BioVLAB-MMIA: A Reconfigurable Cloud Computing Environment for microRNA and mRNA Integrated Analysis

Hyungro Lee*, Youngik Yang†, Heejoon Chae*, Seungyoon Nam‡, Donghoon Choi§, Patanachai Tangchaisin*, Chathura Herath*, Suresh Marru*, Kenneth P. Nephew¶ and Sun Kim‖

*School of Informatics and Computing, Indiana University, Bloomington, Indiana 47404, USA
Email: lee212@indiana.edu, heechae@indiana.edu, ptangcha@indiana.edu, cherath@cs.indiana.edu, smarru@cs.indiana.edu

†J Craig Venter Institute, San Diego, CA, USA
Email: yyang@jcvi.org

‡National Cancer Center, Goyang, Korea
Email: seungyoon.nam@ncc.re.kr

§Supercomputing Center, Korea Institute of Science and Technology Information, Korea
Email: choid@kisti.re.kr

¶Medical Sciences Program, Indiana University School of Medicine, Indiana University, Bloomington, IN, USA
Email: knephew@indiana.edu

‖School of Computer Science and Engineering, Seoul National University, Seoul, Korea
Email: sunkim.bioinfo@snu.ac.kr

***Abstract*—MicroRNAs, by regulating the expression of hundreds of target genes, play critical roles in developmental biology and the etiology of numerous diseases, including cancer. As vast amounts of microRNA expression profile data are now publicly available, the integration of those data sets with gene expression profiles represents an extremely active area of life science research. However, the ability to conduct genome-wide microRNA-mRNA (gene) integration currently requires sophisticated, high-end informatics tools, significant expertise in bioinformatics and computer science to carry out the complex integration analysis. In addition, increased computing infrastructure capabilities are essential in order to accommodate large data sets. In this study, we have extending BioVLAB cloud workbench to develop an environment for the integrated analysis of microRNA and mRNA expression data, named BioVLAB-MMIA. The workbench facilitates computations on the Amazon EC2 and S3 resources orchestrated by XBaya Workflow Suite. The advantages of BioVLAB-MMIA over the web-based MMIA system include: 1) readily expanded as new computational tools become available; 2) easily modifiable by re-configuring graphic icons in the workflow; 3) on-demand cloud computing resources can be used on an as needed basis; 4) distributed orchestration supports complex and long running workflows asynchronously.**

***Keywords*-Scientific workflows; Cloud computing; microRNAs; mRNAs**

I. Introduction

MicroRNAs (miRNA) are small (19-24 nucleotides) non-protein-coding RNAs that function to regulate the expression of specific gene products via hybridization to messenger RNA (mRNA) transcripts [1]. In concert with associated ribonucleoprotein complexes, miRNAs can affect translational blockade or message degradation. Although miRNAs have been implicated in numerous developmental and adult disease states [2], including cancer [3], their impact on distinct biological pathways and phenotypes remains largely unknown. Furthermore, as comprehensive mRNA and miRNA data sets are now publicly available; the ability to bioinformatics systems that can be easily used by the vast majority of scientists to fully utilize these important data sets would be extremely valuable.

Previously, we have developed a web-based system called MicroRNA and MRNA Integrated Analysis (MMIA) [4] which has been used extensively by the research community. However, tightly coupled web-based systems such as MMIA are not able to easily accommodate new tools and databases or incorporate new ways of analyzing integrated data. Furthermore, the ability of users to control complicated workflows of multiple steps is highly desired, which remains a challenge for a web-based system.

To address these key issues, we extended the bioinformatics infrastructure called BioVLAB to provide a bioinformatics systems for the integrated analysis of microRNA and mRNA data. BioVLAB is built upon the Open Grid Computing Environments (OGCE) workflow suite tools [5] and Amazon computing cloud (EC2 and S3) resources. Although BioVLAB has previously been successfully used to develop two prototype systems for protein sequence analysis [6] and gene expression data analysis [7], it has not been tested in comprehensive bioinformatics systems or in real world settings. Thus, we fully expanded the BioVLAB system architecture, implementing a system for integrated analysis of microRNA and mRNA expression data (BioVLAB-MMIA). BioVLAB-MMIA uses the latest

version of the XBaya workflow composer, which includes the following new features: 1) breakpoints that allow users to re-start workflows partially at any node; 2) a color-scheme for input, process, and output components; 3) an enhanced virtual resource control function; 4) importing/exporting workflows that allow a hierarchical workflow using existing workflows in Apache ODE (Orchestration Director Engine) Workflow Engine [8].

In this paper, we present BioVLAB-MMIA, an extended system based on the MMIA web server [4] previously developed for integrated analysis. The BioVLAB-MMIA compliments the web-based MMIA system providing additional features including: 1) seamless integration with emerging new computational tools; 2) reconfiguration and experimentation of the workflow structure and adaptation with new algorithms; 3) on-demand cloud computing resources that can be used as needed; 4) distributed orchestration supporting complex and long running workflows asynchronously.

II. Related work

A. Workflow composers

Scientists have been using scientific workflow composers to describe, construct, run, and share models and processes. A Workflow composer allows them to describe scientific tasks at a high-level abstraction and re-use scientific discoveries, which often involve complicated and time-consuming computations. Its user-centered GUI also gives a way to integrate complex and low-level programs in a simple and intuitive visual form. Nodes and arrows are key building blocks in the workflow. Nodes are executable, including web-based applications and typical application software packages. Applications on remote machines or web servers can easily be wrapped into a list of executable resources in the workflow. Arrows represent data flow between nodes, as node input and output. The user can easily compose a new workflow by dragging-and-dropping and connecting input and output between nodes.

There are several workflow composers, including Kepler [9], Taverna [10], Triana [11], GNARE [12], and XBaya [13]. These workflow tools provide many overlapping features with minor variations. The contribution of this paper is not the workflow tool itself, but applying a proven workflow tool to democratize the science tools to community at large. We chose the XBaya suite over other tools because of its flexibility in incorporating any command line driven science application as a workflow activity; support for various computational environments include on-demand cloud computing;and its interoperability with other workflow tools like Taverna. Moreover, XBaya developers are co-located with the Bioinformatics group, making for a productive collaboration.

B. Cloud computing

Cloud computing is a new paradigm of computing in which dynamically scalable and virtualized resources are provided as a service over the Internet [14]. Infrastructure, platform, and software are a service, paid for by users on an as needed basis. Amazon Elastic Compute Cloud (EC2) [15], Google AppEngine [16], and Microsoft Azure [17] are popular commercial cloud services. Amazon cloud allows a user to have a full control over virtual machines and storages as an infrastructure provider with powerful developer tools and libraries in the Amazon Web Services (AWS) developer community.

C. Cloud-based bioinformatics systems

A number of bioinformatics systems have been developed on top of the cloud computing infrastructure. Notable examples on the cloud include CloudBurst [18] for mapping a large number of short reads efficiently on a cluster of computers on the cloud, a comparative genomics system [19] that implements the reciprocal smallest distance algorithm on the Amazon EC2, and a system [20] that searches for single nucleotide polymorphisms (SNPs). CEO [21] provides a cloud computing environment for a genome-wide association study. BioNimbus [22] provides a comprehensive genome analysis environment for fly genomes and a set of 60 genomes provided by the Complete Genomics. A number of technical issues still need to be addressed for bioinformatics use, as discussed [23], [24].

III. BioVLAB system architecture

The successful deployment of the bioinformatics systems discussed above illustrates the usefulness of cloud computing. However, these systems are tightly coupled with cloud computing and not easily modified. To address this issue, we built a new bioinformatics system architecture, called BioVLAB. This concept is implemented at each layer of three-layer architecture of the BioVLAB system. At the user level, the XBaya graphical workflow lists all tools and databases in the resource panel of XBaya. Whenever a new resource is added at the gateway in the middle layer, the resource list in XBaya is automatically updated. The system developer or administrator can add new resources at the gateway in the middle layer using the OGCE web interface if the resources are installed on the server, e.g., on the Amazon EC2.

A. Three-layered BioVLAB-MMIA

The BioVLAB architecture has three distinct layers (Figure 1) based on their functionality [25]. Users only interact with the front layer, which consists of a graphical workflow composer (XBaya) and a web browser. The actual applications and data of the users are stored and exist in the back layer, which uses cloud environments. The middle layer is a collection of controllers that process user requests by

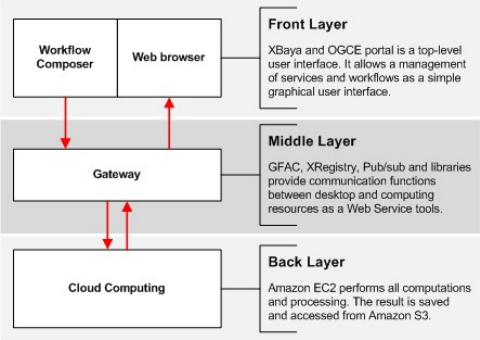

Figure 1. Three-layered BioVLAB-MMIA

communicating through the front layer and the back layer. Cloud computing resources are in the back layer, where the user's applications are executed and results are stored. Because each layer functions independently, management of components in each layer is flexible. For example, an update of a user's application in the back layer has no effect on XBaya usage and web browsing in the front layer. This approach reduces the bioinformatics burden of system management and programming skills. It is also important when developers (advanced users) want to update and implement existing workflows and modules without changing the user's experiences of the BioVLAB environment.

One important clarification is that collecting available processes for a particular bioinformatics task (the integrated analysis of microRNA and miRNA in this case) is not at all trivial. Thus, collecting computational processes and registering them in XBaya are typically far beyond the ability of ordinary users and should be done by bioinformatics experts, which is the main reason why we created the BioVLAB-MMIA cloud on Amazon cloud. For the integrated analysis of microRNA and miRNA, we have collected and registered all necessary processes based on our prior experience developing the MMIA server. However, users can freely modify the workflow with pre-composed components and/or add new components by referring to a manual at the BioVLAB project page [25].

B. Executing a long running workflow

For a workflow component that can last for hours or days, XBaya in BioVLAB-MMIA, by using GPEL engine based on Apache ODE and WS-BPEL standard provides a way to execute the process for long periods of time without connection timeout or out of memory errors. XBaya handles a long running workflow through WS invocation with WS addressing callback URL and GFAC that allows the user to retrieve the status of the job via OGCE portal service.

In addition to supporting the ODE BPEL engine, XBaya features a workflow interpreter service. This enactment service supports multiple Executable Workflow Graph Models that will identify and execute activities which have all dependencies fulfilled. XBaya provides a choice of enactment based on the characteristics of the workflow. These engines, by design, are built on open Service Oriented Architectures and incorporate asynchronous communication and inherently support long execution times. The tool suite monitors and manages any resulting discrepancies due to the long running nature of the applications. The interpreter service also features dynamic workflow execution capabilities, including: rerun or smart re-run with modified inputs; stopping execution at a break point; step through execution; add new activity; remove/replace activity.

The workflow system decouples workflow composition from workflow execution and workflow monitoring. This separation enables the following:

- Late workflow instance binding that supports a large number of applications without having to monitor and maintain persistent services. The binding from a Web service interface to a Web service instance is performed right before or during the workflow execution using the features of on-demand service creation.
- Users can compose workflows and launch them from their laptop, but it is not necessary to keep their laptop running or connected to the network throughout the execution of the workflows.
- The activities in the workflows are executed in parallel or in sequence, according to data flow and control flow dependencies. Appropriate parallelism is introduced when a workflow graph is converted into a workflow script. All dependencies are extracted from the dataflow.

IV. BioVLAB-MMIA

A. miRNA and mRNA integrated analysis (MMIA)

MMIA (microRNA and mRNA integrated analysis) [4] compares inversely correlated expression patterns between miRNA and mRNA (perfect seed-pairing between miRNA and mRNA is associated with mRNA destabilization [26]) and reports the results of two computational analyses. The first result includes disease information associated with deregulated miRNA expression and common transcription factors in upstream regions of the miRNAs. The second result includes functional, pathological and pathway information associated with inversely correlated expressed target mRNAs of the miRNAs.

Currently, MMIA is only fully compatible with human data and has five modules. The first module uses miRNA expression data or deregulated miRNA gene list as data input. The module performs a statistical test for identifying down- or up-regulated miRNAs under different conditions, e.g., control vs. treated. The second module identifies miRNA enriched gene sets based on the deregulated miRNAs. The miRNA gene set database contains two categories: disease-related miRNA gene sets [27] and transcription binding

factor sites in promoter regions of miRNA genes [28]. For example, each disease entry has an annotated miRNA gene list in the miRNA gene set database. The third module performs a significance test for mRNA microarray and reports deregulated mRNA genes. The fourth module detects mRNA targets of the deregulated miRNAs using three target prediction algorithms: TargetScan 4.2 [29], PITA [30], and PicTar [31]. The module supports single algorithms as well as the intersection between two different algorithms. The fifth module inspects inversely correlated expression between the selected miRNAs and their previously identified mRNA targets and also performs gene set analysis for the inversely expressed target mRNAs. The precompiled gene sets are used from MIT MSigDB [32], KEGG [33], and G2D [34], providing functional, pathological, and pathway information.

B. BioVLAB-MMIA cloud

BioVLAB-MMIA is a cloud computing environment that allows users to compose and modify existing workflows from currently available resources. In addition, BioVLAB is an extensible resource repository on Amazon cloud, i.e., the BioVLAB-MMIA cloud, to which we will continue to add resources (maintaining an extensible resource list is a unique feature compared to other bioinformatics systems using the cloud computing [21], [22]). Below is the current list of resources deployed for BioVLAB-MMIA.

Target prediction algorithms include TargetScan release 4.2 (www.targetscan.org) and PicTar (downloaded from the UCSC Genome Browser human NCBI). We also obtained PITA (version 6, 31-Aug-08) Sites catalog (3/15 flank; genie.weizmann.ac.il/pubs/mir07/mir07_data.html).

For gene expression analysis, Affymetrix chip annotation information (for 10 separate microarray platforms) was downloaded from the Affymetrix web page (www.affymetrix.com). The different probe names, such as Ensemble Transcript (www.ensembl.org), NCBI RefSeq, mRNA/protein and Entrez Gene (www.ncbi.nlm.nih.gov), Swiss-Prot (www.ebi.ac.uk/swissprot) and Gene Symbol (www.genenames.org), can also be downloaded.

The chip annotation information and miRNA target information in BioVLAB-MMIA are stored in the MySQL database; the gene sets are stored in MIT GSEA GMT and CHIP file formats (www.broad.mit.edu/cancer/software/gsea/wiki/index.php/Data_formats).

The mRNA gene sets consist of canonical pathways, positional information, chemical/genetic perturbation, GO, cancer genes and inherited diseases [32], [33]. These mRNA sets were from KEGG [33], MIT MSigDB [32] and G2D [34]. MSigDB has positional gene sets, chemical/genetic perturbation, canonical pathway gene sets, cancer gene sets and GO gene sets, while G2D contains inherited disease gene sets.

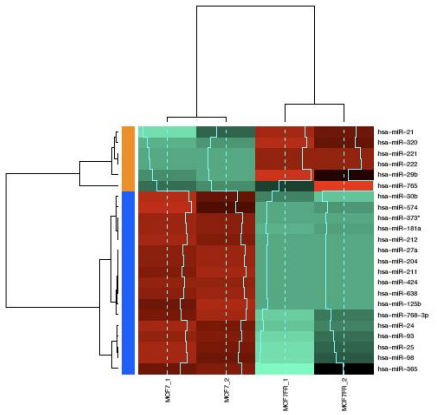

Figure 2. Significantly down regulated microRNA genes in Fulvestrant-resistant MCF7FR breast cancer cells

C. Analysis of fulvestrant drug resistant cell lines using BioVLAB-MMIA

Fulvestrant-resistant MCF7FR breast cancer cells [35] and their drug-sensitive MCF7 cells were previously developed by us (Fan et al., 2006) and used in the current experiment. Two replicates for each group (MCF7_1, MCF7_2, MCF7FR_1, MCF7FR_2 in Fig 8. a) were used for the miRNA microarray. For the mRNA microarray, three replicates for each group were obtained. MMIA assigns the first sample column to group 1, and the other sample columns (different from group 1) are assigned to group 2 in its predefined expression data format (see details in the MMIA documentation web pages [36] and [37]). In the example, group 1 is MCF7 and group 2 MCF7FR. Down-regulated miRNAs in group 2 compared to group 1 are inspected.

Figure 2 [36] is a heatmap that displays significantly down-regulated genes in the miRNA experiment (human genes). In the MMIA workflow, the heatmap is obtained after the execution of the first component (microRNAExpressionAnalysis). The UCSC Genome Browser [28] result is obtained by clicking an entry in the summary table of the second component (microRNAGeneSetAnalysis). The forth component (TargetGeneExtraction) shows an example of mRNA target prediction and the details of a gene from NCBI. The KEGG map is obtained by clicking a pathway entry in a result of combined analysis of microRNA and mRNA from executing the fifth component (GeneSetEnrichmentAnalysis). Additional details can be obtained from the BioVLAB project page [25] and the original paper [4].

D. Use of BioVLAB-MMIA

There are three types of potential BioVLAB users: a ordinary user, a workflow developer, and an advanced system developer. A ordinary user downloads a pre-composed workflow and runs the workflow with an Amazon cloud account and their data on a desktop machine. A workflow

developer can create a new workflow by adapting existing components in the resource entries instead of creating new components. An advanced system developer has the ability to manage existing or new computational components, which far exceeds the expertise of ordinary users.

E. Input and output for the analysis using BioVLAB-MMIA

BioVLAB-MMIA provides users a way to upload miRNA expression data via Amazon S3 Interface. The result reports are also publicly available on Amazon S3 buckets, which are accessible by clicking the View of the output component of XBaya Workflow GUI. Users who want to provide their data should have an Amazon cloud account and the data input should be on Amazon S3 as a SIP formatted file. SIP is a simply customized tab-delimited data format, followed by two header lines for the group and dataset descriptions. Once users have the input files on Amazon S3, Access Control for the Amazon S3 bucket (ACLs) needs to be readable for the BioVLAB-MMIA cloud. The HTTP URLs of the input files on Amazon S3 (which is formatting http://s3.amazonaws.com/USER-BUCKET-NAME/USER-INPUT-FILE-NAME) are used as input parameters of the BioVLAB-MMIA workflow. The reports are also addressable using HTTP URLs.

V. Conclusion

In this paper, we introduce the BioVLAB system, a new bioinformatics system architecture using the XBaya graphical workflow composer and Amazon cloud computing EC2 and S3. We describe an implementation of a full bioinformatics system, BioVLAB-MMIA, for the integrated analysis of mRNA and microRNA expression data. BioVLAB-MMIA has several advantages over the web-based MMIA system. It is readily expandable (new computational tools can easily be added), modifiable (graphic icons can be re-configured in the workflow), elastic (Amazon dedicated cloud computing) and stubborn (distributed orchestration supporting complex and long running workflows). Also, collecting resources on the Amazon EC2 is a resource repository that we call, the BioVLAB-MMIA cloud.

BioVLAB is particularly useful for small research labs that do not have access to high performance computing infrastructures and resources to hire full time bioinformatics and computer system specialists. BioVLAB requires users to only perform two tasks: create a virtual machine (Amazon VM in this case) and then download and execute a pre-composed workflow (BioVLAB-MMIA workflow in this case) consisting of their own data. As omics data continues to be generated by major research institutions and becomes more readily available to the life sciences community as a whole, we firmly believe that BioVLAB will significantly increase the ability and likelihood of small labs using such high-throughput data.

In the future, the resource list for BioVLAB-MMIA will be extended by implementing hierarchical workflows and managing large amount of data by Hadoop, open-source framework for distributed applications.

VI. Acknowledgements

The project is partially supported by the MetaCyt Microbial node from the Lily Foundation and National Cancer Institute grants U54 CA113001 to S.K. and K.N. S.N is supported by the Walther Cancer Foundation, Indianapolis, IN.

This work is also partially funded by the National Science Foundation through the award: OCI-1032742 (SDCI NMI Improvement: Open Gateway Computing Environments - Tools for Cyberinfrastructure-Enabled Science and Education) and Korea Institute of Science and Technology Information (KISTI), Republic of Korea.

References

[1] D. Bartel, "Micrornas: Target recognition and regulatory functions," *Cell*, vol. 136, no. 2, pp. 215–233, 2009.

[2] D. Sayed and M. Abdellatif, "Micrornas in development and disease," *Clinical Genetics*, vol. 91, no. 3, pp. 827–887, 2011.

[3] C. M. Croce, "Causes and consequences of microrna dysregulation in cancer," *Nature Reviews Genetics*, vol. 10, no. 10, pp. 704–714, 2009.

[4] S. Nam, M. Li, K. Choi, C. Balch, S. Kim, and K. P. Nephew, "Microrna and mrna integrated analysis (mmia): a web tool for examining biological functions of microrna expression," *Nucleic Acids Research*, 2009.

[5] J. Alameda, M. Christie, G. Fox, J. Futrelle, D. Gannon, M. Hategan, G. Kandaswamy, G. von Laszewski, M. A. Nacar, M. Pierce, E. Roberts, C. Severance, and M. Thomas, "The open grid computing environments collaboration: portlets and services for science gateways: Research articles," *Concurr. Comput. : Pract. Exper.*, vol. 19, pp. 921–942, April 2007.

[6] J. Y. Choi, Y. Yang, S. Kim, and D. Gannon, "V-lab-protein: Virtual collaborative lab for protein sequence analysis," in *IEEE International Conference on Bioinformatics and Biomedicine*. San Jose, CA: IEEE Computer Society, 2007, pp. 183–190.

[7] Y. Yang, J. Y. Choi, K. Choi, M. Pierce, D. Gannon, and S. Kim, "Biovlab-microarray: Microarray data analysis in virtual environment," in *Proc IEEE E-science*. Ieee, 2008, pp. 159–165.

[8] A. S. Foundation, "Apache ode," pp. –. [Online]. Available: http://ode.apache.org/

[9] B. Ludscher, I. Altintas, C. Berkley, D. Higgins, E. Jaeger, M. Jones, E. A. Lee, J. Tao, and Y. Zhao, "Scientific workflow management and the kepler system: Research articles," *Concurrency and Computation Practice and Experience*, vol. 18, no. 10, pp. 1039–1065–, 2006.

[10] T. Oinn, M. Addis, J. Ferris, D. Marvin, M. Senger, M. Greenwood, T. Carver, K. Glover, M. R. Pocock, A. Wipat, and P. Li, "Taverna: a tool for the composition and enactment of bioinformatics workflows," *Bioinformatics*, vol. 20, no. 17, pp. 3045–3054, 2004.

[11] S. Majithia, M. Shields, I. Taylor, and I. Wang, "Triana: a graphical web service composition and execution toolkit," pp. 514–521, 2004.

[12] D. Sulakhe, M. D'Souza, M. Syed, A. Rodriguez, Y. Zhang, E. M. Glass, M. F. Romine, and N. Maltsev, "Gnare–a grid-based server for the analysis of user submitted genomes," *Nucleic Acids Research*, p. gkm366, 2007.

[13] S. Shirasuna and D. Gannon, "Xbaya: A graphical workflow composer for the web services architecture," Linked Environments for Atmospheric Discovery (LEAD), Tech. Rep., 2006.

[14] "Gartner says cloud computing will be as influential as e-business," 2008. [Online]. Available: www.gartner.com

[15] "Amazon elastic compute cloud." [Online]. Available: http://aws.amazon.com/ec2

[16] "Google app engine." [Online]. Available: http://code.google.com/appengine/

[17] "Microsoft azure." [Online]. Available: http://www.microsoft.com/azure

[18] M. C. Schatz, "Cloudburst: highly sensitive read mapping with mapreduce," *Bioinformatics*, vol. 25, no. 11, pp. 1363–1369, 2009.

[19] D. Wall, P. Kudtarkar, V. Fusaro, R. Pivovarov, P. Patil, and P. Tonellato, "Cloud computing for comparative genomics," *BMC Bioinformatics*, vol. 11, no. 1, p. 259, 2010.

[20] B. Langmead, M. Schatz, J. Lin, M. Pop, and S. Salzberg, "Searching for snps with cloud computing," *Genome Biology*, vol. 10, no. 11, p. R134, 2009.

[21] Z. Wang, Y. Wang, K.-L. Tan, L. Wong, and D. Agrawal, "Ceo: A cloud epistasis computing model in gwas," in *IEEE International Conference on Bioinformatics & Biomedicine*, 12/2010 2010.

[22] "Bionimbus." [Online]. Available: http://www.bionimbus.org/

[23] L. Stein, "The case for cloud computing in genome informatics," *Genome Biology*, vol. 11, no. 5, p. 207, 2010.

[24] M. C. Schatz, B. Langmead, and S. L. Salzberg, "Cloud computing and the dna data race," *Nature Biotechnology*, vol. 28, no. 7, pp. 691–693, 2010.

[25] H. Lee, "Biovlab-mmia project homepage." [Online]. Available: http://microbial.informatics.indiana.edu/biovlab

[26] W. Filipowicz, S. N. Bhattacharyya, and N. Sonenberg, "Mechanisms of post-transcriptional regulation by microrna: are the answers in sight?" *Nature Reviews Genetics*, vol. 9, pp. 102–114, 2008.

[27] Q. Jiang, Y. Wang, Y. Hao, L. Juan, M. Teng, X. Zhang, M. Li, G. Wang, and Y. Liu, "mir2disease: a manually curated database for microrna deregulation in human disease," *Nucleic Acids Research*, vol. 37, no. suppl 1, pp. D98–D104, 2009.

[28] D. Karolchik, R. M. Kuhn, R. Baertsch., G. P. Barber, H. Clawson, M. Diekhans, B. Giardine, R. A. Harte, A. S. Hinrichs, F. Hsu, K. M. Kober, W. Miller, J. S. Pedersen, A. Pohl, B. J. Raney, B. Rhead, K. R. Rosenbloom, K. E. Smith, M. Stanke, A. Thakkapallayil, H. Trumbower, T. Wang, A. S. Zweig, D. Haussler, and W. J. Kent, "The ucsc genome browser database: 2008 update," *Nucleic Acids Research*, vol. 36, no. Database issue, pp. D773–779, 2008.

[29] B. P. Lewis, C. B. Burge, and D. P. Bartel, "Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microrna targets." *Cell*, vol. 120, pp. 15–20, 2005.

[30] M. Kertesz, N. Iovino, U. Unnerstall, U. Gaul, and E. Segal, "The role of site accessibility in microrna target recognition," *Nature Genetics*, vol. 39, pp. 1278–1284, 2007.

[31] A. Krek, D. Grun, M. N. Poy, R. Wolf, L. Rosenberg, E. J. Epstein, P. MacMenamin, I. d. Piedade, K. C. Gunsalus, M. Stoffel, and N. Rajewsky, "Combinatorial microrna target predictions," *Nature Genetics*, vol. vol. 37, pp. 495–500, 2005.

[32] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, pp. 15 545–15 550, 2005.

[33] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori, "The kegg resource for deciphering the genome," *Nucleic Acids Research*, vol. 32, pp. D277–280, 2004.

[34] C. Perez-Iratxeta, P. Bork, and M. A. Andrade-Navarro, "Update of the g2d tool for prioritization of gene candidates to inherited diseases," *Nucleic Acids Research*, pp. W212–W216, 2007.

[35] M. Fan, P. S. Yan, C. Hartman-Frey, L. Chen, H. Paik, S. L. Oyer, J. D. Salisbury, A. S. Cheng, L. Li, P. H. Abbosh, T. H.-M. Huang, and K. P. Nephew, "Diverse gene expression and dna methylation profiles correlate with differential adaptation of breast cancer cells to the antiestrogens tamoxifen and fulvestrant," *Cancer Research*, vol. 66, no. 24, pp. 11 954–11 966, 2006.

[36] "Mmia." [Online]. Available: http://cancer.informatics.indiana.edu/mmia

[37] F. Xin, M. Li, C. Balch, M. Thomson, M. Fan, Y. Liu, S. M. Hammond, S. Kim, and K. P. Nephew, "Computational analysis of microrna profiles and their target genes suggests significant involvement in breast cancer antiestrogen resistance," *Bioinformatics*, vol. 25, no. 4, pp. 430–434, 2009.