Bioinformatics Advance Access published January 16, 2012

IMA: An R package for high-throughput analysis of Illumina's 450K Infinium methylation data

Dan Wang[1], Li Yan[1], Qiang Hu[1], Lara E. Sucheston[2], Michael J. Higgins[3], Christine B. Ambrosone[2], Candace S. Johnson[4], Dominic J. Smiraglia[5] and Song Liu[1,*]

[1]Department of Biostatistics, [2]Cancer Prevention and Control, [3]Molecular and Cellular Biology, [4]Pharmacology and Therapeutics, [5]Cancer Genetics, Roswell Park Cancer Institute, Buffalo, New York 14263, USA.

Associate Editor: Janet Kelso

ABSTRACT

Summary: The Illumina Infinium HumanMethylation450 BeadChip is a newly designed high-density microarray for quantifying the methylation level of over 450,000 CpG sites within human genome. IMA (Illumina Methylation Analyzer) is a computational package designed to automate the pipeline for exploratory analysis and summarization of site-level and region-level methylation changes in epigenetic studies utilizing the 450K DNA methylation microarray. The pipeline loads the data from Illumina platform and provides user-customized functions commonly required to perform exploratory methylation analysis for individual sites as well as annotated regions.

Availability: IMA is implemented in the R language and is freely available from http://www.rforge.net/IMA.

Contact: song.liu@roswellpark.org

1 INTRODUCTION

As a major epigenetic modification, DNA methylation plays a vital role in transcriptional regulation and chromatin remodeling. The aberration of DNA methylation profile has been found to be associated with many human diseases including cancer (Portela and Esteller, 2010; P. A. Jones and Baylin, 2007). Use of DNA methylation microarray is a popular approach in studies to characterize the epigenetic landscape of human cells (Laird, 2010). Two widely used commercial platforms to perform methylation profiling are the GoldenGate Methylation Beadarray and Infinium HumanMethylation27 BeadChip provided by Illumina Inc. These two arrays quantitatively target 1,505 CpG loci covering around 800 genes and 27,578 CpG sites targeting around 14,000 genes, respectively. Since their release, many analytic methods have been developed to process and analyze the Illumina DNA methylation array data [for a recent summary, see (Siegmund, 2011).

Compared with previously released Illumina DNA methylation platforms, the recently launched Infinium HumanMethylation450 BeadChip represents a significant increase in the CpG site density for quantifying methylation events. At the gene level, the 450K microarray covers 99% of RefSeq genes with multiple sites in the annotated promoter (1,500 bp or 200 bp upstream of transcription start site), 5' UTR, 1st exon, gene body and 3' UTR. From the CpG context, it covers 96% of CpG islands with multiple sites in the annotated CpG Islands, shores (regions flanking island) and shelves (regions flanking shores) (Bibikova et al., 2011). While the role of DNA methylation in promoter and/or CpG island regions is long been appreciated, the importance of DNA methylation in gene body or shore regions for transcription regulation and tumor initialization has recently come to attention (R. A. Irizarry et al., 2009; Maunakea et al., 2010). The significantly increased coverage makes 450K microarray a powerful platform for exploring methylation profile in these annotated regions. As each targeted region contains at least one CpG site, treating the region as a unit in the differential methylation analysis might help identify regions with consistently coordinate methylation changes. From a statistical point of view, region-based differential methylation analysis will reduce the burden of multiple comparisons and increase the power to catch differentially methylated regions associated with the phenotypes of interest. To this end, we have developed a pipeline, IMA, for automatic site-level and region-level methylation analysis using the 450K microarray. While the pipeline is primarily designed as an automatic tool for exploratory analysis and summarization, it is flexible for users to tailor within R statistical computing and graphics environment for their specific needs.

2 DESCRIPTIONS

IMA is implemented in R and can be run on any platform with an existing R and Bioconductor installation. The user can run the pipeline with default settings or specify optional routes in the parameter file. An overview of the IMA pipeline is provided below:

Preprocessing: IMA takes as input the beta values representing the methylation levels of individual sites reported by Illumina BeadStudio or GenomeStudio software. It allows user to choose several filtering steps or modify filtering criteria for specific quality control purposes. By default, IMA will filter out loci with missing β value, from the X chromosome or with median detection P-value greater than 0.05. As probe containing SNP(s) at/near the targeted CpG site might not be sufficient to measure DNA methylation level (but rather genomic variation), users can choose to filter out the loci whose methylation levels are measured by probes containing SNP(s) at/near the targeted CpG site. The option for sample level quality control is also provided (B. C. Christensen, A. A. Smith, et al., 2011). Although the raw β values will be analyzed as recommended by Illumina, the user can choose Arcsine square root

[*]To whom correspondence should be addressed.

transformation when modeling the methylation level as the response in a linear model (Rocke, 1993; C. J. Marsit et al., 2011). Logit transformation is also available as an option (Kuan et al., 2010). The default setting of IMA is that no normalization will be performed, and quantile normalization is available as an alternative preprocessing option. It has been shown that quantile normalization is not sufficient for removing all unwanted technical variation across samples (Teschendorff et al., 2009). The development of normalization strategy for DNA methylation study is an active area of ongoing research (Aryee et al., 2011).

Methylation Index Calculation: The promoter, 5' UTR, 1st exon, gene body and 3' UTR are gene-based regions. The CpG island and its surrounding shore and shelve regions are not necessary gene-based, depending on their distance to the nearest genes. For each specific region (*e.g.*, 1^{st} exon), IMA will collect the loci within it and derive an index of overall region methylation value. Currently, there are three different index metrics implemented in IMA: mean, median, and Tukey's Biweight robust average. By default, the median beta value will be used as the region's methylation index for further analysis.

Differential Methylation Analysis: For each specific region, Wilcoxon rank-sum test (default), Student's t-test and empirical Bayes statistics are available for inference in differential testing. General linear models are available as an option to infer methylation change associated with continuous covariate (*e.g.*, age), as well as to adjust confounding factors (*e.g.*, batch). A variety of multiple testing correction algorithms are available, including stringent Bonferroni correction and widely used false discovery rate control. Users can specify the significance criteria in the parameter file. The same statistical inference and multiple test correction procedures described above can also be applied to each single site to obtain site-level differential methylation inference.

Output: Detailed output files are provided for each of the three modules above. For the preprocessing module, the output contains a matrix of methylation value for qualified loci across qualified samples. For the methylation index calculation module, there is a matrix of methylation index across the samples for each region category of interest (*e.g.*, South Shore). For the differential methylation analysis module, the differential methylation values (*e.g.*, delta β) together with both raw and adjusted P-values of each region (or site) of interest will be provided.

3 DISCUSSION

The major differences between IMA and existing R packages for Infinium methylation analysis (Du et al., 2008) are that IMA provides a pipeline which automates the tasks commonly required for the exploratory analysis and summarization of 450K DNA methylation data at both site-level and region-level. The package makes use of Illumina methylation annotation for region definition, as well as several Bioconductor packages for various preprocessing and differential testing steps (Gentleman et al., 2004).

Instead of providing recommendations about which specific analysis method should be used, the main purpose of developing the IMA package is to provide a range of commonly used DNA methylation microarray analysis options for users to choose for their exploratory analysis and summarization in an automatic way. Written in open source R environment, it provides the flexibility for users to adopt, extend and customize the functionality for their specific needs. It can be used as an automatic pipeline of methylation level index and differential analysis for downstream functional exploration and hypothesis generation. For example, the matrix of methylation index for shore regions produced by IMA can be used as the input for model-based clustering (Houseman, B. Christensen, et al., 2008) to identify clustered shores associated with the phenotype of interest.

Analytic methods for DNA methylation microarray analysis are still under rapid developments (Siegmund, 2011; Laird, 2010). Future development of IMA package will include the extension of its functionality by incorporating the latest preprocessing and differential analysis methods. For example, options will be added to filter out defective bead types (*e.g.*, mismatched or non-uniquely aligned probes) detected from systematic re-annotation efforts (Barbosa-Morais et al., 2010).

Funding: National Institute of Health (grant R01-CA133264 to C.A. and M.H.; R01-CA095045 to C.J.)

Conflict of Interest: none declared.

REFERENCES

Aryee,M.J. et al. (2011) Accurate genome-scale percentage DNA methylation estimates from microarray data. *Biostatistics*, **12**, 197 -210.

Barbosa-Morais,N.L. et al. (2010) A re-annotation pipeline for Illumina BeadArrays: improving the interpretation of gene expression data. *Nucleic Acids Res*, **38**, e17.

Bibikova,M. et al. (2011) High density DNA methylation array with single CpG site resolution. *Genomics*, **98**, 288-295.

Christensen,B.C., Smith,A.A., et al. (2011) DNA Methylation, Isocitrate Dehydrogenase Mutation, and Survival in Glioma. *Journal of the National Cancer Institute*, **103**, 143 -153.

Du,P. et al. (2008) lumi: a pipeline for processing Illumina microarray. *Bioinformatics*, **24**, 1547 -1548.

Gentleman,R.C. et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, **5**, R80-R80.

Houseman,E.A., Christensen,B., et al. (2008) Model-based clustering of DNA methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions. *BMC Bioinformatics*, **9**, 365.

Irizarry,R.A. et al. (2009) The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat Genet*, **41**, 178-186.

Jones,P.A. and Baylin,S.B. (2007) The Epigenomics of Cancer. *Cell*, **128**, 683-692.

Kuan,P.F. et al. (2010) A statistical framework for Illumina DNA methylation arrays. *Bioinformatics*, **26**, 2849 -2855.

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

Marsit,C.J. et al. (2011) DNA Methylation Array Analysis Identifies Profiles of Blood-Derived DNA Methylation Associated With Bladder Cancer. *Journal of Clinical Oncology*, **29**, 1133 -1139.

Maunakea,A.K. et al. (2010) Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature*, **466**, 253-257.

Portela,A. and Esteller,M. (2010) Epigenetic modifications and human disease. *Nat Biotech*, **28**, 1057-1068.

Rocke,D.M. (1993) On the beta transformation family. *Technometrics*, **35**, 72–81.

Siegmund,K.D. (2011) Statistical approaches for the analysis of DNA methylation microarray data. *Human Genetics*, **129**, 585-595.

Teschendorff,A.E. et al. (2009) An Epigenetic Signature in Peripheral Blood Predicts Active Ovarian Cancer. *PLoS One*, **4**.