Vol. 28 no. 4 2012, pages 587–588
doi:10.1093/bioinformatics/btr699

Advance Access publication December 22, 2011

MeQA: a pipeline for MeDIP-seq data quality assessment and analysis

J. Huang[1,2,*], V. Renault[2], J. Sengenès[3], N. Touleimat[2], S. Michel[4], M. Lathrop[2,3] and J. Tost[2,3]

[1]School of life science, Tongji University, 200092 Shanghai, China [2]Fondation Jean Dausset-CEPH, 75010 Paris, [3]CEA, Centre National de Génotypage, 91000 Evry, France and [4]Hannover Medical School, 30625 Hannover, Germany

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: We present a pipeline for the pre-processing, quality assessment, read distribution and methylation estimation for methylated DNA immunoprecipitation (MeDIP)-sequence datasets. This is the first MeDIP-seq-specific analytic pipeline that starts at the output of the sequencers. This pipeline will reduce the data analysis load on staff and allows the easy and straightforward analysis of sequencing data for DNA methylation. The pipeline integrates customized scripting and several existing tools, which can deal with both paired and single end data.

Availability: The package and extensive documentation, and comparison to public data is available at http://life.tongji.edu.cn/meqa/

Contact: jhuang@cephb.fr

Received on July 8, 2011; revised on December 4, 2011; accepted on December 18, 2011

1 INTRODUCTION

Methylated DNA immunoprecipitation (MeDIP) enables the rapid identification of genomic regions containing methylated cytosines. MeDIP, in combination with hybridization to high-resolution tiling microarrays or high-throughput sequencing (HTS) techniques, is a useful method for the identification of methylated CpG-rich sequences (Jacinto *et al*., 2008). Recently, several benchmark publications reported on the use of MeDIP-seq for genome-wide DNA methylation analysis and compared it to several others methods, for example, whole-genome bisulfite-sequencing (BS-seq) and methyl-binding protein-based enrichment of methylated sequences (MBD-seq) (Bock *et al*., 2010; Harris *et al*., 2010; Li *et al*., 2010). Though MeDIP is not substitute for BS-seq to obtain a methylome at single-nucleotide resolution, the generation of genome-wide data derived from MeDIP-seq provides a major tool for epigenetic studies in health and disease (Harris *et al*., 2010; Li *et al*., 2010). Further, MeDIP is specific for methylated cytosines and results are not confounded by the presence of hydroxymethylated cytosines unlike bisulfite-based methods. The main challenges resulting from the rapidly advancing technology development in DNA methylation analysis is now the computational analysis of the genome-wide sequencing data (Laird, 2010).

Two methods (Batman and MEDIPS) have been developed for MeDIP-seq data analysis (Chavez *et al*., 2010; Down *et al*., 2008). However, both of them do not include quality control of sequencing data or the read mapping. These methods require, therefore, considerable effort to prepare the data and run several other sequencing and quality control packages separately, increasing analysis time and potentially introducing processing errors.

To fill the gap between experimental throughput and processing speed, we developed the MeQA pipeline for pre-processing, data quality assessment and distribution of sequences reads and estimation of DNA methylation levels of MeDIP-seq datasets. The pipeline will also generate files for the UCSC browser. The pipeline presented here runs on the Unix/Linux platform and was written in the popular bioinformatics languages shell script, Python and R. It can be run locally on a single Linux/Unix or Mac server. We have tested this pipeline on our cluster with qsub and bsub commands with excellent performance. On a DELL 16 CPUs (each 2.67 GHz) and 32 GB memory computer server, it takes ∼20 h to run the entire pipeline for a mouse genome-wide DNA methylation estimate that contained 20 million 50-bp length single end reads. When several lanes are combined, which would provide a similar sequencing depth as HiSeq2000 data, it will take ∼30 h. We recommend running this pipeline on a computer with at least 16 GB memory.

2 METHODS

The execution of MeQA is straightforward and easy. After preparation of the configuration file, a simple command line calls the pipeline. We incorporated several existing computer packages into the MeQA pipeline. As installation of these packages requires some effort, we prepared a script to install these packages conveniently.

The pipeline is described in two parts. Each part can be run independently and be easily exchanged with other software if required. Part A performs the quality control of the DNA sequence information. First, MeQA provides a quick overview of sequence problems and data quality can be quickly assessed since it provides summary graphs. These results are exported to a pdf-based permanent report. The raw sequence is then aligned to a reference sequence genome using BWA (Li and Durbin, 2009). The alignments are saved in the standard SAM format, converted to BAM format and sorted with SAMtools (Li *et al.*, 2009). A shell script provides automatic download of references and index files from UCSC (Fujita *et al*., 2011), or a local directory can be provided for the alignment to a custom user provided reference. Quality of the sequence read mapping is accessed by SAMStat, and results are presented as a HTML report that includes unmapped, as well as poorly and accurately mapped reads, separately.

[*]To whom correspondence should be addressed.

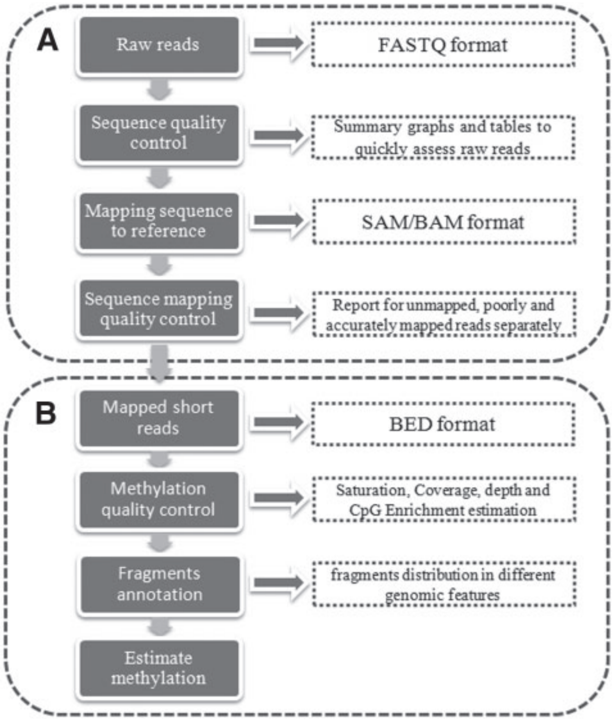
A
Raw reads
FASTQ format
Sequence quality control
Summary graphs and tables to quickly assess raw reads
Mapping sequence to reference
SAM/BAM format
Sequence mapping quality control
Report for unmapped, poorly and accurately mapped reads separately
B
Mapped short reads
BED format
Methylation quality control
Saturation, Coverage, depth and CpG Enrichment estimation
Fragments annotation
fragments distribution in different genomic features
Estimate methylation

Fig. 1. The MeQA analysis pipeline. (**A**) General sequence quality control. (**B**) Assessment of read distribution and estimation of DNA methylation levels.

Part B deals with the read distribution for different genomic features and estimates of the DNA methylation level. Methylated mapped region can be extracted as BED format by SAMtools and BEDtools (Quinlan and Hall, 2010). These BED files are supplied to MEDIPS that estimates the reproducibility of the genome-wide DNA methylation profile with respect to the total number of given short reads and to the size of the reference genome. MEDIPS also analyzes the coverage of genome-wide DNA sequence patterns (e.g. CpGs) by the given reads, and calculates a CpG enrichment factor as a quality control for the immunoprecipitation. For annotation of the methylated regions, CEAS (Shin *et al.*, 2009) is used to calculate the percentages of regions that correspond to: (i) promoters, (ii) bidirectional promoters, (iii) downstream of a gene and (iv) gene (3′UTRs, 5′UTRs, coding exons and introns). Lastly, MEDIPS summarizes the DNA methylation levels for genome-wide windows of a specified size or for user-defined regions of interest.

A flow chart of the analysis pipeline is shown in Figure 1.

3 DISCUSSION

The MeQA pipeline consists of two parts: the first one provides the general quality control of the sequence reads, and the second assesses the quality of the DNA methylation analysis experiment and provides comprehensive analysis. The first part could in principle also be applied to other sequence data, e.g. for quality assessment of RNA-seq data. The MeQA package allows users to obtain methylation estimates from raw sequence files with a single python function call. The main advantage of MeQA is its ease of use and its availability as open source, as all programs added to the pipeline are open source programs. Widely used file formats are used for each step of the pipeline, e.g. FASTQ, SAM/BAM and BED files. Each step of the pipeline can be replaced without compromising the workflow allowing users to update components, replace some packages and to extend and customize the pipeline for their needs.

This pipeline permits to estimate methylation levels in differentially methylated regions and genes using MEDIPS. Further functional analysis such as GO enrichment and KEGG pathway analysis for differentially methylated genes can be performed using additional Bioconductor packages.

ACKNOWLEDGEMENTS

The authors thank Prof. Howard Cann and Zhongwen CHANG for helpful discussions on this paper.

Funding: National Natural Science Foundation of China (grant no. 30900838); National Basic Research Program of China (973 Program No. 2010CB944904); European Union framework 7 integrated project READNA (contract no. HEALTH-F4-2008-201418).

Conflict of Interest: none declared.

REFERENCES

Bock,C. *et al.* (2010) Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nat. Biotechnol.*, **28**, 1106–1114.

Chavez,L. *et al.* (2010) Computational analysis of genome-wide DNA methylation during the differentiation of human embryonic stem cells along the endodermal lineage. *Genome Res.*, **20**, 1441–1450.

Down,T.A. *et al.* (2008) A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat. Biotechnol.*, **26**, 779–785.

Fujita,P.A. *et al.* (2011) The UCSC Genome Browser database: update 2011. *Nucleic Acids Res.*, **39**, D876–D882.

Harris,R.A. *et al.* (2010) Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat. Biotechnol.*, **28**, 1097–1105.

Jacinto,F.V. *et al.* (2008) Methyl-DNA immunoprecipitation (MeDIP): hunting down the DNA methylome. *Biotechniques*, **44**, 35, 37, 39 passim.

Laird,P.W. (2010) Principles and challenges of genomewide DNA methylation analysis. *Nat. Rev. Genet.*, **11**, 191–203.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li,N. *et al.* (2010) Whole genome DNA methylation analysis based on high throughput sequencing technology *Methods*, **52**, 203–212.

Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

Shin,H. *et al.* (2009) CEAS: cis-regulatory element annotation system. *Bioinformatics*, **25**, 2605–2606.