Genome analysis

**MIRA: an R package for DNA methylation-based inference of regulatory activity**

**John T. Lawson**¹,², **Eleni M. Tomazou**³, **Christoph Bock**⁴ and **Nathan C. Sheffield**²,*

¹Department of Biomedical Engineering, ²Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA, ³Children’s Cancer Research Institute, St. Anna’s Kinderkrebshausung, 1090 Vienna, Austria and ⁴CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, 1090 Vienna, Austria

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on August 25, 2017; revised on November 3, 2017; editorial decision on February 10, 2018; accepted on February 27, 2018

**Abstract**

**Summary:** DNA methylation contains information about the regulatory state of the cell. MIRA aggregates genome-scale DNA methylation data into a DNA methylation profile for a given region set with shared biological annotation. Using this profile, MIRA infers and scores the collective regulatory activity for the region set. MIRA facilitates regulatory analysis in situations where classical regulatory assays would be difficult and allows public sources of region sets to be leveraged for novel insight into the regulatory state of DNA methylation datasets.

**Availability and implementation:** [http://bioconductor.org/packages/MIRA](http://bioconductor.org/packages/MIRA).

**Contact:** nsheffield@virginia.edu

DNA methylation interacts with other regulatory features to control gene expression (Stadler *et al.*, 2011). The connection between methylation and transcription factor (TF) binding goes both ways: TF binding affects and is affected by DNA methylation (Zhu *et al.*, 2016), making it difficult to infer the causative factor; nevertheless, independent of directionality, the inverse correlation between DNA methylation and gene expression indicates that regulatory information can be derived from DNA methylation data.

Multiple approaches have been used to relate DNA methylation to regulatory activity; for example, correlating differential methylation with expression of nearby genes (Yao *et al.*, 2015), or testing enrichment of TFs in differentially methylated regions (Wijetunga *et al.*, 2017; Yao *et al.*, 2015). These approaches are limited by arbitrary thresholds for differential methylation and do not make full use of genome-wide data. Also, factors other than DNA methylation levels, such as the shape of the DNA methylation profile around a site, may be important to the site’s activity (Kapourani and Sanguinetti, 2016).

We recently introduced and validated a novel method called MIRA (Methylation-based Inference of Regulatory Activity), which takes advantage of genome-scale DNA methylation data to assess regulatory activity (Sheffield *et al.*, 2017). We now present the MIRA R package which enhances this method and makes it broadly available.

MIRA requires two inputs: (i) single-nucleotide-resolution DNA methylation data; and (ii) a set of genomic regions (Fig. 1A). The DNA methylation data could come from sources such as whole genome or reduced representation bisulfite sequencing (WGBS or RRBS), or microarrays. MIRA has been successfully tested with coverage as low as 450k array data. Genomic regions can be derived from sequencing assays such as ChIP-seq, DNase-seq, or ATAC-seq. Many region sets are publicly available through large-scale genomics projects and may be conveniently accessed through R packages like LOLA (Sheffield and Bock, 2016).

Using these two inputs, MIRA aggregates the DNA methylation of individual CpGs to create a summary profile through several steps: First, each region (Fig. 1B) is split into n bins. Second, the DNA methylation level (0–100%) within a bin is averaged (Fig. 1C). Third, the regions are aggregated into a single summary profile by averaging the DNA methylation levels of each bin across all regions (Fig. 1D). MIRA thus creates a ‘meta-region profile’ that provides general information about the activity of that region type across the genome. Through aggregation, MIRA handles sparse DNA methylation data well. This makes MIRA well suited for low-coverage bisulfite sequencing (e.g. Farlik *et al.*, 2015).

Once an aggregate profile is constructed (Fig. 1E), it is scored to quantify the regulatory activity (Fig. 1F). MIRA assumes that
genomic regions with lower DNA methylation levels have higher regulatory activity and gives a score based on the depth of the ‘dip’ in the middle of the ‘meta-region profile’. MIRA automatically determines the location of the edges of the dip and calculates the score as the natural logarithm of the ratio between the DNA methylation level of the edges of the dip and the DNA methylation level of the center of the dip (Fig. 1F). The score reduces the DNA methylation profile to a number, which predicts the region set’s aggregate regulatory activity. MIRA scores can be compared between samples to identify regulatory differences.

MIRA supports a variety of applications depending on the context and what type of region set is used. For example, MIRA can be used to compare the chromatin states of different types of cells (Sheffield et al., 2017). MIRA makes analysis of regulatory activity possible in cases where it would otherwise be infeasible. When sample amount or quality would not allow ATAC-seq or ChIP-seq but DNA methylation data can be obtained, regulatory analysis can be done with MIRA using existing ATAC-seq or ChIP-seq data (e.g. from a database). MIRA is also valuable for cases where it would be impractical in terms of time or cost to perform traditional regulatory assays, such as for large-scale cohort studies. The MIRA R package can be accessed via Bioconductor, and comes with multiple vignettes demonstrating how to apply it to biological data. MIRA provides a novel tool to enhance analysis of DNA methylation and leverage existing data from regulatory assays to gain new regulatory insights.

**Funding**

This work has been supported by the University of Virginia and by an NIH training grant to J.L. (NLM; ST32LM012416).

**Conflict of Interest** none declared.

**References**

Farlik,M. et al. (2015) Single-cell DNA methyolme sequencing and bioinformatic inference of epigenomic cell-state dynamics. Cell Rep., 10, 1386–1397.

Kapourani,C.-A. and Sangunnetti,G. (2016) Higher order methylation features for clustering and prediction in epigenomic studies. Bioinformatics, 32, i405–i412.

Sheffield,N.C. and Bock,C. (2016) LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. Bioinformatics, 32, 587–589.

Sheffield,N.C. et al. (2017) DNA methylation heterogeneity defines a disease spectrum in ewing sarcoma. Nature Medicine, 23, 386–395.

Stadler,M.B. et al. (2011) DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature, 480, 490–495.

Wijetunga,N.A. et al. (2017) A pre-neoplastic epigenetic field defect in HCV-infected liver at transcription factor binding sites and polycomb targets. Oncogene, 36, 2030–2044.

Yao,L. et al. (2015) Inferring regulatory element landscapes and transcription factor networks from cancer methylomes. Genome Biol., 16, 105.

Zhu,H. et al. (2016) Transcription factors as readers and effectors of DNA methylation. Nat. Rev. Genet., 17, 551–565.