MEDIPS: genome wide differential coverage analysis of sequencing data derived from DNA enrichment experiments

Matthias Lienhard¹,*, Christina Grimm¹,², Markus Morkel¹,³, Ralf Herwig¹ and Lukas Chavez¹,⁴

¹Max Planck Institute for Molecular Genetics, Ihnestrasse 63-73, 14195 Berlin, Germany
²Charité Universitätsmedizin Berlin, Department of Rheumatology and Clinical Immunology, Berlin, Germany
³Charité Universitätsmedizin Berlin, Laboratory of Molecular Tumour Pathology, Berlin, Germany
⁴La Jolla Institute for Allergy and Immunology, La Jolla, California 92037, USA

ABSTRACT

Motivation: DNA enrichment followed by sequencing is a versatile tool in molecular biology, with a wide variety of applications including genome-wide analysis of epigenetic marks and mechanisms. A common requirement of these diverse applications is a comparison of read coverage between experimental conditions. The amount of samples generated for such comparisons ranges from few replicates to hundreds of samples per condition for epigenome-wide association studies (EWAS). Consequently, there is an urgent need for software that allows for fast and simple processing and comparison of sequencing data derived from enriched DNA.

Results: Here, we present a major update of the R/Bioconductor package MEDIPS which allows for an arbitrary number of replicates per group and integrates sophisticated statistical methods for the detection of differential coverage between experimental conditions. Our approach can be applied to a diversity of quantitative sequencing data. In addition, our update adds novel functionality to MEDIPS, including correlation analysis between samples, and takes advantage of Bioconductor's annotation databases to facilitate annotation of specific genomic regions.

Availability: The latest version of MEDIPS is available as version 1.12.0 and part of Bioconductor 2.13. The package comes with a manual containing detailed description of its functionality and is available at http://www.bioconductor.org.

Contact: lienhard@molgen.mpg.de

1 INTRODUCTION

DNA enrichment methods are widely used for genome wide identification of many different kinds of epigenetic marks. These techniques include Chromatin-Immunoprecipitation (ChIP) for localizing transcription factor binding sites, or for revealing the genomic distribution of different histone modifications. Methylated DNA Immuno-Precipitation (MeDIP) (Weber et al., 2005) and methyl-CpG binding domain (MBD) protein capture (Serre et al., 2010) are similar techniques, but target the enrichment of DNA fragments containing methylated cytosines. Similarly, 5-hydroxymethylcytosines can be detected by antiserum specific to CMS (Pastor et al., 2011). It can be expected that further affinity methods will be developed for immunoprecipitation (IP) of known or novel kinds of epigenetic marks. In order to provide a general framework for efficient genome-wide differential coverage analysis of IP-sequencing data, we have improved the user-friendly MEDIPS package. In contrast to the previous version, the MEDIPS update is capable of processing an arbitrary number of replicates or samples per condition. Furthermore, MEDIPS now integrates an elaborated statistical framework developed for the digital nature of count data which includes a model for biological variation across replicates (Robinson et al., 2010a), and has greatly reduced runtime and memory requirements.

2 MEDIPS WORK FLOW

The MEDIPS package provides functions for the quality control and analysis of data derived from IP-seq samples. It starts with the aligned reads (typically bam files), and can be used for any genome of interest. Fig. 1 gives an overview of a typical workflow.

2.1 Preparation

In the first step, the alignment files (single- or paired-end) are imported, and the fragments overlapping previously specified genomic regions are counted. These regions can be either genome wide windows of regular width, or any given regions of interest (ROI). To control for PCR artifacts, MEDIPS optionally replaces reads with the same position and orientation by one representative.

2.2 Quality Control

The saturation analysis helps to verify, whether the given set of mapped reads is sufficient to generate a saturated and reproducible coverage profile of the reference genome. This is done by extrapolation of the correlation of subsets (see Fig. 1 C).
3 APPLICATION

To demonstrate the functionality of the MEDIPS package, we processed recently published MeDIP-seq data (Grimm et al., 2013), that was generated to assess genome-wide epigenetic changes in mouse intestinal adenoma. For this study, differential methylation was inferred for the sample groups by calculating Wilcoxon rank tests for the rpm values of each window. DMRs were determined by applying filters for p-values, minimal coverage, and ratios (Grimm et al., 2013).

Here, we process the same data but employing the presented MEDIPS package version 1.12.0. The commented R script, showing the function calls of this analysis, can be found in the supplementary information. From the 5 adenoma and 7 normal control mouse samples, 14M to 22M MeDIP-seq reads were uniquely mapped to the mm9 reference genome using bowtie (Langmead et al., 2009), of which about 93% remain after replacing reads with the same position and orientation by one representative. The saturation analysis indicates sufficient sequencing depth, and the CpG coverage indicates an effective MeDIP enrichment (see Fig. 1 C and Suppl. Fig. 1 and 2). Comparison of the normalized rams values with bisulphite validation showed a good overall correlation of 0.69 to 0.79 with a set of bisulphite validation assays previously performed by Grimm et al. (2013) on the same genomic samples (see Fig. 1 E and Suppl Fig. 3). The edgeR test for differentially methylated regions finds 51.722 DMRs (p - value < 0.01), which correspond to 0.5% of the genome. Correction for multiple testing leads to 110 regions at 10 % false discovery rate (FDR). Fig. 1 F shows the methylation logFC vs average log methylation (MA-plot). DMRs are depicted as orange points (p - value < 0.01) and red crosses (FDR < 0.1). The result table containing the DMRs can be found in Suppl. Tab 1. About 60% of the DMRs identified by Grimm et al. (2013) overlap with the DMRs identified by MEDIPS 1.12.

A detailed comparison between the two approaches can be found in the Supplementary Material.

Although the overall number of hypo- and hypermethylated regions is balanced, preferential hypermethylation was found in functionally important subgenomic regions, such as promoters and CpG islands. In particular, CpG-rich promoters showed a substantial enrichment of hyper- over hypomethylation (5:1; see Fig. 1 G). The identification of CpG-rich promoters as preferential targets for hypermethylation may provide important leads for further wet lab experiments. For instance, the analysis can be helpful to identify binding patterns of epigenetic modulator complexes and can be suited to identify candidate genes for epigenetic transcriptional silencing.

The processing of the aligned reads took about 90 minutes on an AMD Opteron 6380 2.5 GHz x computer, using 1 CPU core, and allocating a maximum of 20 GB RAM.
ACKNOWLEDGEMENT

The work was funded by the German Federal Ministry of Education and Research with the grant EPITREAT (No. 0316190A) and by the Max Planck Society with its International Research School program (IMPRS-CBSC). L.C. is supported by a Feodor Lynen postdoctoral Research Fellowship from the Alexander von Humboldt Foundation.

REFERENCES


