MethmiRbase: a Database of DNA Methylation and miRNA Expression in Human Cancer

David Agustriawan[#], Ezra B. Wijaya[#], Chien-Hung Huang[†], Erwandy Lim[†], I-Chin Hsueh, Nilubon Kurubanjerdjit, Ke-Rung Tzeng, Ka-Lok Ng

***Abstract*—Experimental studies suggested that epigenetic gene silencing play an important role in cancer formation. Aberrant DNA methylation is one of the epigenetic silencing mechanisms related to downregulation or upregulation of miRNA expression in human cancer cells. However, there is still few comprehensive *in silico* study concerning DNA methylation-mediated miRNA aberrant expression events. Currently there are a number of epigenetic databases but seldom of them integrate DNA-methylation and miRNAs expression to identify epigenetic-regulated miRNA which targeted oncogene or tumor suppressor genes in cancer. We present MethmiRbase (http://ppi.bioinfo.asia.edu.tw/MethmiRbase) a database that made use of meta-analysis to identify potential epigenetic-regulated miRNAs for three human cancers, i.e. lung adenocarcinoma, lung squamous carcinoma and ovarian cancers. We also included experimentally verified cancer-associated miRNA target genes in MethmiRbase.**

***Index Terms*—DNA methylation, microRNA, epigenetic, lung cancer, ovarian cancer**

I. Introduction

MicroRNA (miRNA), a class of small RNA molecules of about 22-nt long, have been shown to post-transcriptionally regulate target genes expression through translation repression of mRNA cleavage [1]. In human cancer, miRNAs can act as oncogenes or tumor suppressor genes, exerting a key function in tumorgenesis [2].

DNA methylation is an important epigenetic regulator of gene and miRNA expression in cancer formation. *In vitro* studies, which had reported the presence of DNA methylation-associated silencing of miRNA involved in human cancer. For instances, hypermethylated CpG islands of tumor suppressor miRNA such as, miR-127 and miR-124a which can be reversed by epigenetic drugs such as DNA demethylating agents [3]. In addition, regulation of aberrant methylation patterns in non-CpG island promoters may also plays an important role contribution to tumorgenesis and should therefore be included in analyses of cancer epigenetic. It has been showed that miRNA expression could be silenced by direct DNA methylation mechanism, and has potential contribution to tumorgenesis [4].

Microarray and next-generation sequencing have produced copious amounts of DNA methylation and miRNA expression data. It will facilitate the discovery of potential methylation events in human diseases. Several DNA methylation databases had been constructed. MethDB is a database that stores manually annotated experimental data on DNA methylation and environmental epigenetic effects [5]. PubMeth is an annotated and reviewed database of methylation in cancer based on text mining of the published literature [6]. Methycancer included DNA methylation, cancer-associated genes and cancer information from public data resources and large-scale experimental data sets produced from the cancer epigenome project in China [7]. NGSMethDB provides DNA methylation data for humans, chimpanzees, mice and *Arabidopsis thaliana* from publicly available data sets using the NGS bisulfate sequencing technique [8]. DiseaseMeth provides both locus-specific and high-throughput datasets for many human diseases [9]. MENT is the first database to provide both DNA methylation and gene expression information from diverse tumor tissues [10]. MethHC provides DNA methylation miRNA and gene expression information [11]. However, there is no or very few databases that provides rigorous statistical information about epigenetic-mediated miRNA events in human cancer. Also, our database identify two epigenetic-regulated miRNA pathways that can lead to cancer: (i) hypermethylation inhibits miRNA-targeted tumor suppressor genes and (ii) hypomethylation activates miRNA-targeted oncogenes. Thus, there is a great need for an integrated database that provides DNA methylation, miRNA expression, oncogene and tumor suppressor gene information.

This paper describes a DNA methylation, miRNA and gene expression database in human cancer, MethmiRbase (a database of DNA methylation and miRNA expression in human cancer. MethmiRbase made use of the meta-analysis (MA) method by using the Spearman correlation coefficient

Manuscript received December 22, 2015; revised January 21, 2016. The work of Ka-Lok Ng is supported by the grants of MOST 102-2632-E-468-001-MY3, MOST 104-2221-E-468-012 and Asia University, under the grant of 103-asia-06. The work of Chien-Hung Huang is supported by the grant MOST 104-2221-E-150-039, and National Formosa University, under the grant of EN103B-B1005.

David Agustriawan, Ezra B. Wijaya, Erwandy Lim, I-Chin Hsueh and Ke-Rung Tzeng are with the Department of Bioinformatics and Medical engineering, Asia University, Taiwan 41354. (e-mail: d.agustriawan25@gmail.com, erwandy.dev@gmail.com, ezrafisioterapi@gmail.com, chean0847@gmail.com, ncrazj@hotmail.com).

[#] David Agustriawan and Ezra B. Wijaya are the co-first authors.

[†] Chien-Hung Huang and Erwandy Lim are the co-second authors.

Nilubon Kurubanjerdjit is with the School of Information Technology, Mae Fah Luang University, Chiang Rai, Thailand 57100 (e-mail: sendtoopal@gmail.com).

*Corresponding author, Ka-Lok Ng is with the Department of Bioinformatics and Medical Engineering, Asia University, Taiwan 41354. Department of Medical Research, China Medical University Hospital, China Medical University, Tai-chung 40402, Taiwan (e-mail: ppiddi@gmail.com).

to identify significant anti-correlate DNA methylation and miRNA expression events (hypermethylation status with low expression miRNA or hypomethylation status with high expression miRNA) to construct epigenetic-mediated miRNA target gene pathways. We used Spearman correlation and MA instead of Pearson correlation because Pearson correlation is heavily influenced by outlier.

For a preliminary study, this work consider CpG islands which are located upstream and downstream of the miRNA location within a distance less than or equal to 5,000 bps. Here, we employed The Cancer Genome Atlas (TCGA) to obtain large scale of sample batches of raw data for non-small cell lung cancer (NSCLC) and ovarian cancer. To obtain high confident results, this work made use of MA to combine different batches of data together. Meta-analysis can help us to address the problems of (i) data heterogeneity, (ii) combining different datasets with sample sizes, and (iii) data dependence. This study also combined experimentally verified cancer-associated miRNA target genes and tissue specific cancer genes. Therefore, we can provided tissue specific cancer-associated target genes for miRNAs in which their expressions are significant inverse correlated with DNA methylation from meta-analysis computation.

II. MATERIAL AND METHODOLOGY

A. Database Content

Fig. 1 shows the flowchart to construct DNA methylated miRNA pathways which comprise two parts: (i) DNA-methylated miRNA and (ii) miRNA-regulated oncogenes (OCGs) or tumor suppressor genes (TSGs). A detail description is provided below.

Fig. 2 shows two possible epigenetic regulation mechanisms of miRNA pathways that can lead to cancer. Our database provides significantly inverse correlation between miRNA expression and DNA methylation, which suggested hypermethylation inhibits miRNA expression or hypomethylation activates miRNA expression. Then, for all the miRNA statistically have inverse correlation with DNA methylation, we provided their target genes which function as oncogenes or tumor suppressor genes.

B. Methods

Data Collection and Data Pre-Processing

DNA methylation and miRNA expression data were downloaded from TCGA database. NSCLC data, i.e. lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) and ovarian cancer data were downloaded from TCGA. We retrieved 'level 3' DNA methylation microarray (it represents the methylated sites/genes per sample) data and 'level 3' miRNA expression profile (it represents the calculated expression calls for genes and miRNAs per sample) and 'level 3' mRNA expression (it represents the miRNA/gene expression per sample). The platforms for DNA methylation are Human methylation 27 which has 27,578 probes and Human Methylation 450 which has 485,577 probes. The platform for miRNA are Agilent two-channel array which has 722 human mature miRNA expression and Illumina HiSeq_miRNASeq platform which has 1,046 human mature miRNA expression.

We selected six and 12 batches of patient samples for lung cancer (LUAD & LUSC) and ovarian cancer respectively. Each batch has at least 30 samples (patients). Batch samples which have DNA methylation, miRNA and mRNA expression measurements were included in our analysis. Batch IDs for LUAD are: 144, 166, 213, 238, 264 and 119, batch IDs for LUSC are 193, 214, 293, 283, 140, and 181 and batch IDs for ovarian cancer are: 9,11,12,13,14,15,17,18,19,21,22,24,27, and 40.

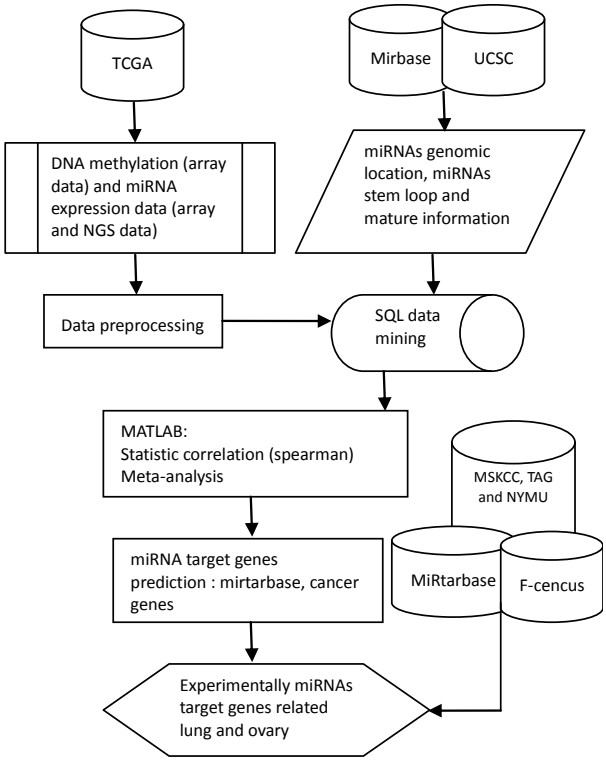

Fig. 1. Flowchart - construct DNA-methylated miRNA pathways.

Selected miRNA-methylation sites pairs

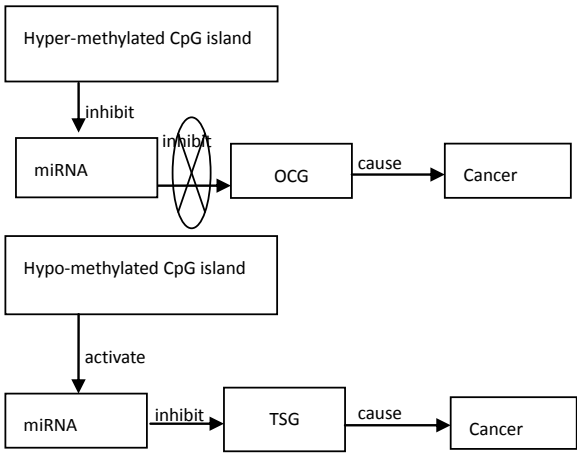

Fig. 2. Two possible epigenetic regulation mechanisms of miRNA pathways that can lead to cancer.

ISBN: 978-988-19253-8-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

TCGA database used the hg18 version to record the DNA methylation sites location. We collected miRNA's chromosome location, start location, and end location from the UCSC genome browser NCBI36 (hg18) [12]. By mapping miRNA chromosome location with DNA methylation sites, we identified miRNA-DNA methylation sites pairs where DNA methylation sites are located within 5,000 bps from the miRNA upstream and downstream regions.

Correlation and meta-analysis

This study performed the correlation study (Spearman Rank Correlation Coefficient, SRCC) between DNA methylation level and miRNA expression profile for each batch by using MATLAB. It returns the SRCC value, *rho* and *p-value* for each batch. Since each batch has different SRCC and sample size, MA was conducted in order to obtain confidence results [13]. By using MA, we can summarize and remove heterogeneity for the DNA methylation and miRNA correlation results.

The correlation score *rho* is converted to the Fisher's *z* scale and all analyses were performed using the transformed values [14]. The transformation from sample correlation *r* to Fisher's *z* is given by

$$Z = 0.5 \times ln \frac{1+r}{1-r} \qquad (1)$$

The variance of *z* is defined by $Vz = 1 / (N - 3)$, where *N* denotes the sample size. To decide which models in MA that we want to use (fixed-effect model and random-effect model), heterogeneity statistical analysis was performed. We employed the statistical test, such as, the *Q* statistics and I^2 statistics, to quantify the heterogeneity, to test it, and to incorporate it into the weighting scheme. The I^2 value is defined by,

$$I^2 = \frac{Q - df}{Q} \times 100\% \qquad (2)$$

where *df* denotes the degree of freedom, and *Q* is given by,

$$Q = \sum_{i=1}^{k} W_i (Y_i - M)^2 \qquad (3)$$

where *k*, *W*, *Y* and *M* represent the number of studies, the study weight, the study effect size and the summary effect respectively. In the fixed-effect model it is assumed that there is only one true effect size, and that all differences among the studies or batches are due to sampling errors only. In contrast, the random-effects model allows the effect size to vary from study to study. Each study estimates a different effect size.

MiRNA Target Genes Query

To obtain experimental target genes for miRNAs which are negative correlated with fixed/random effect scores less than or equal to -0.2 with p-value < 0.05, we used the miRtarbase database [15] to obtained miRNA target interactions. Then, we overlapped all the miRNA target genes from miRtarbase to our collection of cancer-associated genes. For cancer genes, both oncogenes (OCGs) and tumor suppressor genes (TSGs) were retrieved from three different resources, i.e. MSKCC [16], TAG [17] and NYMU [18]. In the final step, we used F-cencus database to determine if those target genes are cancer tissue specific genes in ovarian and lung cancer tissue specific genes [19].

III. Results

To begin the query, user can select a miRNA, methylation distance from miRNA region, meta_analysis score, p-value from the ID pull-down manual in each cancer menu as. We provided a user manual as a complete tutorial for MethmiRbase in our website. The tool will return the DNA methylation-mediated miRNA events that significantly inverse correlated with MA score < -0.2 and p-value < 0.05. Fig. 3 shows the result when user select hsa-let-7g as the input to find miRNA – DNA methylation correlation in LUSC. We set a cut off of MA correlation score to be -0.2, which is based on method published previously[20]. We obtained 192, 266 and 8 DNA methylation-mediated miRNA events that significantly inverse correlated in LUAD, LUSC and ovarian cancer, respectively. It consists of 108 candidate epigenetic-regulated miRNA events and 329 DNA methylation sites. We obtained more results for LUAD and LUSC compare with the ovarian cancer. It is because we used different DNA methylation platform. TCGA only provided Illumina Human Methylation27K beadchip for ovarian cancer which had a total of 27,578 probes. For LUAD and LUSC, TCGA provided Human Methylation450K beadchip which had a total of 485,577 probes.

For the 108 miRNAs that are significant inverse correlated with DNA methylation, users can search experimentally verified miRNA target genes that are related to specific cancer type. Fig. 4 shows when user select hsa-miR-106b as an input to find miRNA target genes related to the lung specific tissue. We collected a total of 4,445 cancer genes with 1,834 known to act as TSGs and OCGs. To find miRNA target cancer genes, we overlap all the miRNA target genes interactions from miRtarbase to our cancer gene collection. According to the F-cencus database we identified a total of 138 and 94 cancer genes that are ovarian and lung tissue specific, respectively.

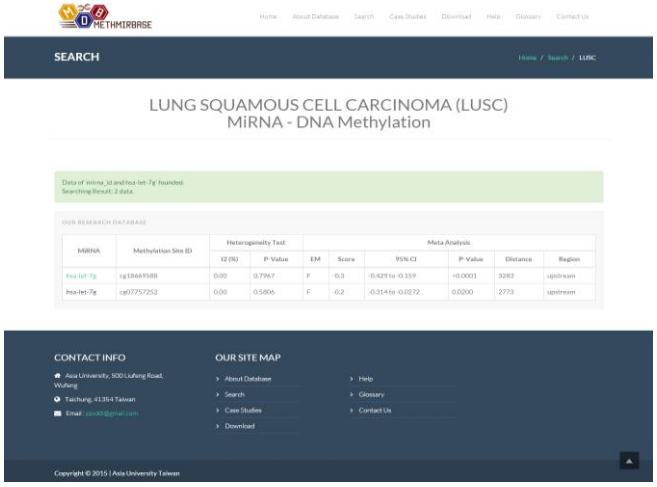

Fig. 3. DNA methylation-mediated miRNA events search menu

ISBN: 978-988-19253-8-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

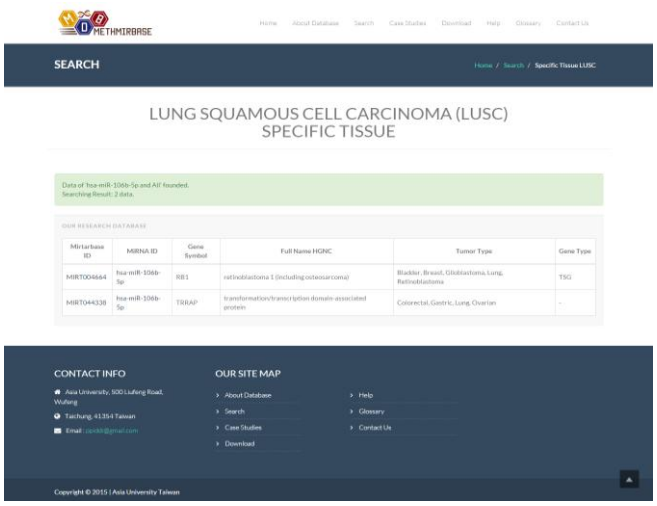

Fig. 4. miRNA cancer target genes search menu

IV. CONCLUSION

Our understanding of how epigenetic changes affect the regulation of miRNAs remains immature. An important contribution of the current study is to conduct genome-wide scale study of epigenetic-mediated miRNAs using lung adenocarcinoma, lung squamous cell carcinoma and ovarian cancers as an illustration. We computed the SRCC for DNA-methylation and miRNA events in several batches of LUAD, LUSC and ovarian cancer, and performed MA to conclude the correlation results. Then, we obtained significantly inverse correlation in DNA methylation and miRNA events. Therefore, this information might suggest that downregulated or upregulated of miRNA is correlated with its DNA methylation status in the upstream or downstream region. In other words, these methylation sites are potential direct DNA methylation-mediated miRNA-targeted gene pathways in lung and ovarian cancers, and demonstrated the present approach might have the effectiveness to predict epigenetic-regulated miRNA pathways.

Our database contributes in identifying two possible epigenetic regulation mechanisms of miRNA pathways that can lead to cancer, i.e. (i) hypermethylation inhibits miRNA-targeted TSGs, and (ii) hypomethylation activates miRNA-targeted OCGs. Methmirbase provides potential epigenetic-regulated miRNA pathways in LUAD, LUSC, and ovarian cancer.

We integrated methylation sites expression, miRNA expression, CpG island location, target genes, cancer genes and lung/ovarian tissue-specific cancer genes information to construct epigenetic-mediated miRNA targeted pathways. This work demonstrated that miRNAs can possibly be silenced or activated by epigenetic mechanisms and involve in cancer formation.

One of the main advantages of the present study on miRNA target information is that all the target genes information are experimentally verified. The results provided here are solid, have fewer false positive events. This platform provides an easy means of investigating epigenetic-regulated miRNAs in cancer study.

In the future study, we will include more cancer types in MethmiRbase. Normal tissue datasets obtained from TCGA will be included and the 5000 bps interval will be extended to 100kbps in future investigation.

V. AVAILABILITY

Methmirbase database will be continuously maintained and updated. The database is now publicly accessible at http://ppi.bioinfo.asia.edu.tw/MethmiRbase.

REFERENCES

[1] L. He and G. J. Hannon, "MicroRNAs: small RNAs with a big role in gene regulation," *Nat Rev Genet*, vol. 5, pp. 522-31, Jul 2004.

[2] A. Lujambio, G. A. Calin, A. Villanueva, S. Ropero, M. Sanchez-Cespedes, D. Blanco, *et al.*, "A microRNA DNA methylation signature for human cancer metastasis," *Proc Natl Acad Sci U S A*, vol. 105, pp. 13556-61, Sep 9 2008.

[3] A. Lujambio and M. Esteller, "CpG island hypermethylation of tumor suppressor microRNAs in human cancer," *Cell Cycle*, vol. 6, pp. 1455-9, Jun 15 2007.

[4] H. Han, C. C. Cortez, X. Yang, P. W. Nichols, P. A. Jones, and G. Liang, "DNA methylation directly silences genes with non-CpG island promoters and establishes a nucleosome occupied promoter," *Hum Mol Genet*, vol. 20, pp. 4299-310, Nov 15 2011.

[5] C. Grunau, E. Renault, A. Rosenthal, and G. Roizes, "MethDB—a public database for DNA methylation data," *Nucleic Acids Research*, vol. 29, pp. 270-274, 2001.

[6] M. Ongenaert, L. Van Neste, T. De Meyer, G. Menschaert, S. Bekaert, and W. Van Criekinge, "PubMeth: a cancer methylation database combining text-mining and expert annotation," *Nucleic Acids Res*, vol. 36, pp. D842-6, Jan 2008.

[7] X. He, S. Chang, J. Zhang, Q. Zhao, H. Xiang, K. Kusonmano, *et al.*, "MethyCancer: the database of human DNA methylation and cancer," *Nucleic Acids Res*, vol. 36, pp. D836-41, Jan 2008.

[8] M. Hackenberg, G. Barturen, and J. L. Oliver, "NGSmethDB: a database for next-generation sequencing single-cytosine-resolution DNA methylation data," *Nucleic Acids Res*, vol. 39, pp. D75-9, Jan 2011.

[9] J. Lv, H. Liu, J. Su, X. Wu, H. Liu, B. Li, *et al.*, "DiseaseMeth: a human disease methylation database," *Nucleic Acids Res*, vol. 40, pp. D1030-5, Jan 2012.

[10] S. J. Baek, S. Yang, T. W. Kang, S. M. Park, Y. S. Kim, and S. Y. Kim, "MENT: methylation and expression database of normal and tumor tissues," *Gene*, vol. 518, pp. 194-200, Apr 10 2013.

[11] W. Y. Huang, S. D. Hsu, H. Y. Huang, Y. M. Sun, C. H. Chou, S. L. Weng, *et al.*, "MethHC: a database of DNA methylation and gene expression in human cancer," *Nucleic Acids Res*, vol. 43, pp. D856-61, Jan 2015.

[12] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, *et al.*, "The human genome browser at UCSC," *Genome Res*, vol. 12, pp. 996-1006, Jun 2002.

[13] A. X. Garg, D. Hackam, and M. Tonelli, "Systematic review and meta-analysis: when one study is just not enough," *Clin J Am Soc Nephrol*, vol. 3, pp. 253-60, Jan 2008.

[14] M. Borenstein, L. V. Hedges, J. P. T. Higgins, and H. R. Rothstein., "Introduction to Meta-Analysis," *United Kingdom,UK : John Wiley & Sons, Ltd*, 2009.

[15] S. D. Hsu, Y. T. Tseng, S. Shrestha, Y. L. Lin, A. Khaleel, C. H. Chou, *et al.*, "miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions," *Nucleic Acids Res*, vol. 42, pp. D78-85, Jan 2014.

[16] M. E. Higgins, M. Claremont, J. E. Major, C. Sander, and A. E. Lash, "CancerGenes: a gene selection resource for cancer genome projects," *Nucleic Acids Res*, vol. 35, pp. D721-6, Jan 2007.

[17] H. H. Chan, "identification of novel tumor associated gene (TAG) by bioinformatics analysis," *MSC. thesis, National cheng kung university Taiwan*, 2006.

[18] C. Y. Huang, "institute of clinical medicine, National Yang Ming University, private communication."

[19] X. Gong, R. Wu, Y. Zhang, W. Zhao, L. Cheng, Y. Gu, *et al.*, "Extracting consistent knowledge from highly inconsistent cancer gene data sources," *BMC Bioinformatics*, vol. 11, p. 76, 2010.

[20] "Integrated genomic analyses of ovarian carcinoma," *Nature*, vol. 474, pp. 609-15, Jun 30 2011.

ISBN: 978-988-19253-8-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)