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

J. Huang1,2,*, V. Renault2, J. Sengenès3, N. Touleimat2, S. Michel4, M. Lathrop2,3 and J. Tost2,3

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

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@ceph.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 other 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.