BioVLAB-MMIA-NGS: MicroRNA-mRNA Integrated Analysis using High Throughput Sequencing Data

Heejoon Chae [1], Sungmin Rhee [2], Kenneth P. Nephew,[3] and Sun Kim [2*]

[1]School of Informatics and Computing, Indiana University Bloomington, IN 47404, USA, [2]School of Computer Science and Engineering, Seoul National University, Seoul, Korea, [3] Indiana University School of Medicine, Indianapolis, IN 46202, USA.

ABSTRACT

Motivation: It is now well established that microRNAs (miRNAs) play a critical role in regulating gene expression in a sequence specific manner and genome-wide efforts are underway to predict known and novel miRNA targets. However, the integrated miRNA-mRNA analysis remains a major computational challenge, requiring powerful informatics systems and bioinformatics expertise.

Results: The objective of this study was to modify our widely recognized web server for the integrated mRNA-miRNA analysis (MMIA) and its subsequent deployment on the Amazon cloud (BioVLAB-MMIA) in order to be compatible with high throughput platforms, including next generation sequencing data (e.g., RNA-seq). We developed a new version called BioVLAB-MMIA-NGS, deployed on both Amazon cloud and on a high performance, publically available server called MAHA. By utilizing next generation sequencing (NGS) data and integrating various bioinformatics tools and databases, BioVLAB-MMIA-NGS offers several advantages. First, sequencing data is more accurate than array-based methods for determining miRNA expression levels. Second, potential novel miRNAs can be detected by using various computational methods for characterizing miRNAs. Third, because miRNA-mediated gene regulation is due to hybridization of a miRNA to its target mRNA, sequencing data can be used to identify many-to-many relationship between miRNAs and target genes with high accuracy.

Availability: http://epigenomics.snu.ac.kr/biovlab_mmia_ngs/

Contact: sunkim.bioinfo@snu.ac.kr, heechae@cs.indiana.edu

1 INTRODUCTION

MicroRNAs are small (19-24nt) single stranded non-coding RNAs that regulate gene expression by specific targeting mechanism to mRNA molecules via complementary sequence pairing. Due to their critical implication in post-transcriptional regulation and impact on developmental process, a number of miRNA-mRNA integrated analysis tools have been developed. MAGIA (Sales *et al.*, 2010) uses miRNA-mRNA expression profiles matrices as input and provides gene set analysis and miRNA target prediction. DIANA-mirExTra (Alexiou *et al.*, 2010) accepts gene sets and computationally compares miRNA associated motifs. miRGator (Cho *et al.*, 2013) provides pre-compiled public resources with browser interface to navigate data. However, several limitation exist for these and other existing tools include: (1) support only microarray or sequencing data but not both; (2) require pre-processing or manual data compiling step; (3) demand cumbersome installation procedures with inter-dependent tools and databases; (4) run on limited computational resources that are not capable of handling large data sets; Here we present the NGS data-compatible BioVLAB-MMIA-NGS, an updated version of our array-based miRNA-mRNA integrated analysis system MMIA (Nam *et al.*, 2009)

2 APPROACH

In order to perform integrated analyses between miRNA and mRNA using NGS data, we completely redesigned MMIA web server. BioVLAB-MMIA-NGS utilizes sequencing data to directly measure miRNA and mRNA expression levels on a genome-scale and accurately detect changes in quantity based on read count. The system accepts raw sequencing data as input without requiring any pre-processing steps. By utilizing RNA-seq and small RNA-seq data, not only can BioVLAB-MMIA-NGS predict novel miRNA candidates, it can also extract new information about miRNA-targeting of intragenic regions, exons, and introns as well as 3 UTRs, which have recently been used in the integrated analysis in plants (Meng *et al.*, 2013). Furthermore, to completely remove the burden of manually installing additional analysis tools, BioVLAB-MMIA-NGS adopts Java Web Start (JAWS), a single click JAVA application deployment technology. Moreover, to support large NGS data analysis, the pre-processing and computational processes in BioVLAB-MMIA-NGS were moved data to Amazon cloud and peta-scale super computing system called MAHA (http://www.etri.re.kr/eng/res/res_06020402.etri).

3 FEATURES

Workflow: In BioVLAB-MMIA-NGS, the integrated analysis workflow begins with extracting differentially and(or) significantly expressed miRNAs (DEmiR) as in our previous published methodology (Xin *et al.*, 2009). To identify miRNAs and their expression level, we adopted the miRDeep (Friedländer *et al.*, 2011) pipeline. The mapper module of the miRDeep package aligns

[*]to whom correspondence should be addressed

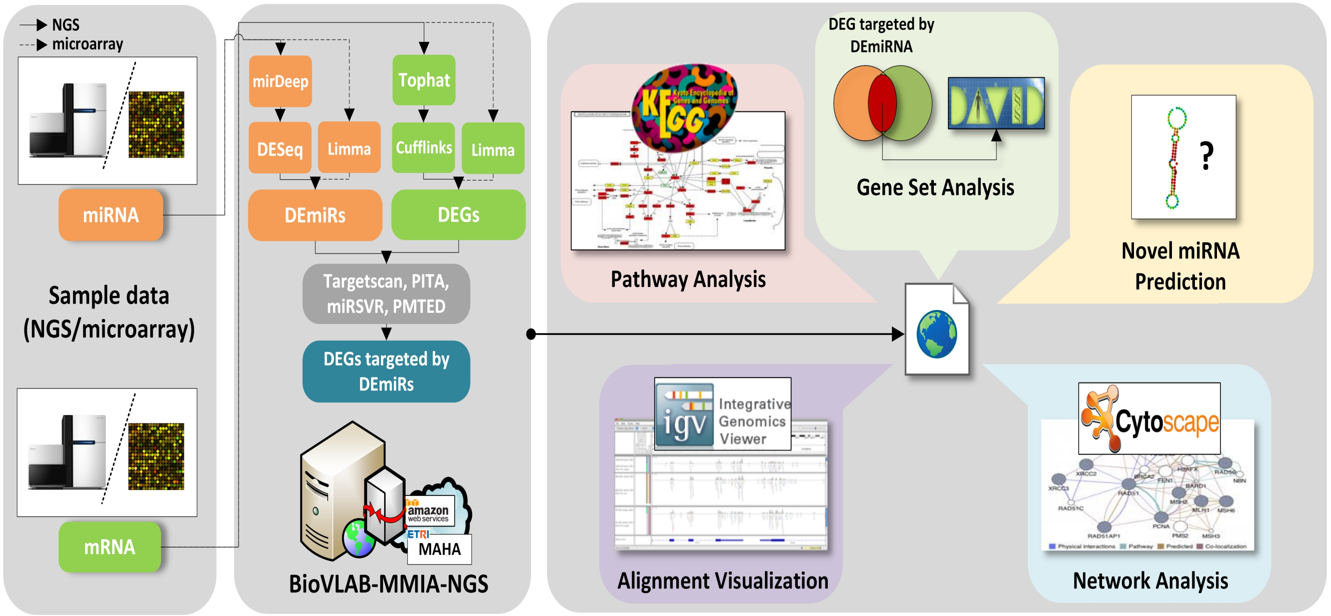

Fig. 1. BioVLAB-MMIA-NGS accepts NGS/ microarray data as input and extracts DEGs targeted by DEmiRs. Through the analysis pipelines, the system produces results of pathway analysis, gene set analysis, novel miRNA prediction, alignment visualization, and constructed miRNA-mRNA target network.

raw sequencing reads to known miRNAs in miRBase (Kozomara and Griffiths-Jones, 2010) and the quantifier module measures expression levels based on read counts. Quality control and adaptor clipping processes are performed if necessary. Significance of expressed miRNAs is tested and visualized by DESeq (Anders and Huber., 2010) based on read count measures. Differentially expressed mRNAs/genes (DEG)s are extracted by using the Tophat-Cufflinks pipeline (Trapnell *et al.*, 2012) with junction aligning based on the RPKM measure. Statistical significance is visualized by cummeRbund (Goff *et al.*, 2012). For microarray data, DEGs and DEmiRs are detected by Limma package (Smyth, 2005). Once DEGs and DEmiRs are extracted, the next step, miRNA-mRNA combined analysis, is performed. For the combined analysis, we utilized miRNA target prediction algorithms/ databases, TargetScan (Lewis *et al.*, 2005), PITA (Kertesz *et al.*, 2007), miRSVR (Betel *et al.*, 2010), and PMTED (Sun *et al.*, 2013), as well as negative correlations between miRNAs and mRNAs, to extract DEGs targeted by DEmiRs. Gene set analysis is performed using extracted DEGs. Gene sets are automatically submitted to DAVID (Huang *et al.*, 2007) to provide functional annotation and clustering, BioCarta and KEGG (Kanehisa and Goto, 2000) pathways mapping, and disease association. Figure 1 shows BioVLAB-MMIA-NGS workflow.

User Interface: BioVLAB-MMIA-NGS keeps the user-friendly web based interface for sample information, analysis options and parameters, and computing nodes. To provide extended analysis interface, we integrated IGV (Thorvaldsdttir *et al.*, 2013) for visualizing the alignment results with zoom-in/out functionality with annotation tracks and Cytoscape (Shannon *et al.*, 2003) for illustrating identified miRNA-mRNA target networks. By using JAWS, IGV and Cytoscape automatically visualize the results; manual installation and data handling processes are not required. A webpage summarizing all the results helps users view and further investigate the data

System: The information system architecture has also adopted several important changes. The analysis begins from the web interface and the graphical workflow composer shows progress status. In addition, BioVLAB, the cloud infrastructure used in our previous system, has been completely rebuilt using Apache Airavata (http://airavata.apache.org/), generating a highly flexible and extensible three-layered BioVLAB-MMIA-NGS architecture. Moreover, BioVLAB-MMIA-NGS now supports human, mouse, and rice genomes.

ACKNOWLEDGEMENT

Funding: This work was supported by the IT R&D program of MSIP/KEIT. [10038768, The Development of Supercomputing System for the Genome Analysis], by Next- Generation BioGreen 21 Program [PJ009037022012]; Rural Development Administration, Republic of Korea, and by Next-Generation Information Computing Development Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT & Future Planning (No.NRF-2012M3C4A7033341).

Conflict of Interest: none declared.

REFERENCES

Sales, G. *et al*. (2010) MAGIA, a web-based tool for miRNA and Genes Integrated Analysis. *Nucleic Acids Res.*, doi:10.1093/nar/gkq423.

Alexiou, P. *et al*. (2009) The DIANA-mirExTra Web Server: From Gene Expression Data to MicroRNA Function, *PLoS One*, doi:10.1371/journal.pone.0009171.

Nam, S. *et al*. (2009) MicroRNA and mRNA integrated analysis (MMIA): a web tool for examining biological functions of microRNA expression, *Nucleic Acids Res.*, doi:10.1093/nar/gkp294.

Cho, S. *et al*. (2013) miRGator v3.0: a microRNA portal for deep sequencing, expression profiling and mRNA targeting, *Nucleic Acids Res.*, doi:10.1093/nar/gks1168.

Xin, F. *et al*. (2009) Computational analysis of miRNA profiles and their target genes suggests significant involvement in breast cancer antiestrogen resistance, *Bioinformatics*, doi: 10.1093/bioinformatics/btn64.

Trapnell, C. *et al*. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks *Nature Protoc.*, doi:10.1038/nprot.2012.016.

Friedländer, M. *et al*. (2011) miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades *Nucleic Acids Res.*, doi: 10.1093/nar/gkr688.

Anders, A., and Huber, W. (2010) Differential expression analysis for sequence count data*Genome Biology*, doi: 10.1186/gb-2010-11-10-r106.

Smyth, GK. (2010) Limma: linear models for microarray data. In: Gentleman R, et al., editors. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, New York : Springer; 2005. p. 397-420.

Goff, L. *et al*. (2012) Analysis, exploration, manipulation, and visualization of Cufflinks high-throughput sequencing data. R package version 2.4.1.

Lewis, B.P. *et al*. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets.*Cell*, doi:10.1016/j.cell.2004.12.035.

Betel, D. *et al*. (2010) Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol.*, doi:10.1186/gb-2010-11-8-r90.

Sun, X. *et al*. (2013) PMTED: a plant microRNA target expression database. *BMC Bioinformatics*, doi:10.1186/1471-2105-14-174.

Kertesz M. *et al*. (2007) The role of site accessibility in microRNA target recognition. *Nature Genet.*, doi:10.1038/ng2135

Kozomara, A., and Griffiths-Jones, S.(2010) miRBase: integrating microRNA annotation and deep-sequencing data.*Nucleic Acids Res.*, doi: 10.1093/nar/gkq1027.

Huang, D. W. *et al*. (2007) DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nature Protoc.*,doi:10.1038/nprot.2008.211.

Kanehisa, M. ,and Goto S. *et al*. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, doi: 10.1093/nar/28.1.27.

Thorvaldsdttir, H. *et al*. (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Bioinformatics*, doi: 10.1093/bib/bbs017.

Shannon, P. *et al*. (2003) Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.*, doi: 10.1101/gr.1239303.

Meng Y. *et al*. (2013) Introns targeted by plant microRNAs: a possible novel mechanism of gene regulation *Rice*, doi: 10.1186/1939-8433-6-8.