis of epigenetics regulations during development (Smith and Meissner, 2013) or cellular reprogramming (Luu et al., 2013). Cost reduction increases the number of generated raw BSR data hosted in public databases such as the Sequence Read Archive (SRA) (Kodama et al., 2011). BSseq processing is cumbersome and requires a five step computational analysis (Krueger et al., 2012). (i) Raw BSRs are quality controlled (QC) with SolexaQA (Cox et al., 2010), FastQC (bioinformatics.babraham.ac.uk), SeqQC (genotypic.co.in),
P3BSseq (Wang et al., 2012) or SAMStat (Lassmann et al., 2011). (ii) Adapters and low quality bases are trimmed with Cutadapt (Martin, 2011), Adapter removal (Lindgreen, 2012), Trimmomatic (Lohse et al., 2012), trim_galore (bioinformatics.babraham.ac.uk), FASTX toolkit (hannonlab.cshl.edu/fastx_toolkit) or SolexaQA. (iii) Trimmed BSRs are aligned to the reference genome with Bismark (Krueger and Andrews, 2011), BRAT-BW (Harris et al., 2012), BSMAP (Xi and Li, 2009), RRBSMAP (Xi et al., 2012), BS Seeker (Chen et al., 2010), MethylCoder (Pedersen et al., 2011), RMAP (Smith et al., 2009) or ERNE-BS5 (Prezza et al., 2012); the most time-consuming step (Harris et al., 2012). (iv) Mapped reads are deduplicated with Picard (picard.sourceforge.net), Bismark or BRAT-BW to avoid PCR amplification bias affecting methylation calls accuracy in WGBS. (v) Methylation calling with BRAT-BW or Bismark.

The tools for performing each step are scattered, operating-system dependent, have incompatible input/output format between steps, and require specification of copious, not always well documented parameters. Some tools are only available for single-end reads (SERs), others are restricted to color space representations or directional libraries. Additionally, they entangle tracking of the results of each step, making the reports of the results non-compliant with the NIH Standards and Guidelines for Whole Genome Shotgun Bisulfite Sequencing (http://www.roadmapepigenomics.org/files/protocols/data/dna-methylation/MethylC-SeqStandards_FINAL.pdf). Moreover, the recent advances in NGS generate reads of length > 1000 bp, resulting longer processing time. Therefore, running time reduction is needed. Since the BSR alignment algorithms are optimized, the process parallelism by taking advantage of memory and multi-core CPUs is a good option to speed up the alignment (Becker et al., 2014).

We developed P3BSseq, parallel processing pipeline for BSR automatic analysis, which integrates existing tools for BSeq reads analysis with additional ad hoc tools for automatic file splitting and format conversion, download and annotation of reference genomes, results tracking, automatic trimming, QC of bisulfite conversion rates, alignment visualization and generation of the methylation ratios of each CG, CHH and CHG of the reference genome in the two DNA strands saving them in BED files. We parallelized each processing step and optimized the parameters. P3BSseq is a user-friendly tool for parallel processing of BSeq data, requiring a minimum number of input parameters and no prior knowledge of parallel processing.

2 System and methods

P3BSseq integrates the core tools: FastQC, trim_galore, Cutadapt, BRAT-BW, SAMtools (Li et al., 2009) with ad hoc developed Python scripts which implement the interfaces for these tools with parallel processes, format the compatible output files of each step, compute bisulfite conversion rates, depth of sequencing, keep track of the intermediate step results, separate and count mapped and unmapped reads. All processing stages, from the initial reference genome indexing to the final report generation, are fully automated, including the cumbersome trimming step, which is performed based on two QCs, base composition and GC content, following the Epigenesys protocol for QC, trimming and alignment of BSeq data (http://www.epigenesys.eu/images/stories/protocols/pdf/20120720103700_p57.pdf). The parameters of P3BSseq are optimized based on the Epigenesys protocol. The datasets used were taken from the SRA database (Supplementary Fig. S11).

3 Results

3.1 P3BSseq optimal trimming improves alignment by mapping uniquely twice more reads

Trimming of reads with low quality, adapters and kmers is essential prior alignment. To fine-tune trimming, we evaluated four schemas using SERs of ff-iPSC69, induced Pluripotent Cells (iPSCs) from human foreskin fibroblast, SRR068336 (Lister et al., 2011): (i) No trim; (ii) Trim reads with base Quality < 30, number of internal undetermined nucleotides Ns > 1, read Length < 24 (Trim QNL); (iii) Trim Adapter after QNL (Trim QNLAD); (iv) Trim Kmer plus QNLA (Trim QNLAK). After alignment of the trimmed reads to the genome, we examined the mapping efficiency. No trim maps uniquely the smallest number of reads, which can be explained by contamination of the reads with low quality bases, Ns, adapters or kmers (Fig. 2). The contamination induces low mapping efficiency and the
other trimming strategies duplicate the number of mapped-uniquely reads. Conversely, the number of reads mapped ambiguously increases with trimming, suggesting that over-trimming reduces mapping efficiency. Therefore, optimal trimming is needed. We implemented an optimized trimming strategy to guarantee efficient and high quality alignment.

3.2 P3BSseq automatically adapts kmers trimming

Trimming decision is taken automatically based on QC. Trimming has a triple positive effect: (i) Improves read quality, therefore increases the number of uniquely mapped reads. (ii) Cleans the reads before mapping from adapters, low quality or undetermined nucleotide contaminations; therefore increasing the accuracy of the methylation ratio. (iii) Removes large amounts of low quality and contaminated reads, therefore reduces alignment time. Trimming kmers (Trim QNLAK) improves alignment efficiency (Fig. 2), however three questions should be addressed before kmer trimming: Should kmers be trimmed before or after alignment? Where on the reads should the kmers be removed? And, how many kmers should be trimmed? To answer these, we conducted four experiments with SRRs of H1 human embryonic stem cells (ESCs), SRR019086 and SRR019088 (Lister et al., 2009). In the first three experiments, we trim QNLA, align the QNLA-trimmed reads, and remove from the unmapped reads the kmers that: (i) anchor at the 3’ and 5’ terminals of the unmapped reads; (ii) anchor at the 5’ terminals and ligate to the 3’ terminals of the unmapped reads (anchor means matching to the terminal sequence of the read; ligate means matching near the terminal sequence of the read); (iii) ligate to the 3’ and 5’ terminals of the unmapped reads. Finally, we realign the kmer-trimmed unmapped reads. In the fourth experiment, we (iv) trim QNL, remove from the QNLA-trimmed reads the kmers that anchor at the 3’ and 5’ read terminals, and align the QNL-trimmed reads. To analyze how many kmers should be trimmed, all experiments were conducted with different percentages of kmers, taken from the whole kmer list generated by the P3BSseq QC report. Experiment (i) has the best performance of mapped percentage in terms of total number of bases and total number of reads (Supplementary Fig. S2). Therefore, we set in P3BSseq a QNLA trimming strategy that trims the kmers with an exact anchor at 3’ and 5’ before and after the first QNLA-trimmed reads alignment, respectively. The mapping efficiency of all four trimming strategies stabilizes after only 1–10% of the listed kmers are used (Supplementary Fig. S2), therefore kmer trimming running time can be minimized by using a small percentage of listed kmers. Since realigning is computationally demanding, such step is implemented in P3BSseq as an additional option, whose activation is especially helpful when the number of mapped reads is small, such as in the HSF1 ESC dataset SRX021425 of Chodavarapu et al. (2010).

3.3 P3BSseq outperforms other software in processing time and computer-resource efficiency

For the most computationally-demanding task, alignment, we evaluated P3BSseq in terms of running time and resource utilization using 20 million 85 bp SRRs from ff-iPSC69 and SRR068329 (Lister et al., 2011). We compared P3BSseq against Bismark with Bowtie2 and Bowtie2 on a 12 multicores and 64 GB main memory computer. With 5 parallel processes using 5 cores, Bismark-Bowtie, Bismark-Bowtie2 and P3BSseq use 12.2, 16, 32.5 GBs of memory with a running time of 58.7, 418.3 and 9.1 minutes, respectively (Fig. 3). P3BSseq uses 2.7 times more memory than Bismark-Bowtie. Thus, P3BSseq takes maximum advantage of the available computer resources in an adaptive manner. Additionally, P3BSseq reduces running time over 6.4 times. Moreover, Bismark cannot run on less than 5 cores or with less than 12 GB memory, while P3BSseq can run even on a one-core computer with 6.5 GB due to the capability BRAT-BW to run with low computational resources. This allows P3BSseq to perform moderate-sized analyzes in short time on a laptop. Under the same settings, when comparing the performance of these aligners for the whole methylome (1000 million 85 bp reads) allowing one mismatch, it takes ~2 days, ~14 days and ~7.58 h with Bismark-Bowtie, Bismark-Bowtie2 and P3BSseq, respectively. Thus, the parallel implementation of P3BSseq significantly boosts alignment due to the adaptive way it maximizes the usage of available resources. P3BSseq has the same parallel adaptive feature for the other BSeq processing steps and different data types.

3.4 P3BSseq automatically generates a BSseq standard report with alignment assessment and visualization

We integrated in P3BSseq an option for automatic BSseq report generation, in compliance with the NIH guidelines, such as reporting the sequencing depth; read information and parameters before and after trimming and alignment; and the percentages of genomic, Cytosine and Cpg coverage. An example report of P3BSseq is provided as Supplementary information. Additionally, P3BSseq converts the mapped read file into SAM and BAM files to facilitate the examination of alignment regions of interest with SAMtools.
Parallel BSseq tool

( Supplementary Fig. S9). The BAM file can be used also for BSseq alignment visualization in the UCSC Genome Browser (genome. ucsc.edu). For assessment of bisulfite conversion treatment, three BSCRs are computed, plotted as histograms for CpG and non-CpG sites within the genomic DNA and for CpA sites within 200 bp promoter CpG islands (Supplementary Fig. S10), and included in the final report.

4 Conclusion

We designed P3BSseq with a ‘mini-max’ philosophy that minimizes the number of user-defined parameters thanks to the parameter optimization, and maximizes the use of the computer resources. P3BSseq identifies the available cores and main memory, and takes maximum advantage of them in function of the dataset requirements. P3BSseq adapts automatically to kmers trimming and thus improves the alignment by mapping uniquely twice more reads. It outperforms other software in processing time and computer-resource efficiency. It works with any combination of type of WGBS or RRBS methylationmes with single-end or paired-end reads and directional or non-directional libraries. P3BSseq is a fully automated parallel processing pipeline for BSseq. The automatized trimming and estimation of the number of processes makes P3BSseq easy to handle. The BSseq standard report of P3BSseq is a helpful resource for QC and downstream analysis. P3BSseq has applications in DNA methylelase analysis of development, cellular reprogramming, aging and diseases, especially cancer, data.

Our Supplementary information provides a complete guide of the experimental design, software installation and use instructions, a complete real-data example ( Supplementary Fig. S8), illustrating software installation, result storage and depiction (Supplementary Figs. S10–S12), and an example of how all steps profit from parallelization in terms of running time ( Supplementary Fig. S13).

Funding

This work was supported by grants DFG10/15 and DFG90/15 from Diputación Foral de Gipuzkoa, and grant 11H400016 from I+D+i National Plan 2013–2016 of Carlos III Health Institute, FEDER funds and IKERBASQUE, Basque Foundation for Science.

Conflict of Interest: none declared.

References


